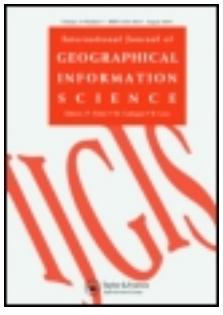

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Volunteered geographic information production as a spatial process

Darren Hardy [a], James Frew [a b] & Michael F. Goodchild [b]

[a] Bren School of Environmental Science & Management, University of California, Santa Barbara, CA, USA

[b] Department of Geography, University of California, Santa Barbara, CA, USA

Available online: 31 Jan 2012

To cite this article: Darren Hardy, James Frew & Michael F. Goodchild (2012): Volunteered geographic information production as a spatial process, International Journal of Geographical Information Science, DOI:10.1080/13658816.2011.629618

To link to this article: http://dx.doi.org/10.1080/13658816.2011.629618

International Journal of Geographical Information Science
i*First*, 2012, 1–22

Taylor & Francis
Taylor & Francis Group

Volunteered geographic information production as a spatial process

Darren Hardy[a]*, James Frew[a,b] and Michael F. Goodchild[b]

[a]*Bren School of Environmental Science & Management, University of California, Santa Barbara, CA, USA;* [b]*Department of Geography, University of California, Santa Barbara, CA, USA*

(*Received 27 May 2010; Revised 22 March 2011; Accepted 23 September 2011*)

Wikipedia is a free encyclopedia that anyone can edit and a popular example of user–generated content that includes volunteered geographic information (VGI). In this article, we present three main contributions: (1) a spatial data model and collection methods to study VGI in systems that may not explicitly support geographic data; (2) quantitative methods for measuring distance between online authors and articles; and (3) empirically calibrated results from a gravity model of the role of distance in VGI production. To model spatial processes of VGI contributors, we use an invariant exponential gravity model based on article and author proximity. We define a proximity metric called a 'signature distance' as a weighted average distance between an article and each of its authors, and we estimate the location of 2.8 million anonymous authors through IP geolocation. Our study collects empirical data directly from 21 language-specific Wikipedia databases, spanning 7 years of contributions (2001–2008) to nearly 1 million geotagged articles. We find empirical evidence that the spatial processes of anonymous contributors fit an exponential distance decay model. Our results are consistent with the prior results on information diffusion as a spatial process, but run counter to theories that a globalized Internet neutralizes distance as a determinant of social behaviors.

Keywords: distance decay; geotagging; user-generated content; volunteered geographic information; Wikipedia

1. Introduction

Volunteered geographic information (VGI) is 'a special case of the more general Web phenomenon of user-generated content' (Goodchild 2007, p. 212). Today, users can share and pool geographic information via the Internet at lower costs, in effect creating a global digital commons of geographic knowledge that is released from traditional mechanisms of production and distribution. Geobrowsers, in particular, use visualization and Internet technologies to share place-based information (Scharl and Tochtermann 2007). The popularity of Google Earth highlights how the geobrowser transforms traditionally heavyweight GIS user interfaces into simple, yet compelling, web browser-like interfaces. Moreover, the utility and growth of geographic user-generated content has led to new efforts in GIScience research (Goodchild 2008), and the production and use of VGI will likely challenge the 'knowledge politics' of spatial data infrastructures (Elwood 2010). For example, VGI may weaken the traditional notions of authoritative sources as the social production of spatial

*Corresponding author. Email: dhardy@bren.ucsb.edu

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2011.629618
http://www.tandfonline.com

data increases (Budhathoki *et al.* 2008). As Sui (2008, p. 4) argues, the 'wikification of GIS is perhaps one of the most exciting, and indeed revolutionary developments since the invention of [GIS] technology in the early 1960s.' In this paper, we discuss the spatial nature of one type of VGI production mechanism, namely 'online collective authorship.' Our main focus is not on whether distance is important for the production of user-generated content about place, as it obviously is, but on exploring exactly how distance affects contributions.

Terms like collective intelligence (O'Reilly 2005), wikinomics (Tapscott and Williams 2006), crowdsourcing (Brabham 2008), online collectivism (Lanier 2006), and peer production (Benkler 2002) refer to the various aspects of this large-scale collective action phenomenon. Hardy (2008) defines new forms of online collective authorship as a 'mass collective effort by individuals to produce information artifacts within a digital commons.' These systems fundamentally recognize the utility of user-generated content, and their common thread is harnessing social behaviors of large online user communities.[1] On the surface, the existence of user-generated content systems seems implausible due to the costs of massive coordination and the seeming lack of extrinsic motivation, and thus, they have been a center of academic debates (e.g., Beer and Burrows 2007, Wang *et al.* 2007, Nov and Kuk 2008, Luo *et al.* 2009, Nov 2009).

1.1. VGI in Wikipedia

Wikipedia is an online collaborative encyclopedia. It is also one of the largest user-generated content sites with 15 million articles from 22 million contributors (Wikimedia 2010), and the sixth most popular Internet site in the United States (Alexa Internet, Inc. 2009). During 2009 alone, Wikipedia had 365 million unique visitors generated 133.6 billion page views (Zachte 2010a). Its impact on the web's content is significant. Fifty-one percent of its site visits come from link-based search engine referrals (Alexa Internet, Inc. 2009). Of those page views that were referred to Wikipedia by external sites, 42% were referred by Google search, maps, and other services, and 8% were made by their *GoogleBot* 'web crawling' software (Zachte 2009).

Through the use of geotagging and web services, Wikipedia integrates its geographic content with web-mapping systems, like Google Earth and OpenStreetMap (Haklay and Weber 2008). Researchers have studied editorial behaviors of Wikipedia contributors but little is known about their socio-spatial behaviors. A basic observation of how Wikipedia production works is that there are (a) two groups of contributors – a small, highly productive set, then everyone else and (b) two groups of articles – a few receiving the majority of contributions, and most receiving a relatively small number of contributions (Voss 2005, Almeida *et al.* 2007). This pattern is commonly referred to as a 'long tail.'[2] The overwhelming majority of contributors do not collaborate with each other in a traditional sense. They do not discuss their contributions with others (e.g. Viégas *et al.* 2007b). Thus, we classify this type of loosely-collaborative production as 'online collective authorship.'

In our study, we found that 8.8% of Wikipedia's articles have VGI content, and these articles are tightly integrated with online mapping services (Kühn *et al.* 2008, Zachte 2010b). Editorial patterns in the production of VGI content are similar to those for nongeographic content. That is, each of the four types of contributors – registered, anonymous, administrative, and bots – exhibits systematic editorial patterns when contributing to geographic articles. For example, using a corpus of place-based Wikipedia articles, Hardy (2008) finds a strong positive correlation between authors and contributions across 21 languages ($R^2 = 0.98$), and an observed power law distribution for anonymous

contributions ($R^2 = 0.90$, $n = 7{,}304{,}171$) and for registered contributions ($R^2 = 0.83$, $n = 11{,}333{,}151$).

1.2. Spatial patterns of contributors

We expect contributors would exhibit systematic *spatial* patterns as well. Despite the advantages of the Internet for collaborative work, contributors are fundamentally engaged in knowledge production processes, which are grounded in social structures and norms, and in turn, physical place. Geographic distance, in particular, should be a significant factor in online VGI production. But, the nature of the Internet in a globalized world has led to debate on whether geographic distance matters (see Cairncross 1997, Goodchild 2004, Friedman 2005, Marston *et al.* 2005, Castells 2010). That is, the Internet may redefine the role of physical place in our lives, due to reduced communication costs and increased ubiquity. Zook (2005, p. 54) summarizes this debate as a new 'geography of electronic spaces,' as the Internet becomes 'a recombinant space for political, cultural, and economic interaction.'

In spatial information theory, an individual's information field is the spatial distribution of the 'knowledge an individual has of the world' (Morrill and Pitts 1967, p. 406), and distance is a dampening factor when modeling socio-spatial behaviors, like the diffusion of innovation, migration, and traffic flows (Hägerstrand 1967, Wilson 1970). In other words, an individual's information field decays as the distance from the individual increases. In quantitative geography, gravity models formalize spatial interaction analysis using this type of distance decay function (Fotheringham and O'Kelly 1989, Sen and Smith 1995). Thus, we expect that when contributors choose to write about a place in Wikipedia, their mean information fields exhibit exponential distance decay effects as found in other socio-spatial phenomena, such as Wilson (1970, 2010) who provided theoretical justification for the exponential nature. Moreover, as contributors write more articles, they expand the overall spatial coverage of Wikipedia articles. As such, we hypothesize that (a) contributors write articles about nearby places more often than distant ones and that (b) this likelihood follows an exponential distance decay function.

2. Methodology

In this section, we discuss our data modeling, extraction, and sampling methods. Then, we introduce gravity models of spatial behavior. Finally, we discuss our model for VGI production and its metrics in our study (Hardy 2010).

2.1. Data

Wikipedia manages hundreds of individual language-specific databases across three data centers in the United States, the Netherlands, and South Korea. Their services use open-source *MediaWiki* software and data models (MediaWiki 2006). Rather than collect data via web-mining methods, we collect our data directly via SQL from near real-time replicas of Wikipedia databases, provided by Wikimedia Deutschland's *Toolserver.org* (Figure 1).

Contributors write articles using *Wikitext*, a loosely structured markup language, and they embed semi-structured metadata *within* the article. The nondeterministic nature of Wikitext's grammar and conventions causes problems for structured data extraction (see Sauer *et al.* 2007). Wikipedia does not support geographic content natively, and the *WP:GEO* project in Wikipedia[3] governs an extension to Wikipedia's base infrastructure

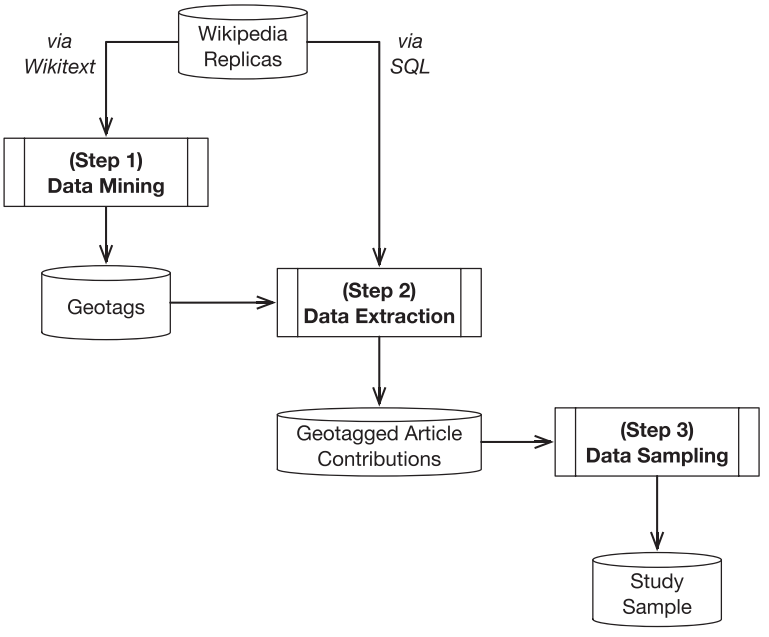

Figure 1. Our data collection process uses *Toolserver* replica databases for direct SQL-based data access.

for adding geographic information to articles. They provide an array of 'wiki templates' that have a semistructured syntax for embedding geographic coordinates. For example, the *coord* template describes a single latitude and longitude, and the *Infobox* template describes an entity with a variety of structured metadata, including latitude and longitude. The templates require that the author first provide location using latitude and longitude. Many authors obtain these coordinates using a web mapping tool like Google Maps. More commonly, they are provided by Wikipedia *bots*, semiautomated programs collect coordinates from gazetteers.

We therefore use data-mining processes from Kühn *et al.* (2008) to extract article geotags.[4] Their software targets a predetermined set of 21 languages[5] and extracted 257,399 unique geographic coordinates from 1,634,264 articles. Our study focuses on the subset of Wikipedia articles that are *about* a particular place, and all our study's Wikipedia articles have geotags with specific geographic locations. As of 23 February 2010, Wikipedia has 15 million articles in 272 languages with 860 million edits, the majority of which is from the 21 languages in our study – 11.3 million articles with 477 million edits (Wikimedia 2010, Zachte 2010b). We limit our data extraction to records associated with the subset of articles available in the geotag data-mining results. We store these records in modified *page* ($n = 997{,}756$), *revision* ($n = 39{,}674{,}239$), *text* ($n = 997{,}756$, most recent version only), and *user* ($n = 712{,}421$) tables. Toolserver provides direct SQL access to replica databases through which we extract our primary source data for these tables. Our modification to the schema includes a language code in each primary key, thereby retaining language distinction within a single table.

Our study's data model provides a single, unified data model spanning these federated databases (Figure 2). It contains a multilingual abstraction layer to Wikipedia articles,

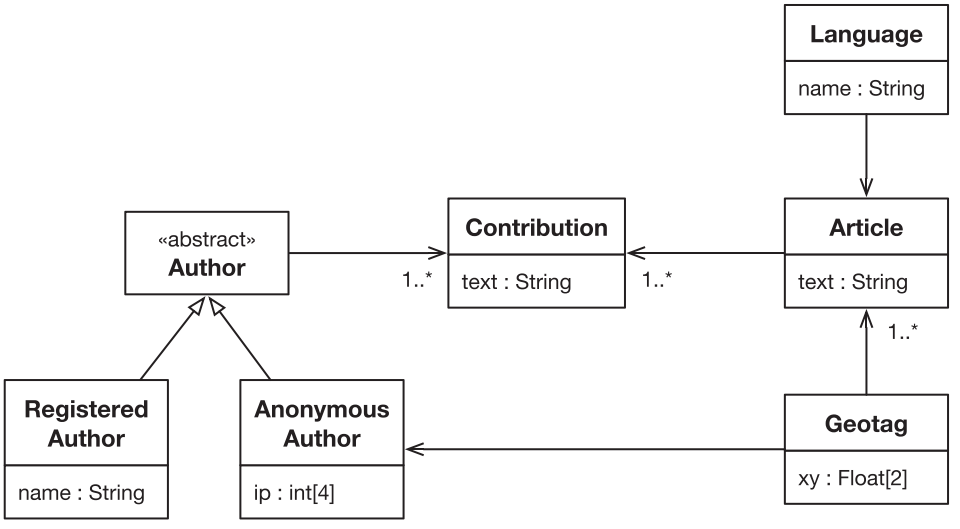

Figure 2. Our data model in UML syntax. In our study, each article has one or more contributions from at least one anonymous author, and all articles and anonymous authors have a single geotag.

authors, and their contributions. Thus, we collect article contribution history data over 7 years (2001–2008) from 21 language-specific databases into a single database. Our data model also provides for spatial analysis, with geographic coordinates of articles and contributors as first-class entities.

To obtain the geographic coordinates of contributors we use *IP geolocation*. We can access IP addresses for anonymous, but not registered, contributors.[6] Location-based services have driven the development of methods to convert IP addresses into geographic coordinates (e.g., Stanger 2008) and to evaluate positional accuracy (e.g., Gueye *et al.* 2007). IP geolocation methods vary and include (a) early attempts to voluntarily publish coordinates in the Domain Name System (DNS) registry (Davis *et al.* 1996), (b) IP address lookups in the Internet registry databases (e.g., *whois*), (c) statistical and probabilistic methods that exploit network topology to estimate geographic distance, such as network delay from established 'landmark' hosts (Gueye *et al.* 2006, Youn *et al.* 2009), (d) methods that exploit 'leaked' user data to improve positional accuracy (Muir and van Oorschot 2009), and (e) various proprietary methods from Akamai, MaxMind, and other companies. Adoption of IP geolocation methods is not widespread nor uniform, and standardized geolocation interfaces are still in development despite significant demand for location-based services. As such, we use the GeoLite City database from MaxMind (2009, 1 July release), which uses proprietary methods and data to convert IP addresses into geographic coordinates.[7]

2.2. Sampling

Our methodology requires locations for both origin (author) and destination (article). We subsample all available data to ensure that our study dataset satisfies these methodological requirements; namely, our geolocation method for anonymous authors requires that articles have at least one anonymous contribution, and our signature distance method (Section 2.3) requires that articles have one and only one geotag. After sampling, our study data include

Table 1. Number of articles, authors, and contributions in our study sample, ranked by article count per language.

		Articles		Authors		Contributions	
1.	English	114,938	58.0%	1,249,835	80.2%	3,244,952	32.4%
2.	German	76,886	74.3%	549,055	88.6%	1,288,200	27.0%
3.	French	48,203	61.5%	214,042	85.9%	553,384	20.4%
4.	Italian	26,465	33.1%	111,208	85.6%	318,113	23.2%
5.	Spanish	23,057	53.5%	178,657	86.3%	462,231	30.3%
6.	Dutch	22,813	17.9%	68,763	82.3%	161,619	15.7%
7.	Polish	20,314	28.9%	75,547	86.0%	206,303	24.6%
8.	Portuguese	17,424	22.4%	78,919	86.9%	175,751	24.4%
9.	Russian	12,141	31.6%	40,072	81.5%	100,287	19.5%
10.	Swedish	12,005	55.6%	30,317	85.1%	75,593	19.2%
11.	Japanese	11,683	51.7%	104,961	86.0%	339,418	39.9%
12.	Czech	7,494	49.3%	14,363	81.0%	34,296	14.3%
13.	Esperanto	7,194	25.0%	8,500	87.8%	22,693	12.8%
14.	Finnish	7,156	61.2%	25,191	82.3%	70,816	21.4%
15.	Norwegian	6,383	34.2%	18,920	79.4%	40,932	15.2%
16.	Catalan	5,986	49.6%	11,472	85.2%	30,411	13.6%
17.	Chinese	5,225	42.1%	23,231	72.6%	60,793	17.6%
18.	Turkish	4,170	55.6%	24,276	86.0%	47,626	27.3%
19.	Danish	3,835	59.9%	11,564	82.0%	27,974	17.1%
20.	Slovak	3,736	25.6%	7,904	83.0%	20,094	16.8%
21.	Icelandic	969	58.4%	1,720	81.9%	3,651	8.9%
	Total	438,077	44.2%	2,848,517	83.4%	7,285,137	27.2%

Note: The percentage shows the how much of the available data qualified for our study sample (e.g., articles with a single geotag and at least one anonymous contribution).

438,077 articles with 7,285,137 contributions from 2,848,517 anonymous contributors (Table 1). These data exhibit structural properties consistent with prior findings (Hardy 2008), such as the growth across languages fits a power law distribution for contributions ($R^2 = 0.92$), authors ($R^2 = 0.87$), and articles ($R^2 = 0.73$).

For author location via IP geolocation, MaxMind's claimed accuracy of the GeoLite City database varies by country, and in terms of the percentage of IP addresses 'correctly resolved within 25 miles of true location,' is the United States (79%), Germany (71%), France (60%), Australia (59%), Japan (54%), and the United Kingdom (54%). Although we cannot substantiate MaxMind's accuracy claims,[8] they are within the resolution of our analysis (≈ 5085 km$^2 \gg 122$ km^2).

When calculating geodesic distance measurements and aggregate statistics in our study, we first round latitude and longitude to the nearest 0.1° to account for imprecision in geotagging and geolocation methods. For geodesic distance, we use a great circle distance method that assumes a spherical earth where 1 arc-min equals 1 nautical mile (1.852 km). For aggregate statistics, we also round geodesic distances to the nearest 10 km. Our sample data have 103,291 and 30,376 unique coordinate locations for articles and anonymous contributors, respectively (Figures 3 and 4). The contribution locations span up to 2.75 million km^2, about 36% of the estimated global land area of human settlement (Elvidge *et al.* 2010). The most active contribution locations ($n = 2{,}652$), which contain 80 percent of all contributions, span up to 0.23 million km^2, only 8.5% of the area from all locations.

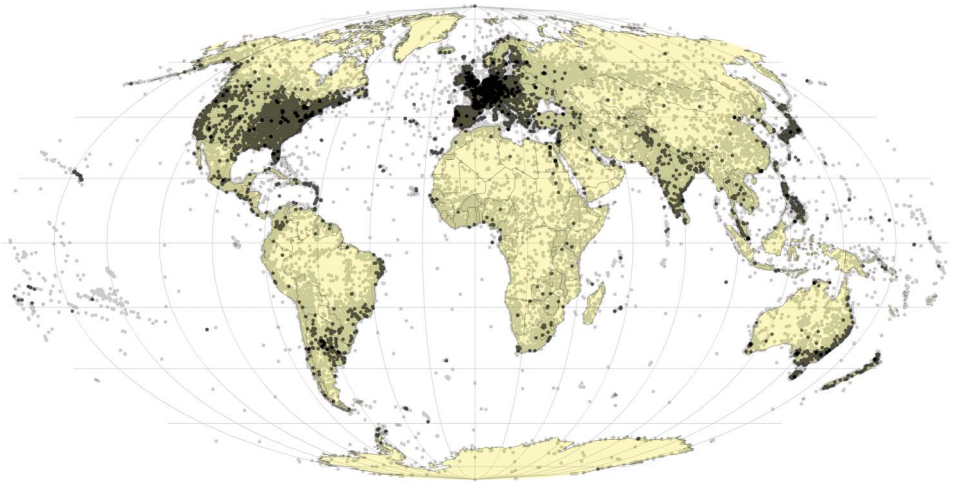

Figure 3. Map of Wikipedia *geotagged article locations* in study with log-scale brightness for density of mean number of articles per language (Projection: Mollweide). The distribution ranked by continent for articles across all languages is Europe (60.6%), North America (13.3%), Asia (9.7%), South America (3.1%), Africa (3.1%), Australia (1.0%), and Antarctica (0.1%).

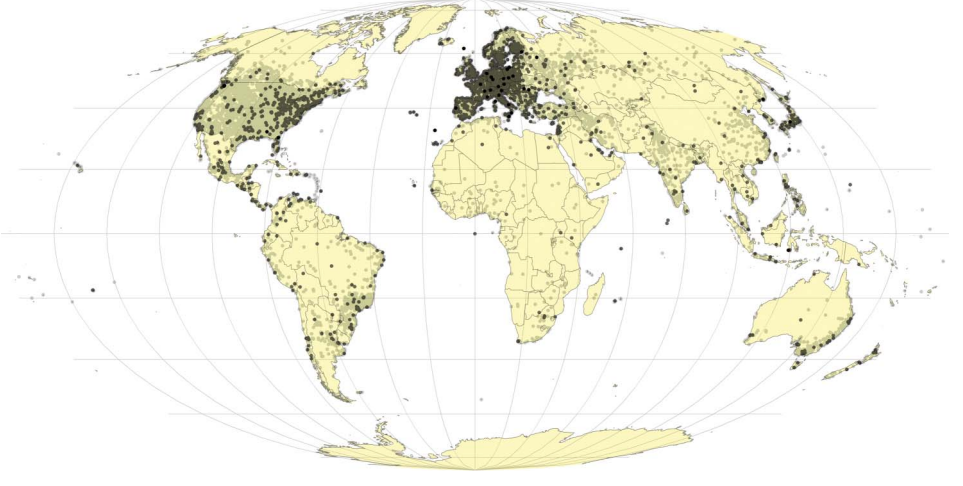

Figure 4. Map of Wikipedia *anonymous author locations* in study with log-scale brightness for the density of mean number of contributions per author per language (Projection: Mollweide). The distribution ranked by continent for contributions by anonymous authors across all languages is Europe (52.9%), North America (27.5%), Asia (9.5%), South America (4.1%), Australia (1.8%), and Africa (0.4%).

2.3. Measurement

In this section, we discuss our proximity metric, the 'signature distance' between authors and articles. Briefly, this metric is the average distance between an article and its authors, weighted by the relative number of contributions from each author.

First, we define the basic properties of an author and an article. Let ρ be an author; α be an article; θ be a geotag with a single (x, y) geographic coordinate; and $\theta(\rho)$ and

$\theta(\alpha)$ be the geotags associated with author ρ and article α, respectively. Second, we define the basic properties of contributions from an author to an article. Let S be our sample set; $P = \{\rho : \rho \in \mathrm{S}\}$ be the sample set of authors; $A = \{\alpha : \alpha \in \mathrm{S}\}$ be the sample set of articles; $\eta(\rho,\alpha)$ be the contribution(s) by an author ρ to an article α; $N(\alpha) = \{\eta(\rho,\alpha) : \rho \in P\}$ be the contribution(s) to article α; $P(\alpha) = \{\rho : \rho \in N(\alpha)\}$ be the author(s) who make contribution(s) to article α; and $N = \{N(\alpha) : \forall\alpha \in A\}$ be all contributions in the sample set. To construct our sample set S, we assert that every article $\alpha \in A$ has one and only one geotag $\theta(\alpha)$ and has at least one author $\rho \in P(\alpha)$ who is anonymous. Finally, let $\delta(\rho,\alpha) = |\theta(\alpha) - \theta(\rho)|$ be the shortest geodesic distance between article α and author ρ.

We define our *signature distance* metric $D(\alpha)$ as the weighted average of distances between an article and its authors. We define work $w(\rho,\alpha) = |\eta(\rho,\alpha)|/|N(\alpha)|$ as the relative frequency of the number of contributions between author ρ and all contributing authors $P(\alpha)$ for article α,[9] and the signature distance $D(\alpha)$ as the average distance between article α and all its contributing authors $\rho \in P(\alpha)$, weighted by the relative work per author $w(\rho,\alpha)$:

$$D(\alpha) = \sum_{\forall\rho\in P(\alpha)} (w(\rho,\alpha)\cdot\delta(\rho,\alpha)) \tag{1}$$

To illustrate, consider an example where two authors, one in San Francisco and another in Los Angeles, made contributions to an article about Santa Barbara. The arithmetic mean distance would be the mean distance between San Francisco and Santa Barbara ($\approx$ 510 km) and Los Angeles and Santa Barbara ($\approx$ 180 km): $\bar{d} = (510 + 180)/2 = 345$ km. But the signature distance would be weighted by their relative contributions, say two from San Francisco and five from Los Angeles: $D(\alpha) = (510 \times 2 + 180 \times 5)/(2 + 5) = 274$ km, resulting in a metric closer to the more active contributor.

2.4. Model

Spatial interaction models, which pertain to flows (interactions) between two or more geographic regions, have decades-long history in geography dating back to 'social physics' in the early twentieth century (Wilson 1969, 1970, 1971, Fotheringham 1981). In regional geography and related disciplines, they form the basis of 'important and useful social theories' (Haynes and Fotheringham 1984 p. 10). Distance decay or 'gravity' models are one type of spatial interaction model, which use 'mass' functions to deal with scale and distance effects. Equation (2) is a general gravity model (Sen and Smith 1995 p. 3), where T_{ij} is the interaction between population centers i and j; $A(i)$ and $B(j)$ are unspecified origin and destination weight (mass) functions; d_{ij} is the spatial factor or distance between regions i and j; and $F(d_{ij})$ is an unspecified distance decay function, for which Equations (3), (4), and (5) are common ones: a power, exponential, and gamma (or combined) distance decay function, respectively (Sen and Smith 1995, pp. 93–99).

$$T_{ij} = A(i) \times B(j) \times F(d_{ij}) \tag{2}$$

$$\text{Power}: \quad F(d_{ij}) = d_{ij}^{-\beta} \tag{3}$$

$$\text{Exponential}: \quad F(d_{ij}) = \exp(-\beta d_{ij}) \tag{4}$$

$$\text{Gamma}: \quad F(d_{ij}) = \exp(-\gamma d_{ij}) d_{ij}^{-\beta} \tag{5}$$

To model VGI production as a spatial process, we define a probabilistic model where the dependent variable is a likelihood for interaction, based on a spatial factor. Specifically, we use a *probabilistic invariant exponential gravity model* (Sen and Smith 1995 p. 102). In terms of Equation (2), T_{ij} is converted to the probability of an interaction based on a spatial factor. The mass terms $A(i)$ and $B(j)$ are combined into a single invariant constant K to allow for uneven distributions of authors and articles over the Earth's surface. Finally, $F(d_{ij})$ is the exponential distance decay function from Equation (4). Thus, in terms of our article signature distance metric $D(\alpha)$, our model in Equation (6) uses the probability $\Pr(D(\alpha) = d)$ as the likelihood that a given article α has a signature distance $D(\alpha)$ equal to a distance d within a range of $d' \pm \varepsilon$, and K and β are empirically derived constants.

$$\Pr\left(D(\alpha) = d\right) = K \exp\left(-\beta d\right), \quad \text{where } d = d' \pm \varepsilon \tag{6}$$

3. Results

In this section, we discuss how our results provide empirical evidence that anonymous contributors contribute to nearby articles more than distant ones and that an exponential distance decay model fits their behavior.

3.1. Distribution of signature distance

Using Equation (1), we compute signature distance $D(\alpha)$ for each article $\alpha \in A$ ($n = 438{,}077$). Our signature distance metric exhibits a strong linear correlation with the (unweighted) arithmetic mean distance ($R^2 = 0.9830$; $\beta = 0.9826$; $p \ll 0.01$), so it does introduce a slight bias, as expected, toward authors who are closer to an article than the arithmetic mean distance. Using Equation (6), we compute $\Pr(D(\alpha) = d)$ as the relative frequency of observed $D(\alpha)$ values at a given distance d, where $\epsilon = 5$ km. We find that $\Pr(D(\alpha) = d)$ is inversely related to distance d for English articles (Figure 5). The distribution of signature distances has 63% of articles with $D(\alpha) \leq 2000$ km.

To test our hypotheses against this distribution, we define them in terms of Equation (6):

$$H_1 : \Pr(D(\alpha) = d_1) \leq \Pr(D(\alpha) = d_2), \quad \text{where } d_1 > d_2 \geq 0 \tag{7a}$$

$$H_2 : \Pr(D(\alpha) = d) = K \exp\left(-\beta d\right), \quad \text{where } \beta > 0 \tag{7b}$$

$\Pr(D(\alpha) = d)$ is weakly decreasing as the distance d increases, and it is monotonically decreasing when the aggregation of d is larger than 10 km, thus supporting H_1. To test H_1 further, we conduct a simulation in which each author's location is chosen at random from all locations within the authorship population. That is, we rewrite Equation (1) to select author location $\theta(\rho)$ at random from the set of all author locations $\{\theta(\rho) : \forall \rho \in P\}$. For the simulation, we use the mean value across 100 randomized iterations to compute $D(\alpha)$ for all articles ($n = 438{,}077$), and for computational simplification, we use $d = d' \pm 50$ km rather than $d = d' \pm 5$ km as in the observed data. When compared to our empirical data, our nongeographic simulation predicts too many distant articles ($D(\alpha) > 3000$ km) and too few nearby articles ($D(\alpha) \leq 3000$ km) by 0.5535, thus supporting H_1. If there were no geographic effects in authorship, the observed values would better match the simulation, such as in the nearby ranges where the simulation predicts very low activity.

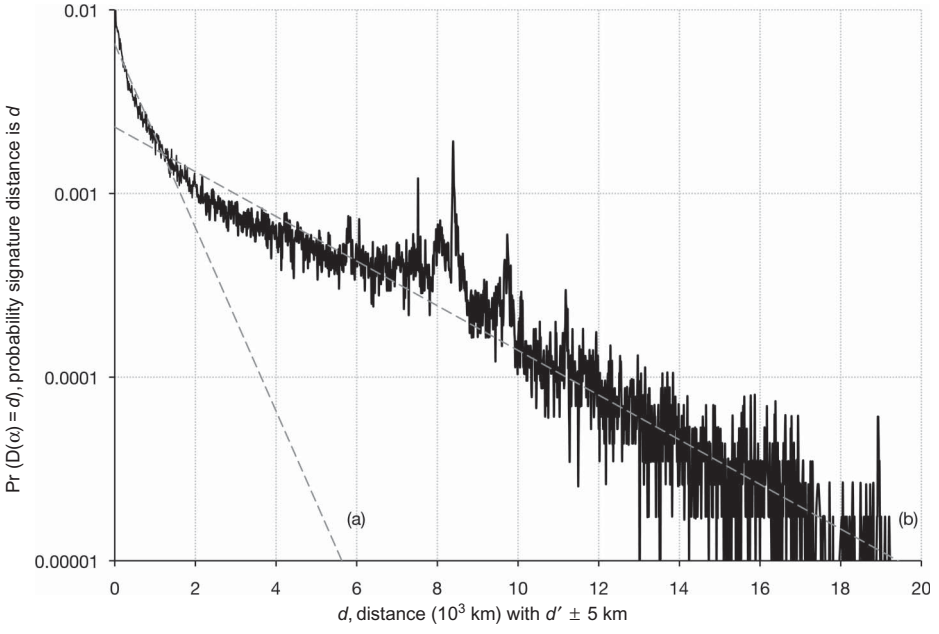

Figure 5. Semi-log plot of $\Pr(D(\alpha) = d)$ versus distance d, where $d = d' \pm 5$ km for observed data in English. Lines (a) and (b) show the regression results for Equation (8) using $\tau = 1500$ km for the head ($n = 55{,}025$; $R^2 = 0.9183$; $K_1 = 0.0065$; $\beta_1 = 1.1498$) and tail ($n = 59{,}913$; $R^2 = 0.9052$; $K^2 = 0.0023$; $\beta_2 = 0.2793$), respectively.

3.2. Model fit

To fit the model in Equation (6) to our study data, we use an ordinary least squares regression method with a logarithmic transformation to a linear model. We convert the units of $D(\alpha)$ from km to 10^3 km and use observed relative frequency for $\Pr(D(\alpha) = d)$. We ran this model for each language (Table 2a) and for a composite of all languages ($n = 438{,}077$; $R^2 = 0.9005$; $p \ll 0.01$; $f = 17{,}480$; $DF = 1{,}930$). The exponential distance decay function in Equation (4) fits best compared with the other distance decay functions. Using a power distance decay function in Equation (3), we find a poorer fit ($R^2 = 0.6383$; $p \ll 0.01$; $f = 3409$; $DF = 1930$). Using a gamma distance decay function in Equation (5), we find no significant improvement in fit ($R^2 = 0.9041$; $p \ll 0.01$; $f = 9102$; $DF = 1929$) despite the added complexity. Thus, we have evidence to support H_2.

4. Discussion

In this section, we first discuss the limitations of an exponential model fit and the methods used to geotag articles and assess the quality of those geotags. Next, we discuss the limitations to geolocating authors, and normalizing their contributions by population density. Finally, we discuss related research on Wikipedia, geotagging, and spatiobehavioral patterns.

4.1. Limitations of exponential model fit

De Vries *et al.* (2009, p. 461) noted that literature assumes an exponential or a power function for distance decay effects[10]. They discuss a power spline function as a more flexible

Table 2. Model fit for each language (in rank order by article count).

		(a) Continuous model				(b) Noncontinuous model						
						$d \leq \tau$			$d > \tau$			
		R^2	K	$-\beta$	τ (km)	R^2	K_1	$-\beta_1$	R^2	K_2	$-\beta_2$	ΔR^2
1.	English	0.92	0.0026	0.29	1500	0.92	0.0065	1.15	0.91	0.0023	0.28	−0.01
2.	German	0.79	0.0014	0.26	2500	0.94	0.0098	1.42	0.73	0.0008	0.22	0.04
3.	French	0.77	0.0012	0.26	1500	0.91	0.0154	2.22	0.77	0.0007	0.21	0.07
4.	Italian	0.63	0.0012	0.23	3000	0.93	0.0098	1.24	0.43	0.0004	0.14	0.05
5.	Spanish	0.72	0.0013	0.19	2000	0.91	0.0084	1.35	0.66	0.0008	0.15	0.07
6.	Dutch	0.59	0.0010	0.19	2000	0.89	0.0100	1.64	0.47	0.0005	0.13	0.09
7.	Polish	0.66	0.0014	0.21	3000	0.91	0.0069	1.03	0.49	0.0008	0.15	0.04
8.	Portuguese	0.64	0.0013	0.18	1000	0.87	0.0088	1.95	0.58	0.0010	0.15	0.09
9.	Russian	0.57	0.0016	0.21	6000	0.77	0.0031	0.46	0.28	0.0013	0.18	−0.04
10.	Swedish	0.57	0.0015	0.18	4000	0.80	0.0043	0.73	0.41	0.0013	0.16	0.04
11.	Japanese	0.35	0.0010	0.12	2500	0.77	0.0038	1.13	0.27	0.0009	0.11	0.17
12.	Czech	0.37	0.0009	0.15	500	0.93	0.0392	6.46	0.35	0.0006	0.11	0.27
13.	Esperanto	0.44	0.0013	0.18	1500	0.76	0.0091	1.62	0.26	0.0007	0.11	0.07
14.	Finnish	0.54	0.0013	0.15	500	0.78	0.0107	3.19	0.51	0.0011	0.13	0.11
15.	Norwegian	0.50	0.0013	0.15	3000	0.79	0.0056	0.87	0.21	0.0006	0.08	0.00
16.	Catalan	0.44	0.0011	0.15	1500	0.85	0.0115	1.69	0.28	0.0005	0.08	0.13
17.	Chinese	0.29	0.0010	0.09	500	0.59	0.0058	3.90	0.24	0.0009	0.08	0.13
18.	Turkish	0.50	0.0016	0.16	7500	0.58	0.0030	0.37	0.10	0.0005	0.05	−0.16
19.	Danish	0.39	0.0014	0.13	4500	0.69	0.0042	0.63	0.10	0.0007	0.05	0.00
20.	Slovak	0.28	0.0010	0.12	1000	0.83	0.0216	3.46	0.13	0.0005	0.04	0.20
21.	Icelandic	0.23	0.0021	0.06	500	0.29	0.0050	2.29	0.19	0.0019	0.05	0.01
	All languages	0.90	0.0022	0.28	1000	0.94	0.0109	1.90	0.89	0.0018	0.27	0.02

Note: The model fit data shown in the table were obtained using (a) an exponential distance decay model from Equation (6), where each result has an adjusted R^2, K and $-\beta$ coefficients; and (b) a non-continuous exponential distance decay model from Equation (8), where the threshold distance τ is for the best fit, and ΔR^2 shows the mean improvement in fit.

functional specification for distance decay. In our data, we also see a likely S-curve distribution as the probability distribution does have a shift in steepness when the distance $d \approx$ 2000 km, and has a spike when the distance d is 7000–10,000 km. For the latter, we suspect that this spike is due to population and interest concentrations in Europe and the United States. The distance between these countries varies in a similar range (e.g., roughly from New York–London at $\approx$ 5600 km to Los Angeles–Berlin at $\approx$ 9400 km).

For the former, we construct a noncontinuous model in Equation (8) based on a threshold τ, that is, a simple dual-point power spline. We ran this model (Table 2b) for each language and for a composite of all languages. The latter fits the model with $K_1 = 0.0081$ and $\beta_1 = 1.2727$ ($n = 275{,}619$; $R^2 = 0.9416$) in the front end, and with $K_2 = 0.0018$ and $\beta_2 = 0.2693$ ($n = 162{,}459$; $R^2 = 0.8785$) in the long tail. The front end has a steeper and tighter fit than the long tail, further supporting H_2. All but three of the languages show improvements in the mean fit between the front end and long tail, although the tail tends to fit more poorly than the stronger fit in the front end. The variance in quality of fit does not correlate with number of articles nor with the quality of fit in the continuous model. This may be due to some language-specific idiosyncratic behaviors, as some degree of behavioral differences in cross-lingual, cross-cultural online populations is expected (e.g., Danet and Herring 2007).

$$\Pr(D(\alpha) = d) = \begin{cases} K_1 \exp(-\beta_1 d), & \text{if } d \leq \tau \\ K_2 \exp(-\beta_2 d), & \text{otherwise} \end{cases} \tag{8}$$

4.2. Limitations of geotagging

The vast majority of geotagging is reportedly done by a variety of Wikibots (Kühn *et al.* 2008), and their ad hoc nature ultimately makes it more difficult to extract geotags from articles. For example, a semiautomated Wikibot (*Anomebot2*) runs periodically to geotag articles or mark those that *may* need a geotag. It cross-references named entities in article titles with online gazetteer services.[11] These bots provide a structural mechanism to integrate existing geographic data sources into articles. But they are not semantic in nature, nor do they generate standardized markup.[12] In fact, they increase the complexity of extracting structured geographic information from articles, because of their chaotic, ad hoc nature and that of *Wikitext* markup and templates themselves. The end result is geotag extraction requires ad-hoc or data-mining approaches to deal with the nondeterministic, semistructured nature of article templates and ad-hoc inclusion of geotags.

There are two dimensions to assess geotag quality of articles – *actual* and *asserted*. That is, either the subject of an article is geographic in nature or not (actual), and either the article has a geotag or not (asserted). Thus, we can describe geotag quality as one of four categories:

(1) The article has a geotag and the subject of the article is geographic in nature. The geotag is valid, although it may have errors in its positional accuracy.
(2) The article has a geotag, but the subject of the article is not geographic in nature. The geotag would thus represent a *false-positive*.
(3) The article does not have a geotag, but the subject of the article is geographic in nature. The absence of a geotag would thus represent a *false-negative*.
(4) The article does not have a geotag, and the subject of the article is not geographic in nature. The absence of a geotag is valid.

Popular articles might avoid the invalid types more easily, but articles in the long tail have no such visibility. In the false-negative case, a workflow may suggest that authors add a geotag (see Cosley *et al.* 2007), and the Wikibot *Anomebot2* does mark articles for geotagging. Automated methods for detecting the geographic nature of an article are difficult and largely rely on matching named entities in a gazetteer.

4.3. Limitations of population data

Our methodology to estimate author location limits our analysis to a subset of available data. We include about one-fourth of the available contributions to geotagged content by restricting to anonymous authors and to articles which already contain a geotag. For authors, the quality dimensions are dependent on author type. Wikipedia does not provide a structure for geographic location for registered authors, and thus their location is unknown. For anonymous authors, however, Wikipedia provides an IP address for which we may estimate geographic location. This structure provides for a paradoxical effect that we know more about the geographic location of anonymous authors than registered authors.[13]

Anonymity in general is thought to have implications on online behavior (Weicher 2006), and we would expect behavioral differences between anonymous and registered contributors in Wikipedia. For example, Viégas *et al.* (2004) found that the proportion of anonymous contributions vary considerably over time, but also found, counterintuitively, that no clear connection exists between anonymity and vandalism in 'edit wars.' Recent group-focused studies of Wikipedia analyze contributor patterns largely along group delineations based on a number of contributions (Priedhorsky *et al.* 2007, Panciera *et al.* 2009), rather than anonymity. For example, the number of contributions metric is a good proxy for anonymity as only 0.08% of anonymous authors have more than 100 contributions, but it also dramatically limits the registered population since only 3.98% of registered authors have more than 100 contributions.

Obtaining geographic information for registered contributors via survey methods is a difficult proposition. Although researchers have used direct survey methods to solicit information from registered contributors, their sample sizes have been small. Wikipedia does not require contact information for registration, so larger sample sizes would require a coordinated outreach effort. For example, Nov (2007) had a sample size of 151 contributors for his survey on motivational factors, but the sample population was limited to 2847 contributors who were listed in the 'Alphabetical List of Wikipedians' article and available via direct email, presumably from listing their full name or email address.[14] An alternative approach to locate registered contributors might be based on a spatial footprint method by Lieberman and Lin (2009) who calculate a contributor's spatial footprint based on their contribution history to geotagged articles. A combination approach of surveying a small of registered contributors for their location and then using that data to calibrate an automated spatial footprint method might be yield sufficient results to include registered contributors in our signature distance method. In future work, however, the adoption of location-based technologies in Web-authoring platforms, such as GPS-enabled smart phones and handheld devices, may fundamentally alter this dynamic as real-time author locations could be integrated into contribution processes.

Due to a lack of available data, we have not been able to normalize the contributor population distributions as appropriate. The spatial distribution of the global Internet population is not available at a sufficient resolution. Although researchers have studied online literacy and divides within Internet populations (e.g., Zook 2005), spatial data of those phenomena are at a regional level. There are higher resolution general population estimations,

such as Balk and Yetman (2004), but Internet adoption and accessibility rates are spatially variant (Billón *et al.* 2008).

4.4. Related work

4.4.1. VGI

In his 1981 book *Critical Path*, Buckminster Fuller described *Geoscope*, a vision of a computer-aided model of the Earth (cited in Foresman 2008). Today, digital earth systems have become 'comprehensive, distributed geographic information and knowledge organization' systems (Grossner *et al.* 2008, p. 145), including Google Earth, Microsoft's TerraServer, NASA Digital Earth Testbed, and Alexandria Digital Library (Barclay *et al.* 2000, de la Beaujardire *et al.* 2000, Frew *et al.* 2000, Butler 2006). Their success is driving increased interest in digital earth research, often technological in nature (Goodchild 1999, 2000, Yongxiang 1999, Maguire 2007, Foresman 2008, Grossner *et al.* 2008). Some of this area's most problematic issues, however, are institutional in nature (Goodchild 2008), and research thus far has paid too little attention to sociological or institutional behaviors that are critical to understand how we use technology in our lives.

In particular, these earlier visions have not anticipated adequately the immense utility of user-generated content and VGI. For example, VGI is increasingly moving into the mobile domain where users leave (often implicitly) digital traces[15] conducive to geolocation methods, such as GPS-enabled smart phones, cellphone tower records, or even georeferenced photos (Girardin *et al.* 2008, Gonzalez *et al.* 2008). VGI and the related phenomena of *neogeography* – where 'people use Web 2.0 techniques to create and overlay their own locational and related data on and into systems that mirror the real world' (Hudson-Smith *et al.* 2009, p. 119) – may also expand the notion of the public from prior work in public participation GIS (PPGIS) to include much larger, distributed civic participation (Sieber 2006, Elwood 2008, Sui 2008, Hall *et al.* 2010). The notion of collective action through new media is at the core of these phenomena, and we encourage further research on behavioral and cultural aspects of VGI production and use.

4.4.2. Wikipedia

Wikipedia is one example within an ecosystem of user-generated content systems on the web (O'Reilly 2005), where open content licenses have challenged intellectual property frameworks for digital content distribution (Lessig 2001). Research on Wikipedia itself fits into three categories of study: (a) production methods and processes (e.g., Pfeil *et al.* 2006, Viégas *et al.* 2007a); (b) use of its articles in lieu of traditional encyclopedic sources (e.g., Lih 2004, Fallis 2008); and (c) use of its content as input to related research, such as semantic analysis (e.g., Chernov *et al.* 2006, Strube and Ponzetto 2006, Völkel *et al.* 2006). For production methods, Wikipedia's success has challenged academic theories of production. Benkler (2002) argued, for example, that in terms of economic models of production, when the efficiency gains of peering exceeds the costs of organizing human capital into a firm or market, then a *commons-based peer production* system will emerge. Its advantage is based not only on the reduced costs of human capital and communications, but also on the nonrival aspects of web-based information artifacts. This effectively eliminates allocation costs (i.e., many people can read a web page simultaneously without degrading its value) and increases the pool of potential contributors, which reduces effects from free-riders.

The most active segments of the contributor population are 91,817 contributors with at least 5 contributions *per month* and 1,076,908 contributors with at least 10 contributions

total (Zachte 2010b). The 'long tail' has 21.1 million contributors, each of whom have less than 10 contributions total. Contributors are one of the four types: registered, anonymous, bots, and administrative. Registered contributors create an account on Wikipedia, and any contributions are explicitly tagged as provided by that author's account. Anonymous contributors do not provide any registration information, and their computer's IP address is used in lieu of an account. 'Bots' and other administrative contributors are both special cases of registered authors who have additional access or permissions to articles. Both contributors and Wikipedia's readership follow power law distributions (Voss 2005, Almeida *et al.* 2007, Priedhorsky *et al.* 2007). Moreover, when Priedhorsky *et al.* (2007) considered quality metrics for contributions, such as longevity or visibility, they found that a small group of elite contributors add more value than predicted by a power law.

Despite its perception in the popular media as a chaotic system, Wikipedia has many policies and mechanisms to govern contributions. Forte and Bruckman (2008), Lih (2009), and Viégas *et al.* (2007b), for example, each discuss the various aspects of rule-making, monitoring, conflict resolution, and norms. The most well-known policy is that contributors must write articles using a neutral point of view (NPOV; for related editorial behaviors, see Viégas *et al.* 2004, Bryant *et al.* 2005). Wikipedia defines their NPOV policy as[16]:

> . . . a fundamental Wikimedia principle and a cornerstone of Wikipedia. All Wikipedia articles and other encyclopedic content must be written from a neutral point of view, representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources. This is non-negotiable and expected of all articles and all editors.

4.4.3. Geotagging

Some approaches to geographic content do not require explicit geotagging. Toponym resolution methods, for example, discover geographic content in unstructured text through named entity recognition (Leidner 2007). Training co-occurrence algorithms on unstructured text – either based on Wikipedia articles or web pages – improves toponym resolution (Buscaldi and Rosso 2007, Jones *et al.* 2008, Overell and Rüger 2008, de Alencar *et al.* 2010). Tagging systems, such as in Flickr and YouTube, are effectively unstructured text and also support toponym resolution. In Flickr, however, 100 million images have geotags (as of 30 March 2010 on http://www.flickr.com/map/), and thus toponym resolution methods could construct a gazetteer from probabilistic place semantics (Rattenbury and Naaman 2009). Goodchild and Hill (2008) discuss the potential of such novel methods to develop gazetteers from a wide variety of nontraditional digital sources.

4.4.4. Spatial behaviors

Others have addressed whether contributors exhibit specific spatial behaviors. Lieberman and Lin (2009) studied whether authors have locality within the range of articles to which they contribute. They estimate an author's location based solely on their contribution history to geotagged articles.[17] They then survey a true location of an author to compare with their model results. This survey method might be useful to provide a quality assessment of our results, and, perhaps more importantly, provide a framework for geolocating registered authors.

Hecht and Gergle (2010) further examined whether user-generated content is local and used Flickr and Wikipedia for their analysis. Using article location, they use a mean distance method to estimate an author's region of interest based on which articles he/she edits. They found that distance is a stronger deterrence factor in Flickr than in Wikipedia. From a

content-centric perspective of Wikipedia, Hecht and Moxley (2009) applied Tobler's Law to the link structure of article content. They find that articles are more likely to link to nearby articles. Zook and Graham (2009, cited by Elwood 2010) also analyzed the spatial distribution of user-generated content in Google Maps.

5. Conclusions

Our results provide some insight into the spatiobehavioral characteristics of VGI processes. We find that the likelihood of an anonymous contribution to a geotagged Wikipedia article exponentially decreases as the distance between the contributor and article locations increases. As a group, anonymous contributors also write about fewer places than registered contributors, despite outnumbering them five-to-one (Hardy 2008). In this article, we have developed empirical methods for the spatial analysis of Wikipedia – a real-world, large-scale VGI system. Our approach takes a user-centric perspective of spatial behavior in VGI production and may provide insight into data-centric perspectives of VGI quality and uncertainty assessments.

Research on VGI production is a nascent area with many unexplored avenues. New forms of collective authorship have emerged that raise sociobehavioral questions for geospatial information. That is, GIS has evolved from a technology to a new media with a global user population (Sui and Goodchild 2001). Moreover, our results provide some evidence that geographic distance still matters despite the reduced communication costs of a globalized Internet. To better understand these geographic effects, we discussed one research direction to characterize the spatial patterns of contributors. We plan further modeling of spatiotemporal constraints on the social network of contributors, and comparison of geographic effects across article categories. We also plan to suggest methods for specifying geotags as first-class metadata, rather than fall back on the most basic common denominator of latitude, longitude coordinates. New approaches to geolocation of contributors would likely have a significant impact on VGI research, as current methods are labor-intensive. Data-mining methods with resolution at subarticle levels, such as sections or paragraphs, would improve the sample size. Moreover, geographic and network visualization methods may enable a visual analytics approach to study our corpus.

Acknowledgments

This research was supported in part by the National Science Foundation (Awards #BCS-0849625 'Collaborative Research: A GIScience Approach for Assessing the Quality, Potential Applications, and Impact of Volunteered Geographic Information' and #IIS-0431166 'Collaborative Research: Integrating Digital Libraries and Earth Science Data Systems') and the US Army Research Office (Award #W911NF0910302). We thank Wikimedia Deutschland, e.V. in Berlin, Germany, for providing the helpful *Toolserver* service (http://toolserver.org). They provided database access, web hosting, and computational resources for our study. Thanks to Tim Alder and Stefan Kühn for comments on geotagging methods in Wikipedia, and for sharing their data-mining software and results. Finally, we also thank Sarah Elwood, Danica Schaffer-Smith, Daniel Sui, and especially the anonymous reviewers for their comments.

Notes

1. Priedhorsky *et al.* (2010) used the term *geographic volunteer work* rather than VGI 'to emphasize the active role of end users.'

2. O'Reilly (2005) attributed the term 'long tail' to Chris Anderson who was describing the 'collective power of the small sites that make up the bulk of the web's content.' Barabási and Albert (1999) described the underlying phenomena behind this web topology.
3. Source: Retrieved 23 February 2010, from http://en.wikipedia.org/wiki/Wikipedia:GEO.
4. A workflow of their data mining processes is available in German (Source: http://de.wikipedia.org/wiki/Datei:Wikipedia_Geodata_Workflow.svg [Accessed 25 February 2010]) We used their 22 June 2008 results data, which are available from any Toolserver account in the *u_kolossos_geo_p* database.
5. These are as follows (with their ISO 639-1 codes used by Wikipedia): Catalan (ca), Chinese (zh), Czech (cs), Danish (da), Dutch (nl), English (en), Esperanto (eo), Finnish (fi), French (fr), German (de), Icelandic (is), Italian (it), Japanese (ja), Norwegian (no), Polish (pl), Portuguese (pt), Russian (ru), Slovak (sk), Spanish (es), Swedish (sv), and Turkish (tr).
6. Reportedly, Wikipedia logs IP addresses for all contributions – from anonymous and registered contributors alike – but they restrict access to those data to authorized administrators.
7. GeoLite City database is a freely available version of their commercial product GeoIP City database. MaxMind described their methods as follows: 'We employ user-entered location data from sites that ask web visitors to provide their geographic location. We then run millions of these datasets through a series of algorithms that identify, extract, and extrapolate location points for IP addresses' (http://www.maxmind.com/app/ip-locate [Accessed 17 February 2010]). "GeoIP and GeoLite draw from different seed data sources to generate the IP location data. GeoLite draws primarily from publicly available data and is less accurate, especially at the city level. GeoIP draws primarily from internally collected sources and is more accurate" (Source: http://forum.maxmind.com [Accessed 21 November 2007]).
8. MaxMind provides an 'Accuracy Radius Database' that includes an estimated average error – which they define as the 'average distance between the actual location of the end user using the IP address and the location returned by the GeoLite City database' – for given IP address blocks (Source: http://www.maxmind.com/app/geolite_city_accuracy [Accessed 17 February 2010]). We suspect that their evaluation methods are both nonscientific and limited in scope. For example, we cannot account for large variations between industrialized countries with significant Internet deployments.
9. We define work in simple terms as an edit count for our study, but we recognize that the literature has many different definitions for work, including edit counts (Kittur *et al*. 2007), edit deltas (Zeng *et al.* 2006), edit similarity (i.e., information distance) (Voss 2005), edit longevity (i.e., age or survival or persistence) (Adler and de Alfaro 2007, Wöhner and Peters 2009), and edit visibility (Priedhorsky *et al.* 2007).
10. Wilson (2010, p. 367) serendipitously noted that this exponential nature 'even now appears [as an article] in Wikipedia'.
11. These services include *GEOnet Names Server* (GNS) and *Geographic Names Information System* (GNIS) (Source: http://en.wikipedia.org/wiki/User:The_Anomebot2 [Accessed 24 February 2010]) Using gazetteers as data sources is common for these automated processes, but there are other data sources in use. Another Wikibot (*Rambot*), for example, use its own database of 3,141 countries and 33,832 cities to create geographic articles (Source: http://en.wikipedia.org/wiki/User_talk:Rambot [Accessed 24 February 2010]).
12. The sytax for geographic markup is varied, spanning rich markup with the Geography Markup Language (GML) to simple HTML-based markup with Dublin Core metadata (Kunze 1999), GEO metadata (Daviel and Kaegi 2007), and geo microformat (Çelik 2005).
13. This paradox, in conjunction with privacy and other concerns, has led to what is known as 'sock puppetry' where a single author will use multiple accounts to protect their privacy or otherwise obfuscate their actions.
14. The article they reference was deleted on 8 January 2008, but the article 'Wikipedians with articles' does list full names (Source: http://en.wikipedia.org/wiki/Wikipedia:Wikipedians_with_articles [Accessed 23 September 2010]).
15. Kisilevich *et al.* (2010) reviewed the spatial data-mining methods for tracking trajectories of individuals or groups with such trace data.
16. Source: http://en.wikipedia.org/wiki/NPOV [Accessed 24 September 2010].
17. From these articles, they create a geographic container for the area bound by these articles. They estimate author locations by clustering contribution histories based on a convex hull of the article locations.

References

Adler, B.T. and de Alfaro, L., 2007. A content-driven reputation system for Wikipedia. *In*: *16th international conference on World Wide Web*, 8–12 May, Banff, AB.

Alexa Internet, Inc, 2009. Alexa Traffic Rank [online]. Available from: www.alexa.com/siteinfo/wikipedia.org [Accessed 9 December 2009].

Almeida, R., Mozafari, B., and Cho, J., 2007. On the evolution of Wikipedia. *In*: *1st international conference on weblogs and social media*, 26–28 March, Boulder, CO.

Balk, D. and Yetman, G., 2004. *Gridded population of the world (GPWv3)* [online]. Available from: sedac.ciesin.columbia.edu/gpw/ [Accessed 21 February 2010].

Barabási, A.L. and Albert, R., 1999. Emergence of scaling in random networks. *Science*, 286 (5439), 509–511.

Barclay, T., Gray, J., and Slutz, D., 2000. Microsoft TerraServer: a spatial data warehouse. *In*: *SIGMOD'00*, Dallas, TX, 15–18 May.

Beer, D. and Burrows, R., 2007. Sociology and, of and in Web 2.0: some initial considerations. *Sociological Research Online*, 12 (5), 17.

Benkler, Y., 2002. Coase's Penguin, or, Linux and the nature of the firm. *The Yale Law Journal*, 112 (3), 369–446.

Billón, M., Ezcurra, R., and Lera-López, F., 2008. The spatial distribution of the Internet in the European Union: Does geographical proximity matter? *European Planning Studies*, 16 (1), 119–142.

Brabham, D.C., 2008. Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence*, 14 (1), 75–90.

Bryant, S.L., Forte, A., and Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *In*: *GROUP'05*, 6–9 November, Sanibel Island, FL.

Budhathoki, N., Bruce, B., and Nedovic-Budic, Z., 2008. Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal*, 72 (3), 149–160.

Buscaldi, D. and Rosso, P., 2007. A comparison of methods for the automatic identification of locations in Wikipedia. *In*: *4th workshop on geographical information retrieval*, 9 November, Lisbon, Portugal.

Butler, D., 2006. Virtual globes: The web-wide world. *Nature*, 439 (7078), 776–778.

Cairncross, F., 1997. *The death of distance: How the communications revolution will change our lives*. Boston, MA: Harvard Business School Press.

Castells, M., 2010. *The rise of the network society*. 2nd ed. Chichester, West Sussex: Wiley-Blackwell.

Çelik, T., 2005. Geo microformat [online]. Available from: microformats.org/wiki/geo [Accessed 22 April 2010].

Chernov, S., *et al*., 2006. Extracting semantic relationships between Wikipedia categories. *In*: *1st workshop on semantic wikis*, 12 June. Budva, Montenegro.

Cosley, D., *et al*., 2007. SuggestBot: using intelligent task routing to help people find work in Wikipedia. *In*: *12th international conference on intelligent user interfaces*, 28–31 January. Honolulu, HI.

Danet, B. and Herring, S., 2007. *The multilingual Internet: Language, culture, and communication online*. New York: Oxford University Press.

Daviel, A. and Kaegi, F., 2007. *Geographic registration of HTML documents* [online]. Available from: tools.ietf.org/html/draft-daviel-html-geo-tag-08.txt [Accessed 22 April 2010].

Davis, C., *et al*., 1996. *A means for expressing location information in the Domain Name System*. RFC 1876. Reston, VA: Internet Society.

de Alencar, O., *et al*., 2010. Geographical classification of documents using evidence from Wikipedia. *In*: *6th Workshop on Geographic Information Retrieval*, 18–19 February. Zurich, Switzerland.

de la Beaujardire, J., *et al*., 2000. The NASA digital earth testbed. *In*: *8th international Symposium on advances in geographic information systems*, 6–11 November, Washington, D.C.

De Vries, J.J., Nijkamp, P., and Rietveld, P., 2009. Exponential or power distance-decay for commuting? An alternative specification. *Environment and Planning A*, 41 (2), 461–480.

Elvidge, C.D., *et al*., 2010. Global urban mapping based on nighttime lights. *In*: P. Gamba and M. Herold, eds. *Global Mapping of Human Settlement: Experiences, Datasets, and Prospects*. London: CRC Press, 129–144.

Elwood, S., 2008. Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72 (3), 173–183.
Elwood, S., 2010. Geographic information science: emerging research on the societal implications of the geospatial web. *Progress in Human Geography*, 34 (3), 349–357.
Fallis, D., 2008. Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, 59 (10), 1662–1674.
Foresman, T.W., 2008. Evolution and implementation of the Digital Earth vision, technology and society. *International Journal of Digital Earth*, 1 (1), 4–16.
Forte, A. and Bruckman, A., 2008. Scaling consensus: increasing decentralization in Wikipedia governance. *In*: *41st Hawaii international conference on System Sciences*, 7–10 January, Waikoloa, HI.
Fotheringham, A.S., 1981. Spatial structure and distance-decay parameters. *Annals of the AAG*, 71 (3), 425–436.
Fotheringham, A.S. and O'Kelly, M., 1989. *Spatial interaction models: Formulations and applications*. Dordrecht, The Netherlands: Kluwer Academic.
Frew, J., *et al*., 2000. The Alexandria digital library architecture. *International Journal on Digital Libraries*, 2 (4), 259–268.
Friedman, T.L., 2005. *The world is flat: a brief history of the twenty-first century*. New York: Farrar, Straus & Giroux.
Girardin, F., *et al*., 2008. Digital footprinting: uncovering tourists with user-generated content. *IEEE Pervasive Computing*, 7 (4), 36–43.
Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782.
Goodchild, M.F. and Hill, L.L., 2008. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22 (10), 1039–1044.
Goodchild, M.F., 1999. Implementing digital earth: a research agenda. *In*: *1st international Symposium on Digital Earth*, 29 November–2 December, Beijing, China.
Goodchild, M.F., 2000. Cartographic futures on a Digital Earth. *Cartographic Perspectives*, 36, 3–11.
Goodchild, M.F., 2004. Scales of cybergeography. *In*: E. Sheppard and R.B. McMaster, eds. *Scale and geographic inquiry: nature, society, and method*. Malden, MA: Blackwell, 154–169.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.
Goodchild, M.F., 2008. Geographic information science: the grand challenges. *In*: J.P. Wilson and A.S. Fotheringham, eds. *The handbook of geographic information science*. Malden, MA: Blackwell, 596–608.
Grossner, K.E., Goodchild, M.F., and Clarke, K.C., 2008. Defining a Digital Earth system. *Transactions in GIS*, 12 (1), 145–160.
Gueye, B., Uhlig, S., and Fdida, S., 2007. Investigating the imprecision of IP block-based geolocation. *In*: S. Uhlig, K. Papagiannaki and O. Bonaventure, eds. *Passive and Active Network measurement*. Berlin: Springer-Verlag, 237–240.
Gueye, B., *et al*., 2006. Constraint-based geolocation of Internet hosts. *IEEE/ACM Transactions on Networking*, 14 (6), 1219–1232.
Hägerstrand, T., 1967. *Innovation diffusion as a spatial process*. Chicago, IL: University of Chicago Press.
Haklay, M. and Weber, P., 2008. OpenStreetMap: user-generated street maps. *IEEE Pervasive Computing*, 7 (4), 12–18.
Hall, G.B., *et al*., 2010. Community-based production of geographic information using open source software and Web 2.0. *International Journal of Geographical Information Science*, 24 (5), 761–781.
Hardy, D., 2008. Discovering behavioral patterns in collective authorship of place-based information. *In*: *Internet Research 9.0*, 15–18 October, Copenhagen, Denmark.
Hardy, D., 2010. *Volunteered geographic information in Wikipedia*. Thesis (PhD). Bren School of Environmental Science & Management, University of California, Santa Barbara, CA.
Haynes, K. and Fotheringham, A., 1984. *Gravity and spatial interaction models*. Beverly Hills, CA: Sage.
Hecht, B. and Moxley, E., 2009. Terabytes of tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. *In*: *COSIT'09*, L'Aber Wrac'h, 21–22 September, France.

Hecht, B.J. and Gergle, D., 2010. On the "localness" of user-generated content. *In*: *ACM CSCW'10*, 6–10 February, Savannah, GA.
Hudson-Smith, A., *et al*., 2009. NeoGeography and Web 2.0: Concepts, tools and applications. *Journal of Location Based Services*, 3 (2), 118–145.
Jones, C.B., *et al*., 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22 (10), 1045–1065.
Kisilevich, S., *et al*., 2010. Spatio-temporal clustering. *In*: O. Maimon and L. Rokach, eds. *Data mining and knowledge discovery handbook*. 2nd ed. New York: Springer, 855–874.
Kittur, A., *et al*., 2007. He says, she says: conflict and coordination in Wikipedia. *In*: *ACM CHI'07*, 28 April–3 May. San Jose, CA.
Kühn, S., *et al*., 2008. Wikipedia-World [in German] [online]. Available from: de.wikipedia.org/wiki/Wikipedia:GEO [Accessed 23 February 2010].
Kunze, J., 1999. *Encoding Dublin Core metadata in HTML*. RFC 2731. Reston, VA: The Internet Society.
Lanier, J., 2006. Digital Maoism: the hazards of the new online collectivism. *Edge: The Third Culture* [online], 183.
Leidner, J.L., 2007. *Toponym resolution in text*. Thesis (PhD). School of Informatics, The University of Edinburgh, Scotland, UK.
Lessig, L., 2001. *The future of ideas: the fate of the commons in a connected world*. New York: Random House.
Lieberman, M. and Lin, J., 2009. You are where you edit: locating Wikipedia contributors through edit histories. *In*: *3rd international AAAI Conference on Weblogs and Social Media*, 17–20 May, San Jose, CA.
Lih, A., 2004. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. *In*: *5th international Symposium on Online Journalism*, 16–17 April, Austin, TX.
Lih, A., 2009. *The Wikipedia revolution: how a bunch of nobodies created the world's greatest encyclopedia*. New York: Hyperion.
Luo, S., *et al*., 2009. Toward collective intelligence of online communities: a primitive conceptual model. *Journal of Systems Science and Systems Engineering*, 18 (2), 203–221.
Maguire, D.J., 2007. GeoWeb 2.0 and volunteered geographic information. *In*: *NCGIA Workshop on Volunteered Geographic Information*, 13–14 December, Santa Barbara, CA.
Marston, S., Jones, J., and Woodward, K., 2005. Human geography without scale. *Transactions of the Institute of British Geographers*, 30 (4), 416–432.
MaxMind, 2009. *GeoLite City database, GeoIP geolocation products* [online]. Available from: http://www.maxmind.com/app/geolitecity [Accessed 23 February 2010].
MediaWiki, 2006. *The technical manual for the MediaWiki software: Database layout* [online]. Available from: http://www.mediawiki.org/wiki/Manual:Database_layout [Accessed 23 February 2010].
Morrill, R.L. and Pitts, F.R., 1967. Marriage, migration, and the mean information field: a study in uniqueness and generality. *Annals of the AAG*, 57 (2), 401–422.
Muir, J.A. and van Oorschot, P.C., 2009. Internet geolocation: evasion and counterevasion. *ACM Computing Surveys*, 42 (1). 4:1–4:23.
Nov, O., 2007. What motivates Wikipedians? *Communications of the ACM*, 50 (11), 60–64.
Nov, O., 2009. Information sharing and social computing: Why, what, and where? *In*: M. Zelkowitz, ed. *Social networking and the web*. Amsterdam: Elsevier, 1–18.
Nov, O. and Kuk, G., 2008. Open source content contributors' response to free-riding: the effect of personality and context. *Computers in Human Behavior*, 24 (6), 2848–2861.
O'Reilly, T., 2005. *What is Web 2.0: Design patterns and business models for the next generation of software* [online]. Available from: http://oreilly.com/lpt/a/6228 [Accessed 23 February 2010].
Overell, S. and Rüger, S., 2008. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22 (3), 265–287.
Panciera, K., Halfaker, A., and Terveen, L., 2009. Wikipedians are born, not made: A study of power editors on Wikipedia. *In*: *GROUP'09*, May 51–60, Sanibel Island, FL.
Pfeil, U., Zaphiris, P., and Ang, C.S., 2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12 (1), [online].
Priedhorsky, R., *et al*., 2007. Creating, destroying, and restoring value in Wikipedia. *In*: *GROUP'07*, 4–7 November, Sanibel Island, FL.

Priedhorsky, R., Masli, M., and Terveen, L., 2010. Eliciting and focusing geographic volunteer work. *In*: *ACM CSCW'10*, 6–10 February, Savannah, GA.

Rattenbury, T. and Naaman, M., 2009. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3 (1), 1–30.

Sauer, C., Smith, C., and Benz, T., 2007. WikiCreole: a common wiki markup. *In*: *Int'l Symposium on Wikis*, 21–25 October, Montreal, Quebec, Canada.

Scharl, A. and Tochtermann, K., eds., 2007. *The geospatial web: how geobrowsers, social software and the Web 2.0 are shaping the network society*. London: Springer.

Sen, A. and Smith, T.E., 1995. *Gravity models of spatial interaction behavior*. Berlin: Springer.

Sieber, R., 2006. Public participation geographic information systems: a literature review and framework. *Annals of the AAG*, 96 (3), 491–507.

Stanger, N., 2008. Scalability of techniques for online geographic visualization of web site hits. *In*: A. Moore and I. Drecki, eds. *Geospatial vision: new dimensions in cartography*. Berlin: Springer-Verlag, 193–217.

Strube, M. and Ponzetto, S.P., 2006. WikiRelate! Computing semantic relatedness using Wikipedia. *In*: *AAAI'06*, 16–20 July, Boston, MA.

Sui, D.Z., 2008. The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS [editorial]. *Computers, Environment and Urban Systems*, 32 (1), 1–5.

Sui, D.Z. and Goodchild, M.F., 2001. GIS as media? [editorial]. *International Journal of Geographical Information Science*, 15 (5), 387–390.

Tapscott, D. and Williams, A.D., 2006. *Wikinomics: how mass collaboration changes everything*. New York: Portfolio.

Viégas, F.B., Wattenberg, M., and Dave, K., 2004. Studying cooperation and conflict between authors with *history flow* visualizations. *In*: *ACM CHI'04*, 24–29 April, Vienna, Austria.

Viégas, F.B., Wattenberg, M., and McKeon, M.M., 2007 a . The hidden order of Wikipedia. *In*: *2nd international conference on Online Communities and Social Computing*, 22–27 July. Beijing, China.

Viégas, F.B., *et al*., 2007b. Talk before you type: Coordination in Wikipedia. *In*: *40th Hawaii international conference on System Sciences*, 3–6 January, Waikoloa, HI.

Völkel, M., *et al*., 2006. Semantic Wikipedia. *In*: *15th international conference on world wide web*, 23–26 May. Edinburgh, Scotland.

Voss, J., 2005. Measuring Wikipedia. *In*: *10th international conference of the international society for scientometrics and informetrics*, 24–28 July, Stockholm, Sweden.

Wang, F.Y., *et al*., 2007. Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems*, 22 (2), 79–83.

Weicher, M., 2006. [Name withheld]: Anonymity and its implications. *In*: *ASIS&T annual meeting*, November. Austin, TX.

Wikimedia, 2010. *List of Wikipedias* [online]. Available from: meta.wikimedia.org/wiki/List_of_Wikipedias [Accessed 23 February 2010].

Wilson, A.G., 1969. Notes on some concepts in social physics. *Papers in Regional Science*, 22 (1), 159–193.

Wilson, A.G., 1970. *Entropy in urban and regional modelling*. London: Pion.

Wilson, A.G., 1971. A family of spatial interaction models, and associated developments. *Environment and Planning*, 3 (1), 1–32.

Wilson, A.G., 2010. Entropy in urban and regional modelling: Retrospect and prospect. *Geographical Analysis*, 42 (4), 364–394.

Wöhner, T. and Peters, R., 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. *In*: *5th International Symposium on Wikis and Open Collaboration*, 25–27 October, Orlando, FL.

Yongxiang, L., 1999. Building up the Digital Earth together: Sharing global data resources each other. *In*: *1st international Symposium on Digital Earth*, 29 November–2 December, Beijing, China.

Youn, I., Mark, B., and Richards, D., 2009. Statistical geolocation of Internet hosts. *In*: *18th IEEE international conference on computer communications and networks*, 6–8 August, San Francisco, CA.

Zachte, E., 2009. *Wikimedia visitor log analysis report: Google requests as daily averages* [online]. Available from: stats.wikimedia.org/wikimedia/squids/SquidReportGoogle.htm [Accessed 23 February 2010].

Zachte, E., 2010a. Wikimedia report card: January 2010 [online]. Available from: stats.wikimedia.org/reportcard/ [Accessed 23 February 2010].

Zachte, E., 2010b. *Wikipedia statistics: overview of recent months* [online]. Available from: stats.wikimedia.org/EN/Sitemap.htm [Accessed 23 February 2010].

Zeng, H., *et al*., 2006. Computing trust from revision history. *In*: *international conference on privacy, security and trust*, Markham, Ontario, Canada, 30 October.

Zook, M., 2005. The geographies of the Internet. *Annual Review of Information Science and Technology*, 40 (1), 53–78.

Zook, M. and Graham, M., 2009. Mapping the GeoWeb: The spatial contours of Web 2.0 cyberspace. *In*: *AAG'09*, 22–27 March. Las Vegas, NV.