Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge

Brent Hecht[1] and Emily Moxley[2]

[1] Electrical Engineering and Computer Science, Northwestern University, Frances Searle
Building #2-419, 2240 Campus Drive, Evanston, IL 60208
brent@u.northwestern.edu
[2] Electrical and Computer Engineering, UC Santa Barbara, Mailbox 217, Harold Frank Hall,
Santa Barbara, CA 93106
emoxley@ece.ucsb.edu

Abstract. The First Law of Geography states, "everything is related to everything else, but near things are more related than distant things." Despite the fact that it is to a large degree what makes "spatial special," the law has never been empirically evaluated on a large, domain-neutral representation of world knowledge. We address the gap in the literature about this critical idea by statistically examining the multitude of entities and relations between entities present across 22 different language editions of Wikipedia. We find that, at least according to the myriad authors of Wikipedia, the First Law is true to an overwhelming extent regardless of language-defined cultural domain.

Keywords: Tobler's Law, First Law of Geography, Spatial Autocorrelation, Spatial Dependence, Wikipedia

1 Introduction and Related Work

When he first posited the statement that "everything is related to everything else, but near things are more related than distant things" [1] almost 40 years ago[1], Waldo Tobler had no idea it would have such staying power. Widely accepted as the First Law of Geography and also frequently known as simply Tobler's Law or Tobler's First Law (TFL), this assertion appears in nearly every geography and Geographic Information Systems (GIS) textbook printed today. Moreover, many social and physical sciences have adopted as existentially essential the ideas of spatial dependence or spatial autocorrelation, both of which are accessibly and succinctly defined by Tobler [2]. TFL has even spawned a First Law of Cognitive Geography, which states that in the context of information visualization "people *believe* closer things are more similar" [3, 4]. Obviously, all of these ideas have proven to be of major applied use in a vast array of well-known work solving a vast array of specific

[1] Although his paper was published in 1970, he first presented his work at a 1969 meeting of the International Geographic Union's Commission on Quantitative Methods, making this year arguably the law's true 40th anniversary.

problems. However, Tobler's statement is high-level, domain-neutral, and problem-independent in scope and it has never been empirically evaluated in these terms. Many authors [5-9] have opined on the topic at a philosophical level, but no experiments have been done. A data-based investigation of such a broad statement has enormous challenges associated with it, and at least part of the reason for this gap in the literature has been the lack of available data to examine.

However, that hurdle was overcome with the development and rise to immense popularity of Wikipedia, the collaboratively authored encyclopedic corpus of unprecedented scale. While it is by no means perfect as a representation of the sum of world knowledge, it is by far the closest humanity has come to having such a data set. As of this writing, Wikipedia, consistently ranked as one of the top ten most-visited websites on the Internet, contains 2.76 million articles in its English edition, and has a total of 25 language editions with over 100,000 articles (see Table 1 for descriptive statistics of the data used in our studies). Each of these articles describes a unique entity.

All of these facts are relatively well known by the Internet-using population. What is less understood is the scope of the quantity of *relations* between these entities present in Wikipedia. The relations, encoded by contributors ("Wikipedians"), and viewed as links to other Wikipedia pages by visitors, number well into the hundreds of millions. Although these unidirectional relations are not typed (except in some demonstration versions of Wikipedia such as "Semantic Wikipedia" [10]), they can still tell us which of the millions of entities are related in some way, and which are not.

We seek to leverage the entities and relations in this enormous data set to examine the validity of Tobler's Law in the very general context described above. While our experiment is, to our knowledge, the most broad empirical investigation of Tobler's Law done to date, it does have its limitations. Critically, we of course do not claim to evaluate the First Law on a representation of *all* spatial data in existence. Rather, due to our data source, our results will only confirm or deny the validity of TFL in the world *as humans see it*. We do assert that Wikipedia data is a reasonable, although flawed, proxy for the world as it is understood by humans. Ignoring this proxy, our experiments will at least determine the validity of TFL in the context of the world knowledge that has been represented by the millions of people who have contributed to Wikipedia (although most of it has been authored by a smaller number of people [11]), is accessed by countless millions more, and is used by dozens of systems in AI and NLP (e.g. [12, 13])

Along the same lines, it is important to at least briefly address the question of accuracy. While it has been found that Wikipedia's reputation for questionable intellectual reliability has been somewhat unfairly earned [14], the nature of our study almost entirely sidesteps the accuracy concern. Because we examine entities and relations in aggregate and rely far more on their existence than their specific details, we can to a large degree ignore accuracy risks. An Internet user would have to very purposely manipulate massive amounts of specific data across many languages of Wikipedia to be able to change the results of our experiments. Non-malicious systemic characteristics of Wikipedia do create their problems, but we describe in detail how we address these, point out when we are unable to, and discuss the problems therein. In summary, like all science, this study is subject to the rule of "garbage in, garbage out." However, judging by the number of papers published in the past few years on Wikipedia or using Wikipedia data in the fields of computer science, psychology, geography, communication, and more [7, 11, 15-17], Wikipedia has been assumed to be far better than garbage by *large* numbers of our peers. More specific to this project, many of the exact same structures leveraged in this study have been used, for instance, to calculate semantic relatedness values between words with much greater accuracy than WordNet [18].

Very rare even in Wikipedia research is our intensely multilingual approach. Less than a handful of papers attempt to validate conclusions in more than one language's Wikipedia. Almost always, the English Wikipedia is considered to be a proxy for all others, a problematic assumption at best. As a way to distinguish the opportunities and challenges of including a double-digit number of languages in our study, we describe our work as using not a multilingual data set, but a "hyperlingual" data set [19] (in analogy to multispectral and hyperspectral imagery). In the end, we consider our results to be much more valid because they are similar across the entire hyperlingual Wikipedia data set rather than being restricted to a single or small number of Wikipedias.

2 Data Preparation

2.1 WikAPIdia

The foundation of our work is WikAPIdia, a Java API to hyperlingual Wikipedia data that we have developed. Available upon request for academic research[2], WikAPIdia provides spatiotemporally-enabled access to any number of Wikipedias (a language edition of Wikipedia is often referred to as simply "a Wikipedia"). WikAPIdia also has a large number of graph analysis and natural language processing features built-in.

WikAPIdia initially takes as input the XML database dumps provided by the Wikimedia Foundation, the non-profit that manages Wikipedia. These XML files contain a "snapshot" in time of the state of the Wikipedia from which they are derived. The user must input the XML dump of each language she wishes her instance of WikAPIdia to support. For this study, we used XML dumps for the 22 different Wikipedias that had around 100,000 articles or more at the time of the dump (see Table 1). Since maximum temporal consistency is desirable in order to minimize external effects, snapshots from as close as possible to the most recent English dump, that of 8 October, 2008, were used.

The parsing of these 22 dump files takes a relatively new, moderately powered desktop PC approximately several days to complete. During this process, a large number of structures are extracted from the very semi-structured Wikipedia dataset, some of which are not used in this study. Those that are of importance to our experiment are described in detail below.

[2] Contact the first author for more details.

Table 1. Descriptive statistics of the Wikipedia Article Graph (WAG) and the number of spatial articles for each of the Wikipedias included in this study.

Language	**Vertices (Articles) =** $|V|$	**No. Edges (Links) =** $|E|$	**Spatial Articles =** $|V_{spatial}|$
Catalan	141,277	3,478,676	13,474
Chinese	203,824	5,566,490	14,177
Czech	112,057	3,089,517	8,599
Danish	97,825	1,714,025	7,118
Dutch	497,902	9,679,755	103,977
English	2,515,908	76,779,588	174,906
Finnish	208,817	3,782,563	11,559
French	716,557	20,578,831	67,042
German	827,318	21,456,176	85,906
Hungarian	120,850	3,009,814	6,939
Italian	516,120	14,968,632	67,433
Japanese	532,496	20,946,112	21,621
Norwegian	193,298	3,774,509	16,607
Polish	555,563	12,678,608	58,367
Portuguese	437,640	8,577,435	79,844
Romanian	118,345	1,434,939	20,349
Russian	341,197	7,762,322	30,346
Slovakian	102,089	1,931,138	7,708
Spanish	443,563	12,576,477	42,431
Swedish	295,605	5,555,219	18,816
Turkish	120,689	2,260,241	5,431
Ukrainian	131,297	1,743,304	4,692
TOTAL	**9,230,237**	**243,344,371**	**867,342**

2.1.1 The Wikipedia Article Graph

From each Wikipedia one of the key structures extracted by WikAPIdia is the Wikipedia Article Graph (WAG). The WAG is a graph (or network) structure that has as its vertices (nodes) the articles of the Wikipedia and the links between the articles as its edges. In graph theory terminology, the WAG is a directed sparse multigraph because its edges have direction (a link from one article to another), each node is connected to a relatively small number of other nodes, and it can contain more than one link from an article to another article. This last characteristic is a problem for our study, and our workaround is described in section three. Even the "smallest" Wikipedias have relatively enormous WAGs. An overview of the basic size characteristic of each WAG is found in Table 1.

2.1.2 Interlanguage Links

Encoded by a dedicated and large band of Wikipedians aided by bots that propagate their work across the various language editions of Wikipedia, interlanguage links are essentially multilingual dictionary entries placed in each Wikipedia article. By parsing out these links, WikAPIdia is able to recognize that the English article “Psychology”, the German article “Psychologie”, and the Chinese article “心理学” all refer to the same concept. Although interlanguage links provide much of the raw material for creating WikAPIdia’s multilingual concept dictionary, a significant amount of detailed post-processing is necessary to mediate conflicts contained within these links.

2.1.3 Spatial Articles

The spatial data used by WikAPIdia all derives from the latitude/longitude tags included in tens of thousands of articles by contributors to several different Wikipedias[3]. The motivation to contribute a tag is to allow readers of a tagged article to click a link to see the location of the subject of the article in any Internet mapping application such as Google Maps. However, WikAPIdia uses these tags for a modified purpose. In addition to providing their specific location in a geographic reference system[4], the tags inform WikAPIdia that the subjects of tagged articles are *geographic* entities. In other words, articles tagged with latitude and longitude coordinates can be called *spatial articles*. For example, the English article “Country Music” is very unlikely to contain a latitude/longitude tag, whereas the article “Country Music Hall of Fame and Museum” does include a tag and can be included in the class of spatial articles.

Although contributors to any of the Wikipedias included in this study are theoretically capable of tagging articles in their language of choice with spatial coordinates, in practice, this is not done in many of the “smaller” Wikipedias with any significant degree of coverage. As such, WikAPIdia uses interlanguage links to copy a tag across all languages’ articles that refer to the same concept. For instance, although the German article “Country Music Hall of Fame” does not possess a geotag, it does include an interlanguage link to the English article mentioned above. WikAPIdia copies the tag from the English article to the German article, the Spanish article “Museo y Salón de la Fama del Country”, etc. The number of spatial articles found for each Wikipedia is listed in Table 1.

2.2 The Scale Problem

Once our instance of WikAPIdia had successfully parsed and processed all 22 language dump files, an additional stage of data preparation was necessary due to what we call the “Geoweb Scale Problem” (GSP). Stated simply, GSP arises because most Web 2.0 spatial data representation schemas only support point vector features. The “blame” for this limited representational expressiveness can probably be split

[3] WikAPIdia has its own spatial data parsers, but also supports the tags collected by the Wikipedia-World Project (http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World/en).

[4] WikAPIdia assumes all latitude and longitude tags are derived from World Geodetic System 1984 (WGS1984) due to its popularity amongst the general public.

between designers' lack of education about geographic information as well as a dearth of popular tools that support vector features of greater than zero dimensions. For many geoweb applications, GSP does not restrict functionality a great deal, but in some cases, the points-only paradigm borders on ridiculous.

We are able to sidestep the smaller inaccuracies introduced by Wikipedia's point-based geotagging system by choosing the appropriate scale for analysis. In our experiments, we adopt a 50-kilometer precision for this purpose. We thus consider, for example, all points between 0 and 50km from each other to be of equal distance from each other. The problem that both all cities and all articles about places within those cities are encoded as points, for instance, can be almost entirely solved as long as we "flatten" our distance function in this way. For example, the spatial article pair "Chicago" and "O'Hare International Airport", which lies within Chicago, would be correctly placed in the 0-50km distance class instead of being assigned the false precision of 25km.

However, certain egregious point spatial representations cannot be reasonably handled using the above methodology. The U.S. state of Alaska, for instance, is encoded as a point (64ºN, 153ºW) in Wikipedia. Cartographically speaking, the only scale at which this representation would be valid is perhaps if one were examining Earth from Mars. Unfortunately, the relatively small number of cases that suffer from representation issues of this degree play a disproportionately large role in our study because these articles tend to have the most inlinks (or relations directed at them). Without further steps, the link from "Anchorage, Alaska" to "Alaska" and thousands like it would be falsely considered in a very high distance class (if the point representations are used, the "Anchorage" / "Alaska" pair would be in the 300-350km distance class, for example).

The only way to tackle this issue is to associate more appropriate representations for this set of particularly problematic point geotags using external data sources. To accomplish this task, we used the polygonal data and toponyms (place names) for countries and first-order administrative districts (like Alaska) included with ESRI's ArcGIS product[5]. We matched this data to Wikipedia articles in all languages by using Wikipedia's "redirect" structures and interlanguage links. Redirects are intended to forward users who search for "USA", for example, to the article with the title "United States." However, they also represent an immense synonymy corpus. The combination of these synonyms and ESRI's toponym-matched polygon data resulted in a relatively effective polygonal gazetteer for countries and first-order administrative districts.

We tried two different methods of leveraging this polygonal information. First, we used a point-in-polygon algorithm to map all point/polygon pairs to a distance of 0, as discussed in section three. Second, we simply stripped all the countries and first-order administrative districts found by our georeferencer from our data set in order to maintain a consistent scale of analysis. Interestingly, the main and secondary characteristics of our results – as described below – were very similar or identical between both methods (Figure 3.3, for example, is nearly identical trend-wise in both cases). As such, since the point-in-polygon algorithm was sufficiently computational complex to limit our sample size extensively (see section 3) and since the details of

[5] Specifically, we used data in the "admin" and "countries" shapefiles.

the choices we made crossing scales are too numerous to fit in this limited space, we discuss our second method in the remainder of this paper and leave analysis of multiple scales to future work. We also note that one of the several proposed "second laws" of geography states "everything is related to everything else, but things observed at a coarse spatial resolution are more related than things observed at a finer resolution." [20] This would suggest that by choosing a more local scale of analysis, we are selecting the option at which less spatial dependence in relations would occur, so TFL's validity should be most tested at this scale.

3 The Main Experiment

3.1 Hypothesis

Let us consider three spatial articles *A*, *B*, and *C* where *distance*(*A*,*B*) < *distance*(*A*,*C*) (assuming the function *distance()* calculates the straight-line distance from the entities described in the articles in the function's parameters)[6]. If Tobler's statement that "near things are more related than distant things" is indeed true, it is expected that spatial article *A* would be more likely to contain a relation/link to *B* than to *C*. In other words, we hypothesize that given three spatial articles *A*, *B*, and *C*, if TFL is valid, $P_{relation}(A,B)$ will be generally greater than $P_{relation}(A,C)$ if *distance*(*A*,*B*) < *distance*(*A*,*C*), where $P_{relation}$ is the probability of the first spatial article containing a link/relation to the second. We also hypothesize, however, given TFL's first clause "everything is related to everything else," that $P_{relation}(A,C) > 0$, even if *A* and *C* are separated by a very large distance.

3.2 Methods

We test this hypothesis on the hyperlingual data set parsed and prepared by WikAPIdia. Stated simply, our basic methodology is to examine all pairs of spatial articles [*A*, *B*] (excluding the identity case, or [*A*, *A*]) and record for each pair the straight-line distance between them and whether or not *A* contains a link to *B* or *B* contains a link to *A*. We perform this analysis *separately* for each Wikipedia in order to compare the results for each language-defined data set.

Taking a page from the field of geostatistics, once we have all the relation existence / distance tuples (examples of tuples include (1) *A* and *B* are 242km apart and *A* links to *B,* and (2) *A* and *C* are 151km apart and *A* does not link to *C*), we group these tuples into distance *lag* classes of 50 kilometers[7] and evaluate the overall probability, P_d, of a link existing between two spatial articles that are separated by

[6] As has been noted by many authors writing about Tobler's Law, straight-line distance is only the simplest of the many possible distance metrics that could be used. We leave experiments with more complex measures to future work.

[7] We identified 50km as the minimum precision possible due to the reasons presented above in our discussion of scale.

each distance class d. We measure the overall probability by calculating the number of existing relations in each distance class and dividing that by the total number possible links, as shown below:

$$\frac{\sum_{i=1}^{|PAIR_d|} relation(PAIR_{di})}{|PAIR_d|} = P_d$$

where $PAIR_d$ is the set of all the spatial article pairs in distance class d, *relation()* evaluates to 1 if there exists an outlink from the first spatial article in the pair to the second and 0 otherwise, and $|PAIR_d|$ is the number of spatial article pairs. It is important to note that [A, B] and [B, A] are considered to be different pairs because A can contain a relation to B, but also vice versa, and both must be considered in evaluating the number of possible links. From these lag-based statistics, we are able to derive empirical functions that form the basis of our results and analysis presented in the following subsections. These functions are effectively "pseudo-covariograms", with probabilities of relation/link existence in the place of covariance. We call these functions representing the relatedness between points spatial "relatograms" ("relate-oh-grams").

Even for the "smaller" Wikipedias, executing the above algorithm is a very computationally complex task. Consider the Catalan Wikipedia, for example. Since it contains 13,474 spatial articles, it was necessary to query WikAPIdia for the distance and existence of links between $|V_{spatial}|$^2- $|V_{spatial}|$ = 13,474^2 – 13,474 = 181, 535,202 pairs. Since we must consider the pairs [A, B] and [B, A] to be different, we are unable – at least in the link existence portion of the measurements – to take advantage of the computational benefits of symmetry.

Since WikAPIdia makes extensive use of MySQL tables to store data, a great number of iterations require disk operations. Even with the large number of optimizations we wrote, this makes the algorithm even more time intensive. As such, for the English Wikipedia, for instance, doing the requisite $|V_{spatial}|$^2 - $|V_{spatial}|$ = ~30.5 billion iterations was simply impractical. Thus, for all Wikipedias in which $|V_{spatial}|$ > 50,000, we used a random subset (without replacement) of size 50,000. Since 50,000 spatial articles represents 28.5 percent at minimum of a Wikipedia's dataset, we are able to attain tractability without risking statistical insignificance.

In order to compare our results across all the Wikipedias, it was necessary to normalize by a measure of each Wikipedia's overall "linkiness." The measure we use is the probability of a link occurring from any random article (not necessarily spatial) X to any other random article Y. We calculated this measure by using the following formula:

$$P_{relation}(X,Y) = \frac{|E_{adjusted}|}{|V|^2 - |V|} = P_{random}$$

where $|V|$ = the number of total articles in the Wikipedia (not just spatial articles), and $|E_{adjusted}|$ is the number of non-duplicate links in the Wikipedia. Duplicate links occur when editors add two or more outlinks in any *X* to any *Y*. We evaluated |E$_{adjusted}$| by calculating the average link duplication over a random sample of 10,000 links in each Wikipedia and dividing |*E*| by this number. The average link duplication ranged from about 1.08 (or, each link from an *X* to a *Y* appears on average 1.08 times) to 1.26, so this was an important normalization.

Finally, in our results and analysis below, we frequently make use of a ratio that allows us to complete the sentence, *"If spatial article A is separated from spatial article B by distance class d, it is ___ times as likely as random to contain a link to B."* Given distance class *d*, this ratio is simply:

$$\frac{P_d}{P_{random}}$$

3.3 Results and Basic Analyses

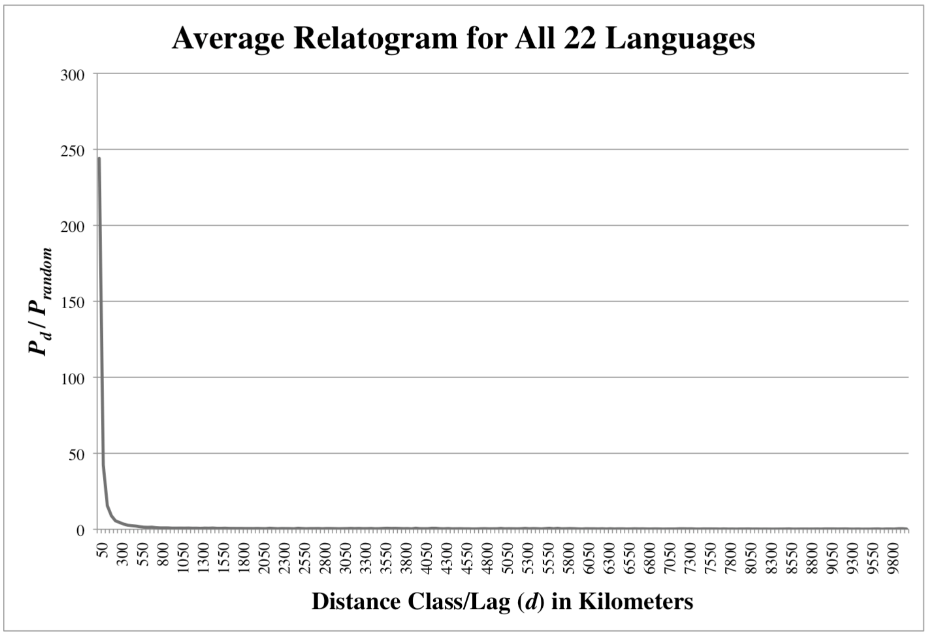

Fig. 3.1. A "relatogram" of the unweighted average of P_d / P_{random} across all 22 Wikipedias included in our study. The *y*-axis thus describes the average multiple of random probability that a link will occur from *A* to *B* given *d*. Along the *x*-axis are the distance classes/lags considered, or all the *d*s (0-50km, 50-100km, etc). The graph looks like that of any variable showing a great degree of spatial dependence: a large amount of relatedness at small distance classes, and a very large drop-off as larger distances are considered.

As can be seen in Figure 3.1, if spatial articles *A* and *B* are within 50km of each other, they are around *245 times* as likely to have a relation connecting them than if they were any two random Wikipedia articles on average[8]. This spatial bias drops off rapidly, however, and by 650km or so, all significant positive spatial dependence goes away. In other words, Figure 3.1 clearly shows that our hypothesis that $P_{relation}(A,B)$ is generally greater than $P_{relation}(A,C)$ if $distance(A,B) < distance(A,C)$ is true[9]. Despite the fact that Wikipedians by no means attempted to create a resource that displayed relation spatial dependence, they nonetheless did so in dramatic fashion, and did so regardless of their language-defined cultural domain. Without exploring further, we can state firmly that Tobler's Law has been validated empirically on a massive repository of world knowledge. In Tobler's words but our emphasis, "near things *are* more related than distant things."

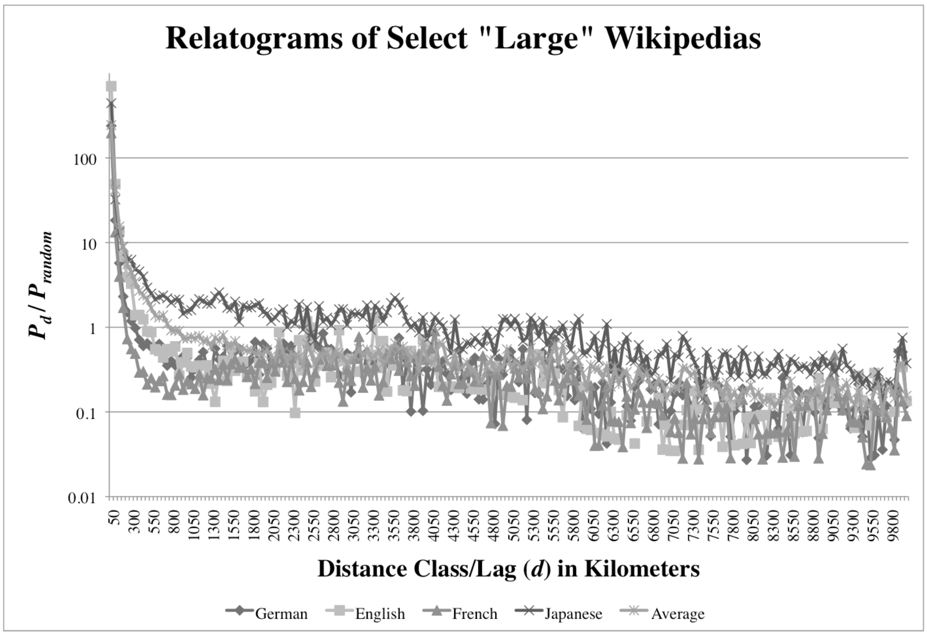

Fig. 3.2. "Relatograms" of selected large Wikipedias and the unweighted average of P_d / P_{random} across all 22 Wikipedias included in our study. As opposed to Figure 3.1, the *y*-axis is displayed on a logarithmic scale, allowing easier discrimination of the variation occurring at higher distance classes and lower probability ratios. A value of $y = 1.0$ means that P_d, or the probability of a link occurring between articles pairs in distance class *d*, is equal to P_{random}, or the probability of any two articles having a link to one another in the Wikipedia being examined.

We also see in figures 3.1, 3.2, and 3.3 that no matter the distance class, the probability of *A* and *B* having a relation is never consistently zero, affirming the accepted meaning of the non-spatial dependence clause of TFL, that "everything is

[8] By "average", we mean the average of all 22 languages' relatograms without weighting by number of articles considered.

[9] Since for most language we consider the whole dataset and for those we do not our *n* is in the millions, we do not show error bars as they would be undefined or microscopically small.

related to everything else." In some of the smaller Wikipedias, P_d occasionally drops to zero, but never does so consistently. This can be seen in the intermittent gaps in the series in figure 3.3 (*log*(0) is undefined, so is displayed as a gap).

Beyond confirming our hypothesis, however, this experiment also produced several second-order observations and, as is usual, raised more questions than it answered. One of the most important secondary patterns we noted is that, in all Wikipedias, beyond a certain threshold $d = \phi$, P_d drops below P_{random} and thus the ratio displayed in the charts becomes less than 1.0. Once this occurs, it nearly always stays that way, as shown in Figures 3.2 and 3.3. In many cases, P_d is consistently less than 25 percent of P_{random}, meaning a spatial article *A* is at least four times more likely to have a link to any *random* article *Y* than it is to a spatial article *B* after *distance*(*A*,*B*) reaches ϕ. In other words, while short distances have a remarkably strong positive effect on the probability that two spatial articles will be related, larger distances actually have a noticeable *negative* effect on these probabilities.

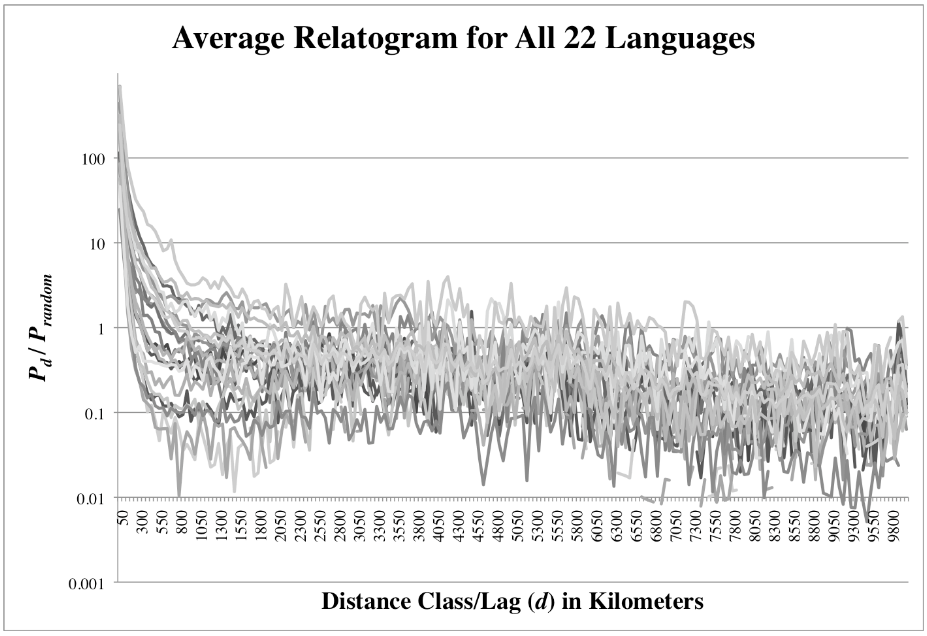

Fig. 3.3. The variation between and similarities in the relatograms in all 22 languages. The purpose of this chart is not to be able to follow individual languages, but to see overall trends.

Additionally, although the primary signal of spatial dependence is obvious, the variation among the different languages is fascinating and somewhat of an enigma. Why does the French Wikipedia demonstrate such an immediate drop to a near-asymptote while the Japanese Wikipedia displays a much more gradual decent? What results in the highly varying initial probability ratios for the 0-50km distance class? No correlation with any common network measures (i.e. number of nodes, etc.) explains these or other notable differences. As such, we have to assume Wikipediacentric, cultural or linguistic variation to be the cause. We discuss our ideas with regard to this beguiling phenomenon and our plan for further research in this area in the future work section.

3.4 Advanced Analyses

In the previous section, we qualitatively discussed our results. In this section, we seek to analyze them with more mathematical rigor so as to better understand the empirical meaning of TFL as suggested by Wikipedia. Our primary aim is to show that the relatograms can be reasonably described as power law distributions, at least as compared to the "distance doesn't matter" model of the uniform distribution.

Power laws are observed frequently in both the manmade and natural world. An observation of a phenomenon, *g*, that follows a power law with scaling exponent *k* over varying *x*, is governed by the equation:

$$g(x) = ax^k + o(k)$$

where $o(k)$ is asymptotically small. The scaling invariance of the distribution, that is $g(cx)$ α $g(x)$, becomes clear when we examine the distribution in log-log space:

$$\log(g(x)) = k\log(x) + \log(a)$$

From this we can see that a necessary condition for a power law is a straight line in log-log space. The slope of the line provides scaling exponent *k*.

Examining the probability of a link, P_d / P_{random}, at varying lags for *some* Wikipedias gives the appearance of a power law. The straight lines seen in log-log plots of the relatograms of these Wikipedias (Figure 3.4) reveal that indeed they appear to follow a power law over a selected distance range. We fit a power law distribution to the data, limiting the lag to 1000km, using Bayesian probability theory to find the best-fit parameters [21]. This amounts to finding the parameters *a,k* in the power law equation above, that given the observed distribution *f(x)* and an assumption H_{pl} of power-law distributed data, give the function that is most in accordance with the data. A maximum likelihood estimation is equivalent to the maximum a posteriori (MAP) optimization in the event of a uniform prior probability *P(a)* and *P(k)*. We assume a Gaussian error model at each point, $f_i(x)$, and use a uniform prior distribution for *a* in log space and a uniform prior distribution of *k* in angular space. Using Bayesian estimation to maximize the posterior probability $P(a,k|f(x),H_{pl})$ as formulated in the equation below, we find the most probable parameters. The best-fit scaling exponents are given in Table 2, and, expectedly, all scaling exponents are negative since the link rate decays as we consider links at further distances.

$$P(a,k \mid f(x),H_{pl}) = \frac{\frac{1}{(2\pi)^{N/2}\sigma^N}\exp\left(\frac{-\sum_i(\log f_i(x) - \log(a) - k\log(x))^2}{\sigma^2}\right)P(a)P(k)}{P(f(x) \mid H_{pl})}$$

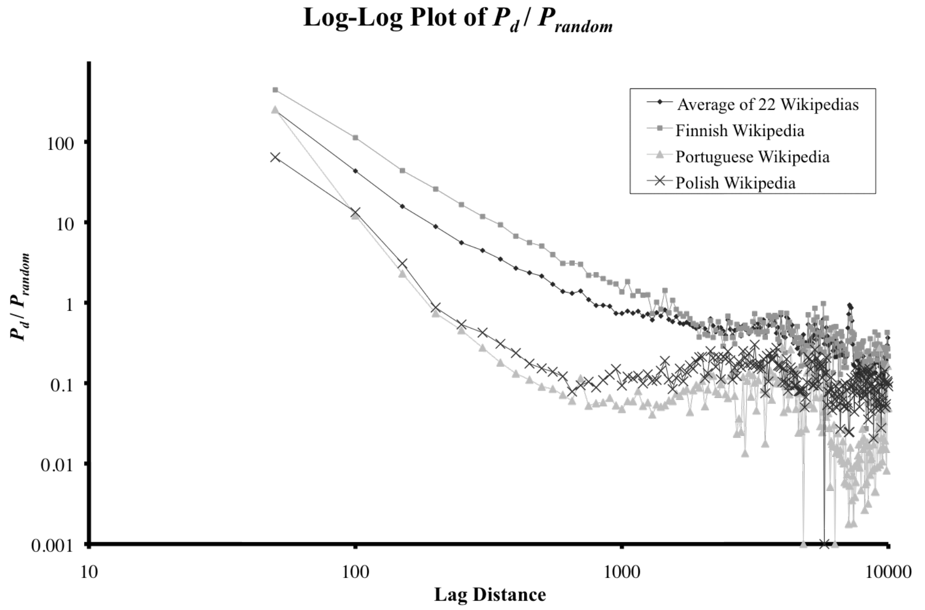

Fig. 3.4 Log-log plots for P_d / P_{random} over lag distance for select Wikipedias. Some Wikipedias fit a power law at small lags better than others. For example, Polish does not very closely follow a power law even at small distances, while Finnish does. Differences in power-law scaling parameter *k* is evident in the varying slopes.

While finding a "best-fit" power law is possible no matter what the underlying distribution of the data, we would like to be able to compare such a fit with other plausible distributions. In particular, we would like to *reject* a uniform distribution, call this hypothesis H_{unif}. A uniform distribution would indicate that the $P_{relation}$ of two points does not change with spatial distance. We can accomplish this easily by finding the odds ratio between the two hypotheses, power-law distributed versus uniformly distributed [22]. This allows us to compare the likelihood that the distribution is drawn from a power law distribution, following the formulation above, versus a uniform distribution, following a formulation *g(x)=c* for some constant *c*. We can specifically compare the two using an odds ratio of the evidence, *P(f(x)|H),* for each model, formulated as:

$$\frac{P(f(x)\,|\,H_{pl})}{P(f(x)\,|\,H_{unif})} = \frac{\int\int P(f(x)\,|\,a,k,H_{pl})\ \mathrm{da\ dk}}{\int P(f(x)\,|\,c,H_{unif})\ \mathrm{dc}}$$

This odds ratio is also given in Table 2, and shows that a power law fits the distribution far better than a uniform assumption. The link rate as we move farther away in geographic space may be reasonably characterized as "failing", or decreasing, as a power law.

Table 2. The scale exponent for the best-fit power law of each Wikipedia over lag classes up to 1000km. Comparison with evidence of a uniform distribution gives an odds ratio that rejects a uniform distribution as compared to a power law.

Language	**Scaling Exponent**	**Odds Ratio** $P(f(x)|H_{pl})/P(f(x)|H_{unif})$
Catalan	-1.6187	2.6971e+055
Czech	-1.3887	6.3573e+057
Danish	-1.7257	1.4333e+067
German	-1.5687	9.8629e+061
English	-2.0467	4.3767e+076
Spanish	-1.6187	6.1007e+058
Finnish	-1.7837	1.6816e+085
French	-1.4747	7.5283e+055
Hungarian	-1.4307	5.2172e+056
Italian	-1.7837	1.0075e+054
Japanese	-1.3487	2.4968e+071
Dutch	-2.2017	5.4446e+058
Norwegian	-1.9077	4.7396e+076
Polish	-1.7257	3.635e+044
Portuguese	-2.1217	2.0573e+059
Romanian	-1.5207	4.6351e+034
Russian	-1.4307	3.0116e+051
Slovakian	-1.1667	5.9496e+041
Swedish	-1.7837	1.1618e+076
Turkish	-1.5207	2.0533e+070
Ukrainian	-1.6187	9.1704e+085
Chinese	-1.3887	1.6234e+075
Average	-1.6187	7.0424e+067

3.5 Network Analogy

Some readers may have recognized the similarity of the unnormalized statistic, P_d, with the *clustering coefficient* statistic frequently used to analyze networks. We have measured the clustering coefficient of a network neighborhood consisting of only edges that exist between binned geographical distances. That is, each data point in Figure 3.2 represents the local clustering coefficient of the neighborhood created by activating only the possible edges that connect nodes at distance *d*=50km, 100km, etc. The clustering coefficient is an important indicator in special types of networks, such as scale-free and small-world. A small-world network is characterized by the necessary conditions of a large average local clustering coefficient and a low path length (number of hops to get from one node to another). While our current analysis lacks analysis of path length, Figure 3.2 shows that, compared to a random network, networks consisting of nodes at small geographical distances have a much higher clustering coefficient than networks representing larger geographical distances. Our work here shows that the literal interpretation of the term *small-world network* may ultimately be provable in the strict geographic sense. After adding an analysis of path length, this dataset may show that networks consisting of nodes at short geographical distances are indeed small-world.

The current formulations lack crucial information to analyze whether they are also *scale-free*, a quality characterized by a power-law distribution in node degree (the number of connections out of a node). Future analysis will incorporate this concept into analysis of spatial Wikipedia.

4 Future Work

The results discussed in this paper have generated numerous further research questions. Most notably, as mentioned, we are actively seeking an explanation for the variation amongst the different languages' relatograms. Our current preliminary hypotheses can be split into two separate but overlapping categories: cultural causes and linguistic causes. On the cultural side, could it be that since the standard activity space of individuals is much smaller than 50km in many of the cultures examined, this causes the 0-50km distance class to have a much lower relation probability in these cultures' Wikipedias? Do certain cultures describe spatial entities in more relational terms, resulting in a higher average probability over large numbers of distance classes? Could we also be seeing differential culturally defined regional cognition effects, as is suggested by [23]? As for the linguistic causes, do certain languages' norms and/or grammatical structures make it more or less difficult to express relations to locations that are closer or further away? Since Wikipedia is a written language corpus and links must occur inline in the corpus, even a slight linguistic proclivity in this area could have a somewhat large effect, relatively speaking. Similarly, given the nature of Wikipedia, does the reference frame used by each language have an effect? Languages that default to a relative reference frame in formal writing will have at least more *opportunities* to encode spatial relations as links than those that use absolute frames. This is simply because contributors to Wikipedia writing in relative frame languages must mention more spatial entities (as opposed to cardinal directions), allowing them the chance to add a links to these entities while they are at it.

We are also working with the hyperlingual Wikipedia dataset to examine another vital and unique aspect of spatial information: scale. For instance, do the WAGs hierarchical structures' mimic urban spatial hierarchies? In other words, can we evaluate central place theory using the hyperlingual Wikipedia? How does this work in "home" countries of languages versus in foreign countries? We are preparing a manuscript repeating this study from a multiple scales perspective, as is discussed in section 2.2.

Also important is to consider more advanced models of relatedness. We have used here a straightforward binary "link existence" approach to avoid the many complications involved with using recently published Wikipedia-based semantic relatedness (SR) measures [7, 12, 18]. However, we are currently working to compare the present results with those from these SR methods. We hope to elucidate how well spatial relations are captured, a very important consideration given that

Wikipedia-based semantic relatedness has quickly become a popular tool in the artificial intelligence community.

Our experiments were focused at a general, theoretical level, but the results do have applied value as a crude quantitative description of spatial relatedness in the absence of more specialized knowledge. While we are by *absolutely* no means recommending that scientists use our "pseudo-covariograms" in place of real covariograms developed on data relevant to their specific research project, at times no such data is available. Take as an example the work of Gillespie, Agnew and colleagues [24], who used a model from biogeography to predict terrorist movement. We assert that our general model of spatial correlation might be more valid in that context than Gillespie et al.'s approach, especially if/when the Arabic and/or Pashto Wikipedias become large enough for our analyses[10]. Future work will involve improving the applied functionality of our methodology even further by including a more sophisticated and/or localized distance function than universal straight-line distance and providing crude relation type information through Wikipedia's category structures.

5 Conclusion

In this paper, we have shown empirically that in the largest attempt to describe world knowledge in human history, the First Law of Geography proves true: nearby spatial entities in this knowledge repository have a much higher probability of having relations than entities that are farther apart, although even entities very far apart still have relations to each other. In other words, we have seen that the very medium that was supposed to oversee the "death of distance" – the Internet – has instead facilitated the reaffirmation of a theory about the importance of distance that is almost 40 years old and that has roots dating back centuries.

Finally, we would also like to reiterate the significance of the fact that TFL proved true in the knowledge repositories constructed by people who speak *twenty-two* different languages. The discussion of what are the universal truths about humanity that span cultural boundaries is a prickly one, but here we have seen at least some evidence that the tendency to see spatial entities as more related to nearer entities than ones that are further away at least deserves mention in that debate.

Acknowledgements

The authors would like to thank Phil Marshall, Ken Shen, Nada Petrovic, Darren Gergle, and Martin Raubal for consulting on the methods and/or idea generation behind this study.

[10] Because of this potential utility, we have made the data from all of our results publicly available at http://www.engr.ucsb.edu/~emoxley/HechtAndMoxleyData.zip. We have also included the algorithms used to perform our experiments in the general WikAPIdia software, which as noted above, is available upon request for academic purposes.

References

1. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Economic Geography **46** (1970) 234-240
2. Longley, P., Goodchild, M., Maguire, D., Rhind, D.: The nature of geographic data. Geographic information systems and science (2005)
3. Fabrikant, S.I., Ruocco, M., Middleton, R., Montello, D.R., Jörgensen, C.: The First Law of Cognitive Geography: Distance and Similarity in Semantic Spaces. GIScience (2002)
4. Montello, D.R., Fabrikant, S.I., Ruocco, M., Middleton, R.: Testing the First Law of Cognitive Geography on Point-Display Spatializations. COSIT '03: Conference on Spatial Information Theory, Kartause Ittingen, Switzerland (2003)
5. Miller, H.J.: Tobler's First Law and Spatial Analysis. Annals of the Association of American Geographers **94** (2004) 284-289
6. Sui, D.Z.: Tobler's First Law of Geography: A Big Idea for a Small World. Annals of the Association of American Geographers **94** (2004) 269-277
7. Hecht, B., Raubal, M.: GeoSR: Geographically explore semantic relations in world knowledge. In: Bernard, L., Friis-Christensen, A., Pundt, H. (eds.): AGILE '08: Eleventh AGILE International Conference on Geographic Information Science, Vol. The European Information Society: Taking Geoinformation Science One Step Further. Springer-Verlag, Girona, Spain (2008) 95 - 114
8. Tobler, W.R.: On the First Law of Geography: A Reply. Annals of the Association of American Geographers **94** (2004) 304-310
9. Goodchild, M.: The Validity and Usefulness of Laws in Geographic Information Science and Geography. Annals of the Association of American Geographers **94** (2004) 300-303
10. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. WWW '06: 15th International Conference on World Wide Web, Edinburgh, Scotland (2006) 585-594
11. Ortega, F., Gonzalez-Barahona, J.M., Robles, G.: The Top-Ten Wikipedias: A Quantative Analysis Using WikiXRay. ICSOFT '07: International Conference on Software and Data Technology (2007) 46 - 53
12. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. IJCAI '07: Twentieth Joint Conference for Artificial Intelligence, Hyberabad, India (2007) 1606-1611
13. Milne, D., Witten, I.H.: Learning to Link with Wikipedia. CIKM '08: ACM 17th Conference on Information and Knowledge Management, Napa Valley, California, United States (2008) 1046-1055
14. Giles, J.: Special Report: Internet encyclopaedias go head to head. Nature **438** (2005) 900-901
15. Halavais, A., Lackaff, D.: An Analysis of Topical Coverage of Wikipedia. Journal of Computer-Mediated Communication **13** (2008) 429-440
16. Pfeil, U., Panayiotis, Z., Ang, C.S.: Cultural Differences in Collaborating Authoring of Wikipedia. Journal of Computer-Mediated Communication **12** (2006)

17. Kittur, A., Chi, E., Pendleton, B.A., Suh, B., Mytkowicz, T.: Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. CHI '07: 25th International Conference on Human Factors in Computing Systems (2007) 1-9
18. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia. WIKI-AI 2008: AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, Chicago, IL (2008)
19. Hecht, B., Gergle, D.: Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. Communities and Technologies 2009: Fourth International Conference on Communities and Technologies, University Park, PA, USA (2009) 10
20. Arbia, G., Benedetti, R., Espa, G.: Effects of the MUAP on image classification. Geographical Systems **1** (1996) 123-141
21. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power law distributions in empirical data. SIAM Review (2009)
22. Goldstein, M.L., Morris, S.A., Yen, G.G.: Fitting to the power-law distribution. The European Physical Journal B - Condensed Matter and Complex Systems **41** (2004) 255-258
23. Xiao, D., Liu, Y.: Study of Cultural Impacts on Location Judgments in Eastern China. In: Winter, S., Duckham, M., Kulik, L., Kupers, B. (eds.): COSIT '07: Conference on Spatial Information Theory. Springer-Verlag, Melbourne, Australia (2007)
24. Gillespie, T.W., Agnew, J.A., Mariano, E., Mossler, S., Jones, N., Braughton, M., Gonzalez, J.: Finding Osama bin Laden: An Application of Biogeographic Theories and Satellite Imagery. MIT International Review (2009)