ASSURING THE QUALITY OF VOLUNTEERED GEOGRAPHIC INFORMATION

Michael F. Goodchild and Linna Li[1]

Abstract

Volunteered geographic information (VGI) is a phenomenon of recent years, offering an alternative mechanism for the acquisition and compilation of geographic information. As such it offers substantial advantages, but suffers from a general lack of quality assurance. We discuss the issues involved in the determination of quality for geospatial data, and trace the history of research on VGI quality. We describe three approaches to quality assurance, which we term the crowd-sourcing, social, and geographic approaches respectively. We discuss the advantages and limitations of each, and the research that will be needed to operationalize the geographic approach.

1. **Introduction**

In the past few years there has been a very rapid growth of interest in volunteered geographic information (VGI; Goodchild, 2007), a version of crowd-sourcing in which members of the general public create and contribute georeferenced facts about the Earth's surface and near-surface to websites where the facts are synthesized into databases. This phenomenon is part of a broader trend, but has special characteristics that are attributable to the geographic nature of the information, in other words to the requirement that every contributed fact specify both a geographic location and a description of one or more properties present at that location. In this context location may refer to a point, a line, an area, or a volume (Longley et al., 2011). VGI has proven very successful as a means of acquiring timely and detailed geographic information at very low cost, but it suffers from several obvious deficiencies. Unlike the traditional authoritative geographic information that it potentially augments and in some cases even replaces, VGI carries no assurance of quality. This paper focuses on the general question of how VGI quality can be assured. As with any discussion of quality or uncertainty in geographic information, it is appropriate to embed the discussion within the general framework of spatial statistics (Zhang and Goodchild, 2002).

Many efforts have been made to study the quality of geographic information. In the 1980s discussions over a potential geospatial data standard for the US Federal Government led to consensus on five fundamental dimensions: positional accuracy, attribute accuracy, logical consistency, completeness, and lineage. This consensus eventually became the basis for the data quality elements of the US Spatial Data Transfer Standard (SDTS; mcmcweb.er.usgs.gov/sdts/) and the Content Standard for Digital

[1]Center for Spatial Studies and Department of Geography, University of California, Santa Barbara, California 93106-4060, USA. {good,linna}@geog.ucsb.edu

Geospatial Metadata (CSDGM; www.fgdc.gov/metadata/csdgm). More recent literature has suggested the addition of further elements, including temporal accuracy and semantic accuracy (Guptill and Morrison, 1995). Definitions of these terms are provided where appropriate later in the paper.

Because the determination of position on the Earth's surface involves measurement, whether by traditional methods of surveying or contemporary GPS (the Global Positioning System), it is inevitable that the reported position of any phenomenon will be inexact. In good scientific tradition, therefore, geographic coordinates should in principle be reported to a precision that reflects the accuracy of the measuring system. In practice, however, geographic information systems (GIS) rarely keep track of measurement accuracy, preferring to process and report coordinates as if they were perfect. It is rare, in fact, for mapping and GIS to track the accuracy of positions, making it difficult to assess the impacts of accuracies on the products and applications of GIS analysis.

In the 1990s a widespread recognition emerged that while a conceptual framework of *accuracy* and error analysis was in many cases sufficient for the measurement of position, it was much less successful as a comprehensive approach to other dimensions of geospatial data quality, because the inherent vagueness of many geospatial concepts, including classes of soil, land cover, and other geographic taxonomies, made it impossible to agree on the true value of many mapped properties or to satisfy the scientific requirement of replicability. The theoretical framework of fuzzy sets was seen as inherently more compatible with vagueness, and the word *uncertainty* was held to be more suitable than accuracy (Zhang and Goodchild, 2002).

Against this background, this paper examines the assurance of data quality in VGI. ~~The next section expands on this introductory discussion of geospatial data quality and discusses some of the specific characteristics of geographic information and their implications for quality assurance.~~ Section 2~~3~~ provides additional background on VGI and its relationship to authoritative geospatial data. Section 3~~4~~ presents three alternative approaches to quality assurance and discusses the advantages and disadvantages of each. Finally, Section ~~5~~4 draws together the paper's conclusions. The paper is intended to provide a conceptual framework for this emerging field, and many of the conclusions lead to recommendations for further research and development.

~~2. Geospatial data quality~~

~~The fundamental item of geographic information is a triple <**x**,*Z*,*z*(**x**)> where **x** is a point location in geographic space (and time, where relevant), *Z* is a property present at that location, and *z*(**x**) is the value of that property at that location. Goodchild, Yuan, and Cova (2007) term this the *atomic* form of geographic information. In practice atoms are observed only in certain domains, such as the measurement of atmospheric temperature or terrain elevation; instead, observations are often made for entire areas (e.g., county name, land-use type), an approach that achieves substantial compression when a property remains essentially constant over an area. Properties may also be attached to lines (e.g., street name, river name) when dealing with areas that can be approximated as linear.~~

Two distinct methods of aggregating these *geo-atoms* are recognized. First, as the previous paragraph suggests, atomic statements can be aggregated into lines, areas, or volumes if properties extend over space, in an approach known as the *discrete-object* conceptualization. These discrete objects may overlap, and are generally countable. Second, atomic statements about the same property can be aggregated into *continuous fields*, when it is believed that the property can be measured at any location in space (and if relevant, time). Continuous fields are commonly used to represent the spatial variation of such properties as terrain elevation, rainfall, land ownership, or population density. On the other hand buildings, individual people, and trees may be better conceptualized as discrete objects.

The duality of discrete objects and continuous fields has immediate relevance to data quality. Positional accuracy of discrete objects can be assessed by comparing the measured position of a point or line, or the boundary of an area, to a reference source of better quality. Point accuracy is often described using the parameters of a circular normal distribution, while there have been many efforts to find comparable measures of positional accuracy for lines and for area boundaries (e.g., Shi and Liu, 2000). Attribute accuracy refers to the difference between the properties associated with the object and the properties recorded in a reference source of better quality. For quantitative attributes the root-mean-square error is an appropriate statistic, while the kappa statistic is often used for qualitative attributes.

Continuous fields raise important conceptual issues. Consider, for example, a field of terrain elevation. Differences between the elevation recorded at a point **x** and the elevation in a reference source at the same point may be caused by any combination of errors of horizontal position and errors of elevation. Thus positional and attribute accuracy are inherently inseparable. Similarly it makes little sense to ask whether the contours of an elevation surface are in the correct locations, since a constant uncertainty in elevation, due for example to the use of GPS, will produce a range of displacements of contours, depending on local slope. A nominal field, often termed an area-class map, also raises a range of problems for accuracy assessment. Areas mapped as having the same class will mask substantial internal variation, making it meaningless to ask whether an area has the correct class. The positions of boundaries between areas will be affected by uncertainty, as will the number of areas and the topology of the boundary network. Thus the only scientifically rigorous method of accuracy assessment will compare the classes mapped at a random sample of locations with the classes mapped at those locations in a reference source.

Spatial dependence is a persistent property of geographic information, often stated as Tobler's First Law of Geography: "All things are related, but nearby things are more related than distant things" (Tobler, 1970). When applied to positional accuracy, it suggests that errors of position will be similar over short distances. Let $\mathbf{e}(\mathbf{x})$ denote the error of position at point **x**; in other words, **e** denotes a vector field of displacements. Many sources of positional error induce strong positive spatial dependence in **e**. For example, misregistration of satellite imagery will be inherited by every point whose

position is determined using that imagery, perhaps by using Google Earth. Errors in GPS positions also exhibit strong spatial dependence over short periods of time due to temporal persistence in satellite ephemerides. When positive spatial dependence is present in the positional error field, errors in distances between nearby points will be reduced, and the shapes of features will be better preserved. In general, relative errors of position of nearby points will be substantially less than absolute errors of position. Thus average positional error may give a greatly inflated impression of relative positional uncertainty. On the other hand, this will not be true if positions have been determined independently, from different sources that have been impacted by their own errors. Thus the lineage component of spatial data quality may have great value in informing the user of the provenance of data.

Goodchild (2009) has summarized many of these arguments in suggesting that the approach to data quality represented by the SDTS and CSDGM needs substantial revision, being based on earlier research that is now decades out of date.

2~~3~~. Volunteered geographic information

As noted earlier, interest in VGI has been growing rapidly, fuelled by a number of factors. In the UK and many other countries much geographic information collected by governmental mapping agencies is available only at high cost. By 2004~~, however,~~ it had become obvious that individuals could create their own digital geographic information, in the form of high-quality, on-line maps, essentially at no cost. Coordinates of recognizable features could be readily obtained using GPS, or by finding features on the fine-resolution images made available by services such as Google Earth. Expertise in cartography was no longer necessary, since open-source software for the production of high-quality maps was now widely available. Moreover a volunteer could create maps of his or her local area, using local knowledge, more effectively than a mapping expert from a distant government agency. OpenStreetMap became the best-known, and clearly most successful, of a number of efforts begun at that time to create an alternative to the products of official agencies.

VGI is a specific instance of a much more general phenomenon, the creation of online information by the general public. Early applications of the Web were essentially top-down, using technology to disseminate a message to a large number of users. The term *Web 2.0* reflects this partial reversal, in which much of the information that now populates the Web comes from those who were previously exclusively users. The earlier distinction between producer and consumer is now somewhat moot, and has led to the coining of the term *prosumer*. In the geographic context, these new arrangements have been termed *neogeography* (Turner, 2006), a term that suggests both a breaking down of the traditional distinction between geographic expert and geographic amateur, and also a new era in which it is possible to produce maps of anything, at any time, for any purpose, from any viewpoint, and at virtually no cost. One of the most interesting questions to emerge from this transition asks: If people can map anything, what will they choose to map? There is now abundant evidence that the answer differs sharply from what has traditionally been mapped by mapping agencies.

~~It is easy but perhaps simplistic to distinguish between the *asserted* information of VGI and the authoritative information of traditional mapping. The latter is subject in many cases to rigorous quality control, and is often accompanied by detailed metadata. VGI on the other hand almost never carries metadata, and has no quality control. However the lack of quality assurance can often be offset by major advantages. VGI is essentially free, and perhaps more importantly it can be much more timely, since a dense network of amateur observers can compete very effectively with a few sparsely distributed experts, especially in time-critical situations such as emergencies. Goodchild and Glennon (2010) note how citizens during an emergency must often balance inaccurate information now against accurate information later.~~

~~Arguments between asserted and authoritative options are also confounded by the possibility of hybrids. Well before the surge of interest in VGI it was common for assemblers of geographic information to rely on networks of amateur observers, especially for updates about new features and corrections to old ones. Google's MapMaker project enlists the help of volunteers to update, correct, and augment its map data, while TomTom invites users of its in-vehicle navigation systems to allow their tracks to be uploaded by TomTom and used to improve the accuracy of its base data. Coleman (in press) and Dobson (in press) provide reviews of similar efforts.~~

Several studies have evaluated the accuracy of VGI against reference sources. Haklay (2010), for example, has compared the quality of OpenStreetMap with data from the Ordnance Survey of Great Britain, the national mapping agency. He found an average positional displacement of 6m, but a substantial geographic variation in both positional accuracy and completeness. Girres and Touya (2010) provide a comparable analysis of French OpenStreetMap data, and Cipeluch et al. (2010) analyze Irish OpenStreetMap against Google Maps and Bing Maps. Such studies give useful insights into the accuracy of VGI, but only indirectly help to identify mechanisms for assuring and improving quality.

More broadly, authoritative data are increasingly out of date in many parts of the world (Estes and Mooneyhan, 1994), and were acquired using older technologies that were less accurate than those available to the general public today. In the US, it is easy for a volunteer to acquire data that are more accurate than the 1:24,000-scale mapping that is most detailed complete coverage available for the continental states. Thus in many circumstances it is easy to show that VGI is of better quality than the best available authoritative data.

These arguments provide the motivation for the remainder of the paper, which addresses the question of how to assure and improve quality in VGI.

3~~4~~. Mechanisms for quality assurance

In this section we discuss three alternative approaches to quality assurance, which we term the *crowd-sourcing*, *social*, and *geographic* approaches. All three provide

mechanisms for triaging VGI, to determine the degree to which a purported fact is likely to be true, and to take appropriate action, perhaps by accepting the contribution, or questioning it, or rejecting it outright. The degree to which such triage can be automated varies; in some cases it might be fully automatic, but in other cases it might require significant human intervention. Our use of the terms truth and fact suggest an orientation towards VGI that is objective and replicable, and for which quality can be addressed using the language of accuracy. Thus our approach is less likely to be applicable for VGI that consists of opinion, or VGI that addresses properties that are vaguely defined.

The quality assurance provided by traditional mapping agencies has two parts: procedures designed to control quality during the acquisition and compilation of geospatial data, and procedures that document quality by taking samples of compiled data and comparing them to reference sources. The latter procedures are typically documented, and the results are attached to the data in the form of metadata. Some degree of generalization is inevitable, of course, since it is impractical to check every item of data.

The research on VGI accuracy discussed in the previous section follows the second part of the assurance process. In this paper we focus on the first: on procedures to enhance quality during the acquisition and compilation steps.

34.1 The crowd-sourcing approach

The term crowd-sourcing has three distinct meanings. The first, stronger meaning has its roots in the terms crowd and out-sourcing, and refers to the solution of a problem by referring it to a number of people, without respect to their qualifications. Surowiecki (2005) in his book *The Wisdom of Crowds* cites examples of how a group of people may converge on the solution to a problem that an individual, even an expert, may be unable to solve. The principle has been widely exploited through the mechanism of the Web, exemplified by sites such as Amazon's Mechanical Turk (www.mturk.com).

However, the second and third meanings are more relevant to quality assurance for VGI: rather than solving a problem, crowd-sourcing in these second and third senses refers to the ability of a group to validate and correct the errors that an individual might make. Our second interpretation of crowd-sourcing is relevant to events such as wildfires, when a single observation of the presence of a fire is strengthened by additional observations from the same or nearby points. Recent work on Twitter reports during emergencies (for a review see Latonero and Shklovski, 2010) focuses on this interpretation, giving greater weight to a spatial clustering of similar reports than to a single report. In such cases one might devise metrics of quality based on the number of independent but consistent reports.

The third interpretation refers to the ability of the crowd to converge on the truth. Linus's Law, named by Raymond (1999) in honor of Linus Torvalds, the originator of Linux, states that "given enough eyes, all bugs are shallow"; in the context of software engineering, the bugs in a piece of software are most likely to be found and corrected if a

large number of software engineers have an opportunity to review the software. The principle provides a form of quality assurance for sites such as Wikipedia that rely on individuals to volunteer encyclopedia content: if one individual contributes an error, others can be expected to edit and correct the error, and the success of this mechanism rises in proportion to the number who look at the contribution. Wilkinson and Huberman (2007) document the effectiveness of this mechanism in the case of Wikipedia.

It is tempting to think that Linus's Law can be applied to quality assurance for VGI. Mooney (2011) has examined patterns of editing of OpenStreetMap content, and has even found instances of "tag wars" in which the type associated with a feature is repeatedly changed by pairs of contributors who disagree on the correct type. In another study of OpenStreetMap, Haklay et al. (2010) have shown that positional accuracy of a feature improves as the number of editors increases, but no further improvement is evident when the number of editors exceeds 13.

In principle one might expect Linus's Law to apply to VGI. But in practice its reliance on multiple "eyes" suggests that it will apply best to geographic facts that are prominent, and less well for facts that are obscure. A fact might be obscure because it refers to a location in a sparsely populated or under-explored part of the world, or because it persists for only a short period of time, or because it refers to a property *Z* that interests few people. Consider, for example, the case shown in Figure 1, which depicts an area in Goleta, California. Wikimapia (www.wikimapia.org) is a project to "describe the whole world" by identifying and providing detailed descriptions of point or area features. At time of writing over 17 million features had been located and described. Wikimapia is in effect a crowd-sourced gazetteer (Goodchild and Hill, 2008), a catalog of placenames, geographic locations, and feature descriptions.

[Figure 1 about here]

Figure 1 shows a highlighted feature labeled Glen Annie Golf Course, contributed in 2009 by an individual who at time of writing had made over 40,000 contributions to Wikimapia. The attached description reads: "405 Glen Annie Rd., Santa Barbara, CA 93117, (805) 968-6400, www.glenanniegolf.com/golf/proto/glenanniegolf/. Glen Annie Golf Club is a championship golf course with first class amenities situated in the rolling foothills of scenic Santa Barbara. This challenging golf layout is enhanced by breathtaking panoramic views of the Pacific Ocean and Channel Islands and is always maintained with the highest standards." Unfortunately none of that information is true of the outlined feature, which is in fact the Ocean Meadows golf course, at a different street address. Glen Annie Golf Club is indeed located at the stated street address, which lies approximately 2km north of the asserted location, and fits the description. But the terrain of the Ocean Meadows course could hardly be described as rolling foothills, and it is impossible to see the Pacific Ocean and Channel Islands from any part of it. In principle the crowd-sourcing mechanism should have corrected this error quickly, but that has not happened; instead, the erroneous entry is still present at the time of writing, and a duplicate entry appeared at the correct location in late 2011, prompted perhaps by the authors' drawing attention to this example. Moreover it is possible to speculate on how

an automated triage system might have checked the street address or computed the viewshed and the terrain roughness, and either queried or rejected the contribution.

Several factors may have contributed to the failure of Linus's Law in this instance. First, Wikimapia may attract people with a greater interest in contributing than in editing; its system of awards and recognition places most of its emphasis on contributions. Second, although this area of California is fairly densely populated, the number of "eyes" with sufficient local knowledge may be very small. Third, the status of the contributor as an Advanced User (see next section) may have given an exaggerated level of implied trust to the contribution. Finally, the architecture of the site may make it too difficult to correct or delete errors.

More generally, however, it is clear that Linus's Law is not as effective for geographic facts as it is for the more prominent information that forms the core of sites such as Wikipedia. Editing in OpenStreetMap is clearly effective according to Haklay's analysis, perhaps because the social network supporting the project is large and active, providing many "eyes". But obscure features may not attract sufficient numbers of "eyes" to assure quality. Lack of "eyes" may also lead to malicious use: why not find a feature in an obscure part of the world and put one's own name on it?

34.2 The social approach

We term the second approach social because it relies on a hierarchy of trusted individuals who act as moderators or gate-keepers. Many studies have shown that voluntary contributions by individuals follow a frequency distribution with a long tail, with a few individuals making large numbers of contributions, and a large number of individuals making only one or a few. For example, Mooney and Corcoran (in press) analyzed OpenStreetMap data for the British Isles, and found that for heavily edited features (at least 15 edits per feature) 84% of edits were made by 12% of contributors.

Tracking contributions by individuals allows for the calculation of metrics of commitment and reliability that may provide a basis for trust. For example, an individual who makes prolific contributions that attract few edits could be assigned a high score, and given a role as moderator or gate-keeper. Adler and de Alfaro (2007) proposed a reputation system for Wikipedia as a way to evaluate authors' reliability according to the lifespan of their edits. Wikimapia provides extensive data on the contributions of individuals, and assigns them awards and a hierarchy of roles and responsibilities. For example, the role of Advanced User "is awarded to a member of Wikimapia who has good standing and it provides additional tools and responsibilities to take care of Wikimapia. It is given by the Wikimapia Team members mostly on the recommendations of the other Advanced Users." Separate awards are given for prolific creating and editing of places and roads, and for adding large numbers of photographs to feature descriptions. Of course being prolific does not necessarily equate with being accurate. Ideally, the hierarchy of the reviewers should also vary from place to place, as local residents tend to have first-hand knowledge.

-This type of hierarchical structure already exists in many self-regulated collaborative efforts. In Wikipedia, there is a group of administrators called *sysops* who have privileges that are not provided to ordinary editors, such as deleting articles, protecting pages from editing, and blocking users. The purpose of the administrators is to assure the quality of articles in Wikipedia by resisting vandalism, removing copyrighted materials and abusive content, and resolving conflicts between edits. Among 15,805,497 registered users in the English Wikipedia, there are 1,517 sysops, of whom about 25 sysops occupy the top level of the administration. Similarly, in OpenStreetMap there are two tiers in the hierarchy of contributors: ordinary users, and the Data Working Group (DWG), currently composed of eight members, who deal with vandalism, copyright violation, disputes, geographic locking, etc. Every registered user can add and edit geographic features; however, if there is any dispute, DWG decides the solution.

In effect, such hierarchies of trust emulate the structure of traditional authoritative mapping agencies, where experience and qualifications act as surrogates for reliability, and promotion leads to greater authority and higher salaries. The social approach to VGI quality assurance replaces the formal structure of an agency with a much less formal, voluntary structure. Moreover the mantle of trust acquired by a mapping agency is as much due to its formal, documented program of accuracy assessment as it is to the stature of its employees or their productivity. Flanagin and Metzger (2008) provide a detailed discussion of trust and credibility in the context of VGI. They identify a pressing need to determine the motivations of contributors, and argue that VGI is more likely to be trusted than other forms of user-generated content because geographic facts are perceived to be objective and replicable.

Some of the more effective applications of the social approach fall into the hybrid area described previously, by combining elements of both volunteered and authoritative practice. Google's MapMaker project, for example, relies on a network of volunteers, promotes prolific contributors, but reserves the top level of the hierarchy and the final gate-keeping role to Google employees. Only when a volunteered fact has been accepted at the top level is it added to Google's own geographic databases. A similar approach is followed by other companies that rely on volunteers to provide updates and corrections.

34.3 The geographic approach

We term the third approach geographic because it relies on a comparison of a purported geographic fact with the broad body of geographic knowledge. Just like a language, the geographic domain is constrained by certain rules of syntax that govern what can and cannot occur at a given location. Some are known because they have been abstracted from reality through scientific research, while others exist in the practices of regulatory agencies. In essence, this approach asks how it is possible to know whether a purported geographic fact is true or false (and in practice, how to order purported facts along a continuum of likelihood).

One such rule has already been discussed. Tobler's First Law of Geography, "All things are related, but nearby things are more related than distant things" (Tobler, 1970), has

two distinct implications. As noted earlier, the second part describes spatial dependence, and suggests that a purported fact about a location should be consistent with what is already known about the location's vicinity; in other words, a purported fact should be consistent with its geographic context. A report of a wildfire, for example, is more likely to be true if wildfire has already been reported nearby. The first part of the principle is also relevant if it is interpreted as saying that a purported fact should be consistent with other facts that are already known about the same location. We might term these *horizontal* and *vertical* context respectively.

Figure 2 provides a convenient example. It shows a contribution to Flickr of a photograph of "Joe's Café" that has been georeferenced to a point in an area of open-space park next to the Santa Barbara Mission. The content of the photograph is clearly inconsistent both with what is already known about that point (it lies in a park) and with what is known about the surroundings (a historic area that contains no businesses). As such, it could be queried or rejected as likely to be mis-located; or the actual location of Joe's Café (536 State St) could be obtained from a Web search.

[Figure 2 about here]

The rules that govern the geographic domain range from the very simple to the highly abstract. Simple rules of feature geometry, for example, may be precise and therefore an effective basis for triage, but may not merit a mention in the scientific literature of geography or related disciplines. Companies such as Navteq, which encourage volunteers to contribute updates and corrections to their databases of streets and roads, have found it helpful to implement hundreds of rules, many of them simple requirements of geometry,. Indeed such rules are essential to provide an automated triage of the very large number of updates and corrections that flow from the consumer base. For example, an off-ramp is required to depart from a freeway at a small angle in the direction of traffic flow, and freeways have strict limits to radius of curvature, based on design speed. Tobler's First Law, on the other hand, is sufficiently general and abstract to merit mention in the academic literature. Other geographic knowledge, such as the commonly observed fractal nature of coastlines and other natural features (Mandelbrot, 1982), or the statistical laws of channel networks (Horton, 1945), or the principles of economic geography (Christaller, 1966), impose much more abstract constraints on the geographic landscape.

While it is clearly possible to identify hundreds of such rules, their implementation in a rule base as a system for triage of VGI remains a challenging and yet clearly well-motivated research problem. An interesting extension arises indistinguishing between real and imagined landscapes: is it possible to determine whether an imagined landscape could exist, as a composite of purported geographic facts? To illustrate the geographic approach to VGI quality assurance we have analyzed the case of Allestone, an imaginary world created around 1800 by the five-year-old child prodigy Thomas Williams Malkin (Ellis, 1970; Hewitt, 2010). The map (Figure 3) came to the attention of the romantic poet and painter William Blake, who engraved it. The map is fully populated by rivers and named places, and its "firm and determinate outline" was much admired by Blake. Somewhat ironically Blake was one of the foremost opponents of the nascent Ordnance

Survey of Great Britain and its maps, which he saw as "enslaving the human mind to reason and universal laws" (Hewitt, 2010, p206), that is, precisely the kind of intellectual activity we explore in the next paragraphs.

[Figure 3 about here]

First we consider the coastline of Allestone. Mandelbrot (1967) describes the use of an analysis attributable to Richardson (1960) to demonstrate the fractal nature of coastlines by creating a double-log plot of measured length against the scale of the measuring device. In Figure 4 we used the dividers method, counting the number of steps needed to traverse the coastline when the dividers are set at a given spacing. The points show a strong adherence to the expected straight line, and the slope of the fitted line indicates a fractional dimension of 1.24. Clearly the coastline of this imagined world follows the same principle as that generally observed for real coastlines, and cannot therefore be rejected as deviating from this known principle of the geographic world.

[Figure 4 about here]

Second, we consider the stream networks depicted on the imagined map. Stream networks are generally observed to form tree graphs, and can be numbered using a simple topological approach: all head streams are of order 1; when two first-order streams join a second-order stream is formed; when two second-order streams join a third-order stream is formed, and so on. When two streams of different order join the resulting stream has the higher of the two orders. Horton (1945) observed that the counts of streams numbered in this way followed a simple exponential distribution, with a constant ratio, known as the bifurcation ratio, between the number of order i and the number of order i+1. Shreve (1970) provided a simple maximum-likelihood explanation. Figure 5 shows the Horton plot of Allestone streams, clearly demonstrating adherence to Horton's Law with a bifurcation ratio of 2.7.

[Figure 5 about here]

Finally, we consider the economic geography of Allestone and the distribution of named places. According to the tenets of Central Place Theory (Christaller, 1966), settlements on a uniform agricultural plain should be regularly spaced in a hexagonal pattern, each one providing a set of services to its surrounding area. We analyzed the distribution of the 321 named places as points using Ripley's k-function. Figure 6 shows the results, comparing the actual function with the function expected in a pattern of points of the same density but distributed according to Complete Spatial Randomness, that is, points independently located and equally likely to occur anywhere. Clearly points are more regularly spaced than random over distances less than 53 "Allestonish miles", falling short of the perfect hexagonal pattern of theory but nevertheless consistent with findings in the literature (Berry and Pred, 1965).

[Figure 6 about here]

These three analyses show that the Allestone map is consistent with three of the more advanced principles on which the geographic landscape is constructed. In other words, we cannot clearly identify Allestone as imagined: it might be a part of the real world.

Although elements of the geographic approach are already in place in at least one corporation and functioning to triage VGI, the rules in use have been assembled pragmatically and without a conceptual or theoretical framework. What is missing, therefore, is research to provide such a framework, and to begin the process of populating it with geographic knowledge. While substantial knowledge exists in the academic literature, it is not formalized in a manner that would make it readily amenable to implementation. Thus although the geographic approach has promise, we are a long way from effective implementation at the current time.

45. Conclusion

VGI has enormous advantages: it is free, vast amounts of it can and are being created, it can be timely, and it can provide types of data that have never figured before in mapping practice. On the other hand its quality is highly variable and undocumented, it fails to follow scientific principles of sampling design, and its coverage is incomplete. As an alternative source of the kinds of data required for scientific research it is often inadequate, though it may play a useful role in the early, exploratory, and hypothesis-generating stages of science. Clearly there would be great benefit if its quality could be improved and assured.

In this paper we have examined three alternative approaches to quality assurance, each of which has merit. We do not assert that these are the only three possible approaches, and would welcome additions to the list. The crowd-sourcing approach is attractive but appears to be less effective for geographic facts than for other types of information, because such facts vary from the very prominent to the very obscure, and errors in geographic facts can clearly persist even in heavily populated areas. The social approach is also attractive, and has proven effective when projects are sufficiently popular and well-structured, and supported by extensive social networks. The geographic approach is attractive to geographers because it taps the heart of geographic knowledge, and motivates a comprehensive effort to formalize such knowledge; and it offers the possibility of automated triage. Although the approach has been implemented by at least one corporation, demonstrations of its potential in the open literature are clearly needed. Research will also be needed to determine how to structure and formalize geographic knowledge within a consistent and comprehensive framework.

These three approaches to quality assurance compare well with the processes of accuracy assessment used by traditional mapping agencies. It is expensive and time-consuming to assemble the reference data needed to check geographic facts, and generalization from a sample of checks is always problematic given the heterogeneity present in the real world. Moreover, by focusing on the first stage of the quality assurance process, these approaches offer direct improvement in quality, rather than merely its posterior assessment and documentation.

References

Adler, B.T. and L. de Alfaro, 2007. A Content-Driven Reputation System for the Wikipedia**.** In *WWW 2007, Proceedings of the 16th International World Wide Web Conference.* Washington, DC: ACM Press.

Berry, B.J.L. and A. Pred, 1965. *Central Place Studies: A Bibliography of Theory and Applications*. Philadelphia, PA: Regional Science Research Institute.

Christaller, W., 1966. *Central Places in Southern Germany*. Translated by C.W. Baskin. Englewood Cliffs, NJ: Prentice-Hall.

Cipeluch, B., R. Jacob, A. Winstanley, and P. Mooney, 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In N.J. Tate and P.F. Fisher, editors, *Proceedings, Ninth International Syposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2010), Leicester, UK, July 20-23*, pp. 337-340.

~~Coleman, D., in press. Potential contributions and challenges of VGI for comprehensive public and private sector base mapping programs. In D. Sui, S. Elwood, and M.F. Goodchild, editors, *Volunteered Geographic Information, Public Participation, and Crowdsourced Production of Geographic Knowledge: New Developments and Applications*. New York: Springer.~~

~~Dobson, M., in press. The role of VGI as a compilation tool for map databases used in navigation and location applications. In D. Sui, S. Elwood, and M.F. Goodchild, editors, *Volunteered Geographic Information, Public Participation, and Crowdsourced Production of Geographic Knowledge: New Developments and Applications*. New York: Springer.~~

Ellis, E.J., 1970. *The Real Blake: A Portrait Biography*. New York: Haskell House.

Estes, J.E. and W. Mooneyhan, 1994. Of maps and myths. *Photogrammetric Engineering and Remote Sensing* 60(5): 517-524.

Flanagin, A.J. and M.J. Metzger, 2008. The credibility of volunteered geographic information. *GeoJournal* 72: 137-148.

Girres, J-F. and G. Touya, 2010. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* 14(4):435-459.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211-221.

~~Goodchild, M.F., 2009. Putting research into practice. In A. Stein, W. Shi, and W. Bijker, editors, *Quality Aspects of Spatial Data Mining*. Boca Raton: CRC Press, pp. 345-356.~~

~~Goodchild, M.F. and J.A. Glennon, 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth* 3(3): 231-241.~~

Goodchild, M.F. and L.L. Hill, 2008. Introduction to digital gazetteer reseearch. *International Journal of Geographical Information Science* 22(10): 1039-1044.

Goodchild, M.F., M. Yuan, and T.J. Cova, 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science* 21(3): 239-260.

Guptill, S.C. and J.L. Morrison, editors, 1995. *Elements of Spatial Data Quality*. Oxford: Elsevier.

Haklay, M., 2010. How good is volunteered geographical information? Acomparative study of OpenStreetMap and Ordnance Survey datasets.*Environment and Planning B: Planning and Design*37(4): 682-703.

Haklay, M., S. Basiouka, V. Antoniou, and A. Ather, 2010. How many volunteers does it take to map an area well? The validity of Linus's Law to volunteered geographic information. *Cartographic Journal* 47(4): 315-322.

Hewitt, R., 2010. *Map of a Nation: A Biography of the Ordnance Survey*. London: Granta.

Horton, R.E., 1945. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Bulletin of the Geological Society of America* 56(3): 275-380.

Latonero, M. and I. Shklovski, 2010. "Respectfully yours in safety and service": emergency management and social media evangelism. In*Proceedings of the 7th International ISCRAM Conference, Seattle, USA, May*.

Longley, P.A., M.F. Goodchild, D.J. Maguire, and D.W. Rhind, 2011. *Geographic Information Systems and Science*. Third Edition. Hoboken, NJ: Wiley.

Mandelbrot, B.B., 1967. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156: 636-638.

Mandelbrot, B., 1982. *The Fractal Geometry of Nature*. San Francisco: Freeman.

Mooney, P., 2011. The evolution and spatial volatility of VGI in OpenStreetMap. Paper presented at the Hengstberger Symposion "Towards Digital Earth: 3D Spatial Data Infrastructures", Heidelberg, September 7-8.

Mooney, P. and P. Corcoran, in press. Characteristics of heavily edited objects in OpenStreetMap. *GeoJournal*.

Raymond, E.S., 1999. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Beijing: O'Reilly.

Richardson, L.F., 1960. *Statistics of Deadly Quarrels*. Pacific Grove, CA: Boxwood.

Shi, W.Z. and W.B Liu, 2000. A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science* 14(1): 51-66.

Shreve, R.L., 1970. Statistical law of stream numbers. *Journal of Geology* 74(1): 17-37.

Surowiecki, J, 2005. *The Wisdom of Crowds*. New York: Anchor.

Tobler, W.R., 1970. A computer movie simulating growth in the Detroit region. *Economic Geography* 46(2): 234-240.

Turner, A., 2006. *Introduction to Neogeography*. Sebastopol, CA: O'Reilly.

Wilkinson, D.M. and B.A. Huberman, 2007. Asssessing the value of cooperation in Wikipedia. http://arxiv.org/abs/cs/0702140.

Zhang, J.-X. and M.F. Goodchild, 2002. *Uncertainty in Geographical Information*. New York: Taylor and Francis.

Figure captions

1**.**An extract of Wikimapia for Goleta, California highlighting the persistent but erroneous entry for Glen Annie Golf Course. Note the subsequent and correct entry for Glen Annie Golf Club approximately 2km to the north.

2.An erroneous entry in Flickr of a photograph showing “Joe’s Café” in a park next to the historic Santa Barbara Mission.

3. The map of the imaginary land of Allestone drawn by Thomas Williams Malkin around 1800.

4.Double-log Richardson plot of the coastline of Allestone, showing length plotted against the spacing of length-measuring dividers, confirming the coastline’s fractal properties.

5. Semi-log plot of Allestone streams, showing a relationship between the number of streams and stream order that is consistent with Horton’s Law

6.Ripley’s *k*-function analysis of the pattern of named places on the Allestone map. The blue line represents Complete Spatial Randomness. The red line indicates that named places are further apart than expected over short distances.

Figure 1:An extract of Wikimapia for Goleta, California highlighting the persistent but erroneous entry for Glen Annie Golf Course. Note the later, correct entry for Glen Annie Golf Club approximately 2km to the north.

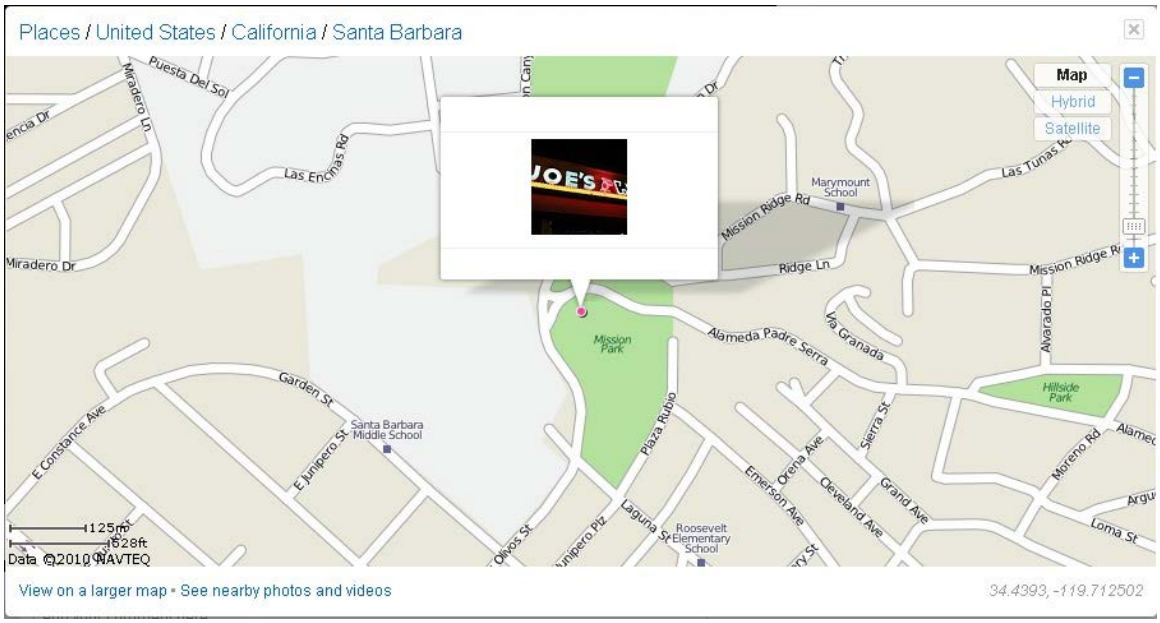

Figure 2: An erroneous entry in Flickr of a photograph showing “Joe’s Café” in a park next to the historic Santa Barbara Mission.

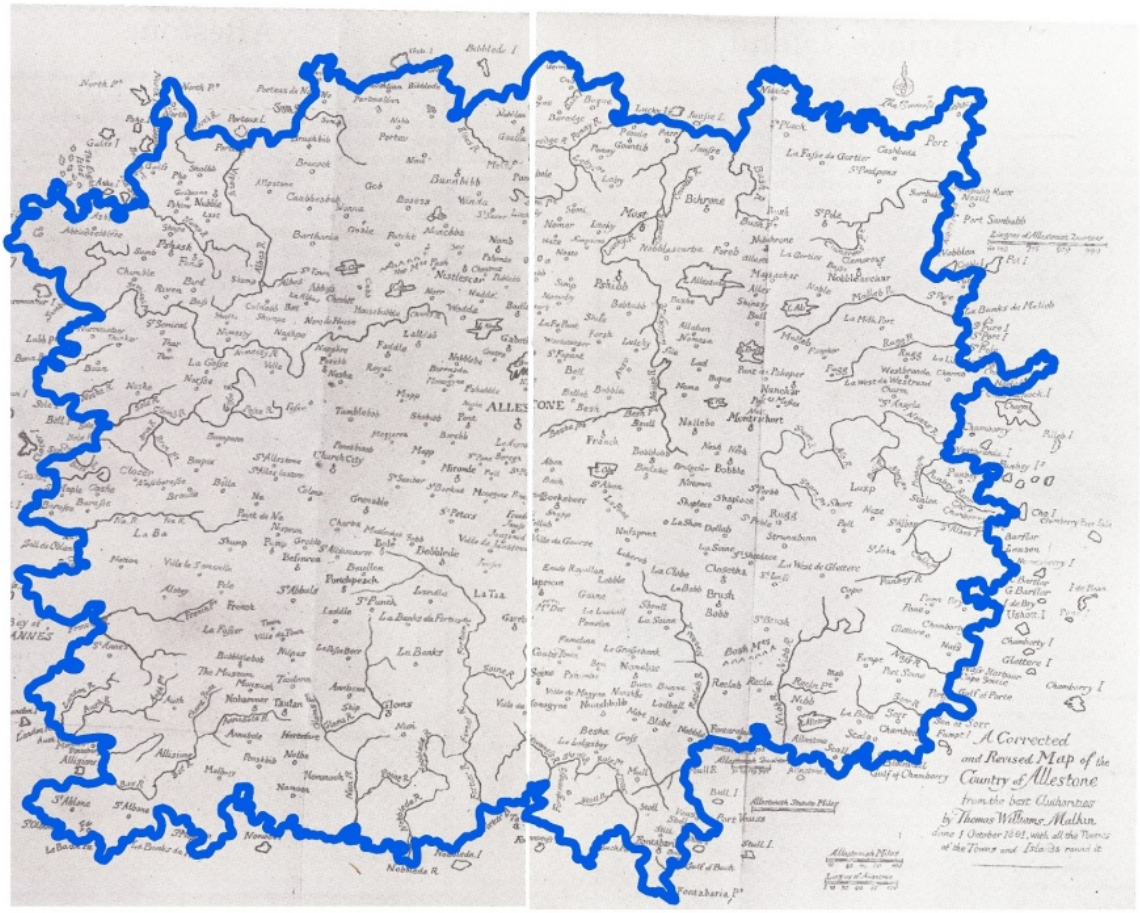

Figure 3: The map of the imaginary land of Allestone drawn by Thomas Williams Malkin around 1800.

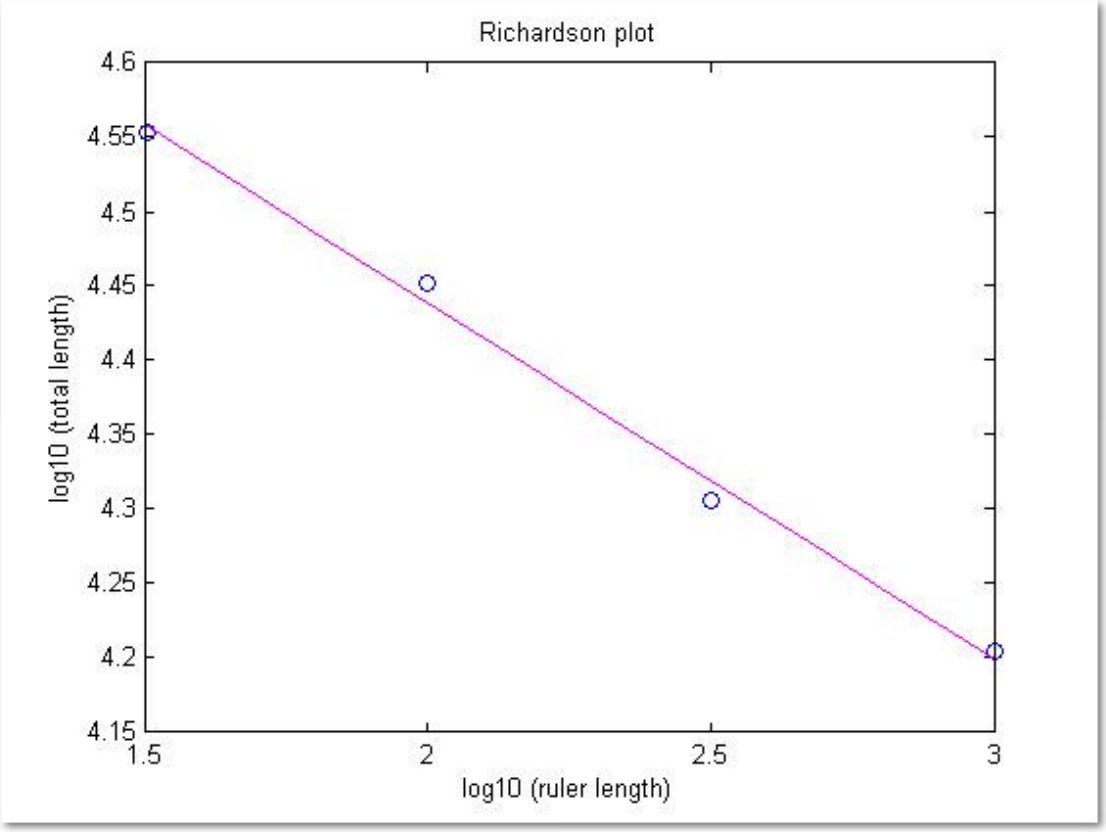

Figure 4: Double-log Richardson plot of the coastline of Allestone, showing length plotted against the spacing of length-measuring dividers, confirming the coastline's fractal properties.

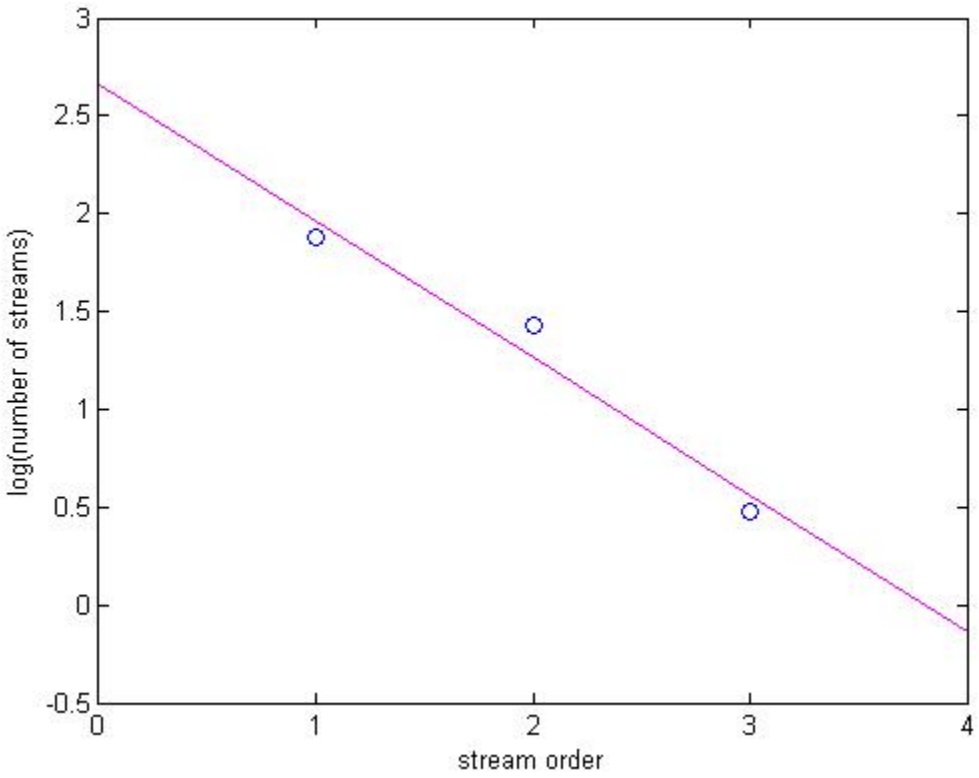

Figure 5: Semi-log plot of Allestone streams, showing a relationship between the number of streams and stream order that is consistent with Horton's Law

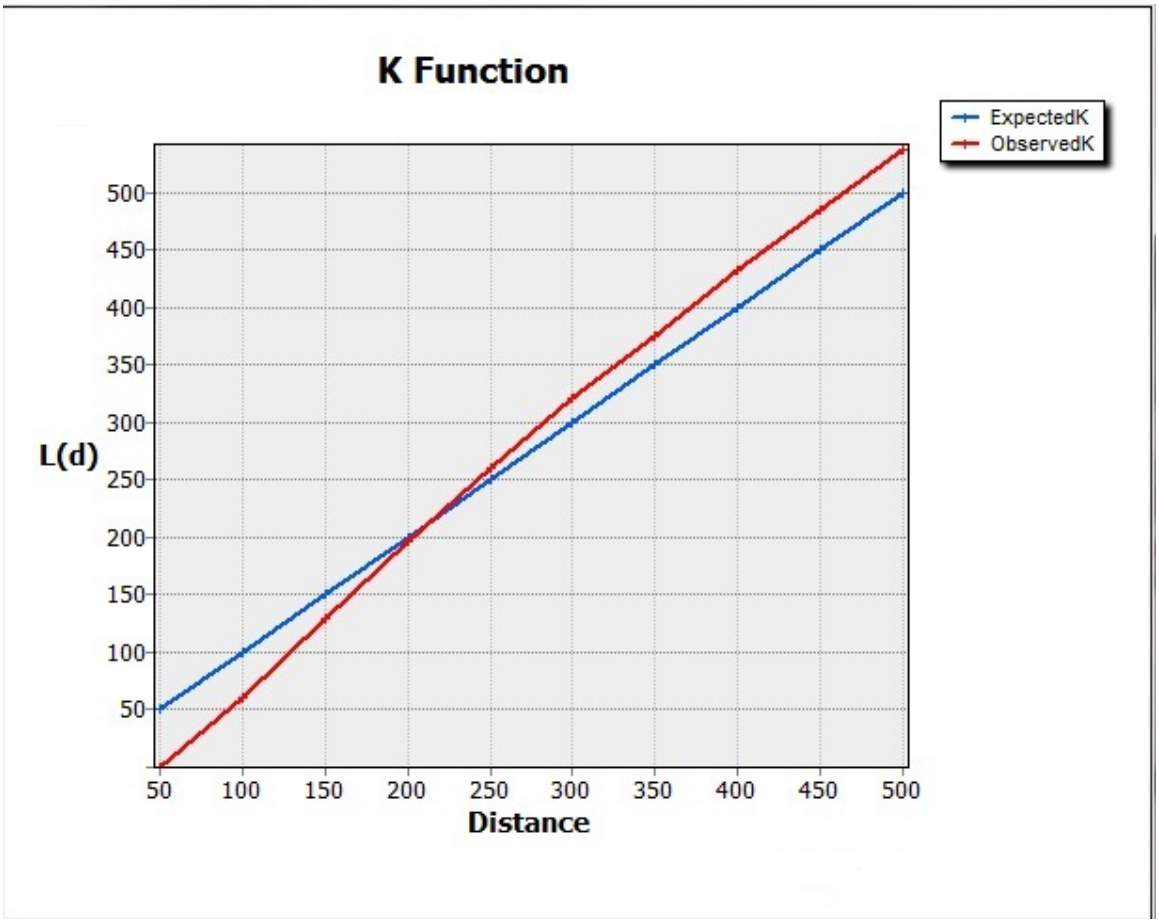

Figure 6: Ripley's *k*-function analysis of the pattern of named places on the Allestone map. The blue line represents Complete Spatial Randomness. The red line indicates that named places are further apart than expected over short distances.