

A Beginner's Guide to Geographic Virtual Communities Research

Brent Hecht

Department of Electrical Engineering and Computer Science, Northwestern University, United States

Darren Gergle

Department of Communication Studies and Department of Electrical Engineering and Computer Science, Northwestern University, United States

INTRODUCTION

Virtual communities have important geographic components. Community participants live, work, and travel to specific places on the Earth's surface, and communities often reflect the characteristics of these places. In addition, community artifacts are often imbued with geographic information.

Researchers can use these often under-appreciated geographic elements to understand important patterns in virtual communities' interaction with the real world. For instance, one could build and study a shared repository for a biking community's geographic knowledge (Priedhorsky & Terveen, 2008), investigate whether community artifact density is biased towards certain areas of the globe (Hecht & Gergle, 2009), or model the particular characteristics of a community's **spatio-social network** (Larsen, Axhausen, & Urry, 2006; Larsen, Urry, & Axhausen, 2006).

Geographic analyses can also allow an investigator to answer questions that are *not overtly geographic in nature*. In such cases, these analyses can provide an efficient alternative or supplement to more traditional methods such as large-scale surveys, interviews, or observational techniques. In many ways, it is this capability of geographical analyses that is more powerful for the virtual communities researcher. The number of research topics here are infinite, but could include modeling the relationship between social networking site usage and socioeconomic status, understanding human photo-taking behavior (Hecht & Gergle, 2010; Yanai, Yaegashi, & Qiu, 2009), modeling and sharing dynamic travel behavior based on interaction within social networks (Pultar & Raubal, 2009), and identifying self-focus bias in wikis (see the case study at the end of the chapter).

This chapter is targeted at the virtual community researcher who wants to quantitatively examine or employ the geography of a community, but has no training in the methodologies necessary to do so. We take the reader from the data collection stage through the application of several simple techniques, suggesting more advanced literature when space limitations prevent us from delving into details. We also take special care to flag important pitfalls that cause hard-to-notice but critical errors. Finally, we close with a brief but illustrative research project case study.

This chapter is effectively an introductory lesson in **Geographic Information Systems (GIS)** and **Geographic Information Science (GIScience)**, customized for the virtual communities

researcher. A GIS is a “set of tools for performing operations on geographic data that are too tedious or expensive or inaccurate if performed by hand”. In doing so, it helps “reveal what is otherwise invisible in geographic information” (Longley, Goodchild, Maguire, & Rhind, 2005b). Another definition many GIS educators find useful describes GIS as a “powerful set of tools for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real world for a particular set of purposes.” (Burrough & McDonnell, 1998) GIScience is the science and engineering behind this “set of tools”. It can be loosely considered analogous to information science but for the well-defined class of geographic information (Longley, et al., 2005b).

While GIS/GIScience and computer science are closely related, this chapter should be accessible to readers with no programming experience at all. However, programming ability (or access to someone with knowledge of programming) will help the reader more readily leverage the tools we mention for their own research. In particular, experience with web-based application programming interfaces (APIs), Java, and/or statistical programming will be useful.

MINING GEOGRAPHIC INFORMATION FROM VIRTUAL COMMUNITIES

Before engaging in any study involving the geographic component of virtual communities, it is necessary to obtain geographic information or to transform pre-existing geographic information into a “usable” form. Usable forms include **latitude/longitude** coordinates, **bounding boxes** around geographic features, and advanced **polygonal** and **polylinear representations** (e.g. the shape of the United States and the path of a road), along with the **attribute** information attached to these data, such as a username, population, etc.

Formally, geographic information is defined as “atomic pairs of the form $\langle x, z \rangle$ where x is a location in space¹ and z is a set of properties [attributes] of that location; or information that is reducible to such pairs.” (M. Goodchild, 2001; M. Goodchild, Yuan, & Cova, 2007). For example, the x in a pair could be a latitude/longitude of a city that is mentioned in a forum posting, and the z could include the average income of the city, the username of the poster, his/her centrality in a social network, and/or the size of the post (Figure 1).

This section discusses important methodologies for obtaining geographic information and making it usable for virtual communities research. We also point the reader to easy-to-use tools for applying these methodologies.

Latitude and Longitude Pairs

A growing number of virtual communities generate community artifacts that contain latitude and longitude coordinates. Assuming this structured information is accurate, it is often immediately “usable” in geographic analyses. Classic examples include the latitude and longitude (“lat/lon[g]”) tags that have been manually associated with hundreds of thousands of Wikipedia articles or online photo collections that have been manually or automatically tagged with lat/lon information. Later in the chapter we discuss challenges that can result from the poor spatial representations inherent in latitude and longitude points (such as inaccurate area and distance calculations). However, if the virtual community being studied explicitly contains lat/lon tags, a researcher can generally consider herself lucky. Geographic information in other forms (covered later in this section) is generally harder and more error prone to extract.

¹ An important topic in cutting-edge GIScience research is the inclusion of the temporal dimension, so x now usually refers to a location in *space-time*, not just *space*.

X		Z	
Location	Username	Years Active	Activity Level
61.433, -142.911	Shania Kroeger	5.2	10
44.939,-93.168	Burton Wainwright	2.4	2



X		Z	
Location	Name	Population	# of Users
	Ontario	12,929,000	5,432
	California	36,757,000	8,321

FIGURE 1: Examples of geographic information datasets. Each row represents an $\langle x, z \rangle$ pair. Note the variety of representations that can make up x (in this case there are both latitude and longitude coordinates and complex polygonal representations), as well as the diversity of possibilities for z attributes.

Street Addresses

Street addresses require a quick and relatively accurate process known as **address geocoding** before they can be used by most geographical analyses. This process, which generally turns a street address into latitude and longitude coordinates, is usually quite exact. However, the returned coordinates can sometimes contain inaccuracies about the size of a city block or the locations may be inaccurately positioned on the wrong side of a street (although this situation is improving). Google², Microsoft Bing³, Yahoo!⁴ and MapQuest⁵ all provide web-based address geocoding APIs.

² <http://code.google.com/apis/maps/documentation/services.html#Geocoding>

³ <http://msdn.microsoft.com/en-us/library/cc981067.aspx>

⁴ <http://developer.yahoo.com/maps/rest/V1/geocode.html>

Geographic Information in IP Addresses

One form of geographic information that is frequently available to virtual communities researchers is that contained within IP addresses. Through the process of **IP geolocation**, a user's location can be determined with a certain degree of precision and accuracy. Usually, the more one pays for the geolocation software, the better the precision and accuracy. One cannot expect to achieve sub-city level precision at any reasonable level of accuracy. Country-scale research, on the other hand, is generally very suited to IP geolocation.

MaxMind's⁶ free GeoLite Country, for instance, advertises 99.5 percent accuracy at a country scale (99.5 percent of country identifications are correct), while its GeoLite City package offers 79 percent accuracy for the US within a 25-mile radius (different countries may be more or less accurate). IP geolocation companies frequently offer free online "sample" versions of their software that can be used to geolocate a small number of IP addresses.

Readers should be somewhat cautious when using and interpreting IP geolocation data as some of the causes for IP geolocation inaccuracies can add significant systematic error to certain studies. For example, if you were examining a community of distributed software developers and that group of users primarily connected via a VPN (virtual private network) to their companies then you might have a bias in the results you would get back from IP geolocation.

Geographic Information in Natural Language

Very frequently, community discussions and other artifacts contain vast amounts of geographic information in the form of **toponyms**, or place names, in natural language. Yahoo! describes this information as "geographically relevant, but not [easily] geographically discoverable." (Yahoo! Developer Network, 2009)

Geotagging is the process of identifying toponyms in text and matching them with structured geographic information. It is composed of two parts – **geoparsing** and **geocoding** (a generalized form of address geocoding) (Pasley, Clough, Purves, & Twaroch, 2008) – each of which is a difficult process and can introduce error. The goal of the geoparsing process is to disambiguate toponyms from non-geographic named entities (solving "geo/non-geo" ambiguity). Consider the case of "Washington", for example. Without context, geoparsing is impossible to do, as "Washington" can be a place (e.g. "Washington Park"), a former U.S. president ("George Washington"), part of a newspaper title (e.g. "Washington Post"), etc. Natural language processing (NLP) techniques are generally used to partially solve this problem.

Once toponyms have been identified with a certain degree of accuracy, the geocoding process can begin. Geocoding associates a toponym with a **spatial footprint** of structured geographic information using a **digital gazetteer** (M. Goodchild & Hill, 2008; Hill, 2000). A spatial footprint can be a latitude and longitude point, a bounding box around a city's borders, or even a detailed polygonal representation. In other words, whereas geoparsing resolves geo/non-geo ambiguity, geocoding resolves geo/geo ambiguity. Again, "Washington" presents an interesting example. Even if we are sure that we are operating in the geographic domain, "Washington" can refer to a U.S. state, the capital of the United States, or even a street in Albany, California. Without

⁵ http://www.mapquest.com/features/developer_tools_oapi_quickstart

⁶ <http://www.maxmind.com>

additional assistance, it is not clear which footprint should be matched with the term “Washington”. The case of “London” presents similar problems.

Contextual clues can help the disambiguation process. Chances are that if a community member writes about how much she enjoys visiting the Tate Modern and Buckingham Palace on the weekends, the “London” she refers to will be that of London, England. Once this is recognized, a spatial footprint (i.e. latitude/longitude pair) for London, England can be used in a geographic analysis. However, if she writes that she is a student at the University of Western Ontario, then London, Ontario is likely correct, and London, Ontario’s (very different) spatial footprint is used.

Virtual communities researchers will often perform the entire geotagging process, but in some cases only the geocoding step is necessary. The latter is true for getting geographic information from data in necessarily geographic database fields such as the “hometown” field in Facebook. The strict typing of the field means that its value is nearly guaranteed to be a geographic entity, thus there is no geographic ambiguity and the geoparsing stage can be skipped.

Both Yahoo! and MetaCarta offer web-based APIs for geotagging. Metacarta’s GeoTagger API⁷ has the advantage of advanced natural language processing, meaning it is capable of correctly interpreting the expression “10 miles North of Phoenix” as more than just “Phoenix”. Yahoo!’s Placemaker⁸ geotagging API, however, may be more familiar to a developer already working with Yahoo!’s APIs, and is better suited to handle high volumes of text.

Generally speaking, if geocoding alone is required, either the address geocoding or geotagging web APIs can be used to extract spatial footprints. Knowing that only geocoding is needed allows the researcher to use the Google, Mapquest, and/or Bing APIs, instead of being restricted to any particular functionalities and foibles of Metacarta and Yahoo! (such as traffic limits).

Once geographic data has been collected, it is important to understand its limitations. The following section identifies the largest of these limitations for virtual communities researchers, as well suggesting tips for getting around it.

THE GEOWEB SCALE PROBLEM: ALASKA ON THE HEAD OF A PIN

Scale is a fundamental concept in the study of geographic information. Patterns observed at one scale, for instance, are not necessarily observed at other scales. In addition to the many other scale-related concerns in geographic research (such as the **ecological fallacy** and the **modifiable area unit problem**), online geographic research usually faces a distinctive scale problem: the **Geoweb Scale Problem** (GSP) (Hecht & Moxley, 2009). Stated in the virtual communities context, the GSP occurs when the spatial footprints available are at too coarse a scale for a given research problem. This can occur when the community itself embeds structured geographic information or when this information is derived using techniques such as IP geolocation or geotagging.

How does this manifest in virtual communities research? Consider a researcher aiming to uncover the relationship between the socioeconomic status of neighborhoods in Chicago with the number of Facebook users in those neighborhoods. In this case the researcher will likely run up against the GSP because Facebook users typically specify their current city (e.g. “Chicago”), and not their neighborhood (e.g. “Hyde Park”). In our work reported in (Hecht & Gergle, 2010), we

⁷ <http://ondemand.metacarta.com/?method=GeoTagger>

⁸ <http://developer.yahoo.com/geo/placemaker/>

were unable to specify the proximity of Flickr users to their photos with a precision better than 50km for the same reason.

An even nastier instance of the GSP occurs when *some* spatial footprints are encoded at an appropriate scale for a study, but others are not. The English Wikipedia, for instance, encodes all footprints as single points, including, for example, the state of Alaska's. Distance-based studies using this point will be fallacious, especially within the region. For instance, Anchorage and the state of Alaska are around 400km apart according to the English Wikipedia's spatial footprints! Similarly, any study that requires knowledge of containment relations would be impossible using this dataset. To get around this problem, (Hecht & Moxley, 2009) automatically removed the more egregiously coarse spatial footprints in Wikipedia using a list of the geographic features with the largest "real" footprints: countries and first-order administrative districts (i.e. provinces, states, etc.). Taking a similar approach, (Lieberman & Lin, 2009) assumed that coordinates not specified to a certain number of significant digits implied that the geographic features being represented were very large, and filtered them from their analysis. Another approach is to decrease the resolution of the experiment to the lowest common denominator resolution, which is the method described in the case study below.

If you do geographic virtual community research long enough, chances are you will run into the GSP. Unfortunately, there is no easy solution. The two approaches used in the literature are either to (1) redefine your study around the spatial representation limitations of your data or (2) filter your data to remove the most egregious cases. At the very least you need to be aware of this potential problem and think critically about how your study or usage of geographic information can be affected.

PROJECTIONS: YOU KNOW THE EARTH ISN'T FLAT, BUT DO YOUR TECHNIQUES AND METHODS?

It is our hope that most people reading this chapter are aware that the Earth is not flat. However, it is surprising how often this piece of common knowledge gets overlooked in the analysis of geographic data by researchers naive to traditional geography, cartography, and related fields. In order to use latitude and longitude points (or other types of spatial footprints encoded in latitude and longitude), it is essential to fully understand the implications of the shape of the Earth on geographic analyses, especially those done at a global/continental scale and/or those that require great precision.

In order to represent the Earth's surface on a flat plane – such as on a map or a regular grid – distortions must necessarily be introduced. For centuries, geographers, cartographers, mathematicians and others have examined ways to manage these distortions in order to optimize the functionality of planar Earth representations for specific tasks. A vital component of these optimized representations are **projections**.

However, with the invention of GPS, geo-tagging, and Google Earth, centuries of expertise and knowledge have been unwittingly ignored as researchers and practitioners from many fields naively attempt geographical analysis, entranced by these new technologies. Careful attention to projections (and coordinate system issues in general) is a necessary step and unfortunately one that is often skipped. In the place of projection expertise has arisen a "knee-jerk" reaction: considering the Earth's surface to be an *accurate* Cartesian coordinate system with longitudes as the x-coordinates and latitudes as the y-coordinates. A flat earth assumption is inherent to this approach.

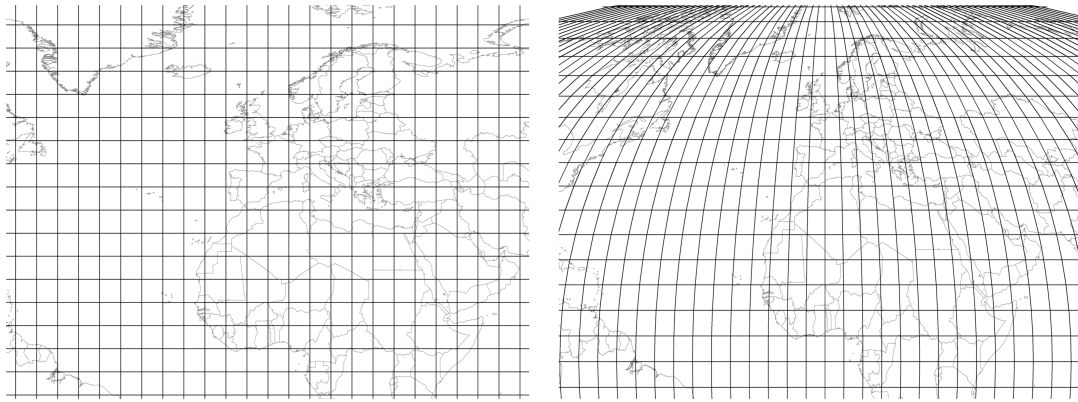


FIGURE 2: In the unprojected projection on the left, the latitude and longitude grid seems to set up “pixels” of identical area across the globe. However, it can be easily seen in an equal area projection like the Mollweide Projection (right), that this is not actually the case. Units of square lat/lon degree are much smaller near the poles than at the Equator because lines of longitude get closer and closer together as they approach the poles.

Geographers have long called this flat-Earth latitude and longitude “projection” the “unprojected projection”⁹ and have strongly cautioned against its use in analyses. Any introductory GIS textbook worth its salt will warn of the “serious problems that can occur” (Longley, Goodchild, Maguire, & Rhind, 2005a) when applying raw latitude and longitude coordinates in analyses. The important thing to remember for virtual communities research about the unprojected projection is that it does not preserve true area, scale, distance, or shape, particularly anywhere far from the equator (i.e. England, Germany, Canada, South Africa, etc.). As a result, most calculations one makes (such as average distance, density measures, etc.) using this projection are significantly distorted.

The most obvious corollary is that researchers who report lengths, densities or areas in units per degree or units per square degree are failing to report findings in a consistent fashion. A degree/square degree has different meanings at different latitudes. As shown in Figure 2, this due to the fact that the real-world length of a degree of longitude varies with latitude. At the equator, one degree of longitude is ~111km, but it is ~70km at 50° latitude and ~38km at 70° latitude. For reference, Berlin, Germany is at ~52°N latitude and Quito, Ecuador is approximately at the Equator (0°). As such, a square degree around Berlin is ~6,200 km² but ~12,300 km² around Quito. Similarly, research that reports distance results in latitude and longitude degrees is equally erroneous. Of course, all of these problems are their most severe for global-scale research, but regional and local analyses will be affected as well if reasonable precision is required.

⁹ Another common name is a “Geographic Coordinate System”, as opposed to a “Projected Coordinate System”.



FIGURE 3: In this screenshot from Google Maps, Greenland appears as large as Canada due to the area distortions inherent to the Mercator projection. Had Google chosen an equal-area projection, Greenland's area would have been accurately depicted as being approximately that of Mexico.

Solving the projection problem for distance calculations is easier than for area-based calculations. Google's Map API and others can calculate driving distance, which for some research problems is the preferred distance metric over straight-line Euclidean distance. For global research problems where local precision is not required, **great circle distance** is a computationally simple proxy for the minimum "as the crow flies" distance. Great circle distances, which differ extensively from Euclidean distances calculated from latitude and

longitude coordinates in nearly all cases, are derived from the same “curved” paths flown by airplanes. These paths (chords of great circles) only look “curved” because of the projection on which they are often drawn; in fact, they are the shortest paths between two points on a sphere. An Internet search will reveal dozens of great circle straight-line distance calculators in many different programming languages and forms¹⁰. Unfortunately, if local precision is required, the Earth-as-sphere assumption behind the great circle calculation becomes a problem, because the Earth is not quite spherical (see next subsection).

Area calculations require transformation of the underlying latitude and longitude coordinates into true linear coordinates (meters, km, etc.)¹¹ using an **equal area** projection. Equal area projections guarantee that “areas on the map are always in the same proportion to areas measured on the Earth’s surface” (Longley, et al., 2005a). This is in stark contrast to the unprojected projection, where an area *A* that appears larger on the map than an area *B* may actually be smaller in “real life”. All full desktop GIS software packages provide extensive projection technology. Those familiar with C can use the famous PROJ.4¹² software package, and Java programmers can take advantage of the excellent open-source GeoTools¹³ code library. GeoTools contains many of the operations of a professional GIS package, albeit only in Java code form. Finally, many statistical packages such as R and MatLab have spatial extensions that are capable of performing projections.

As an important aside, the famous **Mercator projection** is also an example of a projection that is very much *not* equal area. The Mercator projection displays Greenland, for example, as being massively larger than Mexico, but in actuality, the two are approximately equal in area. This may shock anyone who uses Google Maps regularly, as it is encoded in the Mercator projection. Google apparently failed to consult cartographers, who long ago noted that the “use of the Mercator projection for world maps should be [repudiated] by authors and publishers for all purposes” (Boggs, 1947). Of course, performing area-based analyses on data in a Mercator projection (perhaps from data that used a screenshot of Google Maps as a base map) is as problematic as using data in unprojected (latitude and longitude) form. A more appropriate projection for the globe or local areas should be used.

Readers interested in gaining more expertise in projection-related issues (and the datum-related issues discussed below) have many options. The Geographer’s Craft¹⁴ is a well-reputed (albeit a bit long in the tooth) online resource. Introductory GIS textbooks should all have at least one chapter dedicated to coordinate systems. Finally, those who crave the mathematical nitty gritty can turn to John Snyder’s classic text on projections (Synder, 1987).

Latitude and Longitude, According to Whom?

There is yet another major concern regarding the shape of the Earth that can have large effects on research projects that need local accuracy and precision. As noted above in the discussion about great circle distances, the Earth is not a true sphere. In fact, it is not even a spheroid or

¹⁰ <http://www.nhc.noaa.gov/gccalc.shtml> and <http://www.chemical-ecology.net/java/lat-long.htm> both offer easy-to-use manual circle distance calculators.

¹¹ The job of all projections (not just equal area) is converting the “angular” coordinates of latitude and longitude into “linear” coordinates with units like meter, nautical mile, kilometer, etc.

¹² <http://trac.osgeo.org/proj/>

¹³ <http://geotools.codehaus.org/>

¹⁴ http://www.colorado.edu/geography/gcraft/notes/coordsys/coordsys_f.html

ellipsoid, but has an irregular, constantly changing surface. However, for reasons of computational simplicity, the Earth's shape is usually approximated in most GIS analyses with an ellipsoidal model called a **datum**. Latitude and longitude points are always derived on a datum, and each datum is optimized in certain parts of the world. A latitude and longitude coordinate means *nothing* without knowing the underlying ellipsoidal model on which it is based. In other words, a single latitude and longitude coordinate refers to different real-world locations in different datums.

The reason readers should not panic after reading the preceding sentence is that most researchers working with online geographic data will encounter geographic information encoded in one of two datums. **WGS84** (World Geodetic Survey 1984) is the default datum in most GPS devices and web-based APIs, and therefore is the most common datum behind latitude and longitude coordinates. However, with the advent of Google Earth, a new datum has risen in popularity: the **Google Earth datum**. The Google Earth datum deviates from WGS84 due to a problem called (satellite) image misregistration. Goodchild (M. F. Goodchild, 2007) found that in Santa Barbara, California, this error will cause positioning to be off by about 40 meters. Google Earth image misregistration also affects any geographic data layer made using Google Earth as a reference.

Depending on what type of project the reader has in mind, the above two paragraphs should result in one of two reactions:

1. 40 meter error? Why do I care about 40 stinkin' meters?
2. 40 meter error! That ruins my whole project!

The key difference between these two reactions is the required precision and accuracy of the research project, as well as the ratio of the number of data points likely to be affected to those likely not to be affected. A person seeking to count how many Flickr photos' tagged latitude and longitude points lay within each country in the world will likely have the first reaction. Researchers who want to crowdsource gravestone database generation or landmine identification should be in the second camp. These researchers will also have to be extra careful about other coordinate system/projection issues (and other types of precision/accuracy concerns).

SPATIAL AUTOCORRELATION: IF YOU SMELL, IT'S LIKELY YOUR ROOMMATE WILL SMELL TOO

Statistics wonks in the readership may be familiar with temporal autocorrelation, or the tendency of observations made nearby in time to be correlated. Spatial data has an analogous property, albeit in more than one dimension. **Spatial autocorrelation** is so important to the study of geographic information that it is described in the so-called First Law of Geography¹⁵: "everything is related to everything else, but near things are more related than distant things"(Tobler, 1970).

While it is well beyond the purview of this chapter to explain this phenomenon in detail (**spatial statistics** is the field that focuses on spatial autocorrelation issues), it is important that "geo-novices" be aware of spatial autocorrelation. In particular, the virtual communities

¹⁵ While it's called a Law, geography and GIScience researchers agree that it is more of a guideline or rule-of-thumb.

researcher should know that spatial autocorrelation can cause a violation of the standard independent and identically distributed (*iid*) assumption of regression error terms. According to de Smith and colleagues, “many (most) spatial datasets exhibit patterns of data and/or residuals in which neighboring areas have similar values (*positive spatial autocorrelation*) and hence violate the core assumptions of standard regression models.” (de Smith, Goodchild, & Longley, 2009). One approach to addressing spatial autocorrelation is to use **Geographically Weighted Regression** (GWR), which allows parameters in regression models to vary across space. Another is to implement a **mixed regressive spatial autoregressive** model, which explicitly incorporates an autoregressive component, or to apply a **spatial error model**. De Smith and colleagues (de Smith, et al., 2009) provide an excellent overview of these methods and others, along with suggestions of tools that can be used to implement them. Their book is available in online form for free¹⁶.

CASE STUDY: DETECTING SELF-FOCUS IN WIKIPEDIA

In order to ground our geographic information crash course in real virtual communities research, the remainder of this chapter is dedicated to a short case study based on the paper “Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories” (Hecht & Gergle, 2009). We will of course center our attention on the geographic analyses, especially with regard to how we handled many of the issues raised above.

The goal of our study was to examine diversity in knowledge representations across many different language editions of Wikipedia. In other words, is there a global consensus emerging as to the structure and content of world knowledge, or does each Wikipedia contain large amounts of unique information? And if the latter is the case, is this unique information random, or is it self-focused (i.e. centered on the particular interests and realities of speakers of each language)? These research questions were motivated by the implicit “global consensus of world knowledge” assumption in many areas of computer science-based virtual communities research (see (Adar, Skinner, & Weld, 2009) for example). Even Wikipedia’s co-founder Jimmy Wales seems to assume that there is one single “sum” of world knowledge in his famous quote about the Wikipedia project’s end goal:

“Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That’s what we’re doing.”

We had many options in exploring this difficult research question. A typical virtual communities approach would have been to interview or survey Wikipedia authors from several different languages about the type of world knowledge they encode. However, this would be challenging given the need to deal with multi-lingual survey development, encoding and interpretation of the data, and numerous other challenges associated with global surveying. Moreover, Wikipedians are particularly averse to participating in surveys. Another approach would have involved choosing a small sample of articles in several different languages, and examining their particular characteristics. Indeed, after the publication of our article, this was done between English and Polish (Callahan & Herring, 2009) but such an approach is necessarily limited to the sample that is drawn and large-scale / global patterns are more difficult to reveal.

¹⁶ <http://www.spatialanalysisonline.com/>

However, by using the geographic information embedded in many Wikipedia articles, we realized that we could reduce the amount of error-prone human labor, as well as drastically increase the number of languages and articles studied. We ended up examining data from around 8.9 million articles in 15 different Wikipedia language editions. Because hundreds of thousands of these articles are tagged with latitude and longitude coordinates, we could identify the location on the Earth at which these articles exist. We were able to use this information to answer questions such as “Do Russian-speakers tend to write more (relatively) about Russia than anyone else?” and “Do Finnish-speakers blab on and on about Finland relative to Spanish-speakers?” We formalized these inquiries in the geographical analyses that follow.

Before describing these analyses in detail, however, we must highlight an important subtext to the above discussion. One of the much under-appreciated aspects of geographic information is that it can help researchers investigate non-geographic topics. This is particularly true in virtual communities research, where geographic information can provide a unique analytical lens to examine otherwise difficult or impossible questions. Our research question about the diversity of world knowledge representations was in no way explicitly geographic. However, through the use and analysis of geographic information, we were able to provide stronger evidence and expend fewer resources than with a non-geographic approach.

Geographic Data

As noted above, the location component of our geographic information (the x) was the latitude and longitude coordinates embedded by Wikipedia contributors into hundreds of thousands of Wikipedia articles. Of course, the only articles in which a lat/lon tag make sense are those that have a permanent and specific footprint on the surface of the Earth, which we call “explicitly geographic Wikipedia articles”. For instance, explicitly geographic articles include “University of Saskatchewan”, “Toronto”, and “Golden Gate Bridge”. Articles without lat/lon tags are those like “Stephen Colbert”, “Diet Coke”, and “iTunes”.

As noted above, Wikipedia’s latitude and longitude tags are a canonical example of the Geoweb Scale Problem. Latitude and longitude tags are inherently zero-dimensional, while some of the entities described in Wikipedia are quite extensive one- or two-dimensional (on a map) features. It is quite difficult to accurately describe Alaska in a lat/lon point, but that does not stop Wikipedians from doing it. As such, we carefully chose our minimum scale of analysis to circumvent the GSP, a process that will be described below and is repeatable in similar virtual community work.

Geographic Analyses

We used a combination of our open-source, Java-based WikAPIdia Wikipedia analysis software, which is optimized for geographic analysis, and ESRI’s **ArcGIS** software¹⁷. ArcGIS is the industry-standard GIS package, but it is a costly piece of software. Our study could have also been performed – albeit with greater effort – using other software, such as Matlab or R (with their spatial extensions). **GRASS GIS**¹⁸, the most popular open-source GIS software, would have also been possible, but GRASS is notoriously difficult to use. Finally, GeoTools (Java) was another option.

¹⁷ <http://www.esri.com/>

¹⁸ <http://grass.itc.it/>

First, using WikAPIdia, we exported all latitude and longitude tags into the **Shapefile** file format, which is a GIS industry standard¹⁹. We created a separate shapefile for each of the 15 languages. Like all geographic information data formats, shapefiles allow both the storage of location (x) and attribution information (z). In our case, the x was the latitude and longitude pairs, and the z was a measure of how much the article located at each pair was “being written about.” We found that one simple way to quantify the somewhat abstract idea of “being written about” is to use the indegree – or number of inlinks – for each article, because when an author of a given Wikipedia article a links to an explicitly geographic article b , the author must necessarily be writing something about the topic of b in article a . In the end, each of our 15 shapefiles contained a listing of lat/lon coordinates (x) for every explicitly geographic article (in a language edition l) paired with the indegree in l (z) of each of those articles. We also included additional attributes (z), such as article title, in order to help us visually inspect the data.

It was then necessary to aggregate all this information into summary statistics for some set of spatial features that are comparable across all languages. Articles themselves are not comparable because the vast majority of explicitly geographic articles do not exist in all 15 languages. The first concern in our aggregation was to choose a unit that was appropriate given the GSP. This meant that we had to choose first-order administrative districts (states, provinces, etc.) or larger, due to the Alaska problem mentioned above. Had we chosen a smaller unit – counties for example – the article for the state of Alaska would be considered to be within the county²⁰ that the lat/lon tag for Alaska happens to fall within. In the end, we performed our analyses at two scales: first-order administrative district-scale and country-scale.

Similarly, but less obviously, had we decided to use a grid of geographic pixels²¹ – a common choice for researchers new to geographic information – pixels smaller than the state of Alaska would fail to solve the GSP. In general, where possible, it is best to use real spatial units that have inherent semantic meaning to the research question (e.g. states, counties, countries) rather than pixels. This can be done using the **Point-In-Polygon** (PIP) or spatial join algorithms in any of the GIS or GIS-capable software packages mentioned above and geospatial data that is usually available in ESRI’s Shapefile or Google’s **KML** file format (from stakeholder websites²² or via a web search).

Once we executed the aggregation, we were able to perform both statistical and visual analyses of the results. We will leave the rather detailed statistical analyses to readers who download the paper, but the visual reporting both elucidates the power of geographical analyses and presents an opportunity to briefly touch upon appropriate cartographic techniques for reporting these types of results.

¹⁹ Here we used GeoTool’s Input/Output packages

²⁰ Geography trivia sticklers in the readership will note that counties are called “boroughs” in Alaska.

²¹ The geographic pixels methodology refers to dividing up the geographic study area into arbitrarily-sized square area units (i.e. 10km-by-10km).

²² The U.S. Census (<http://factfinder.census.gov>) and/or Statistics Canada (<http://www.statcan.gc.ca/>) are good places to start looking.

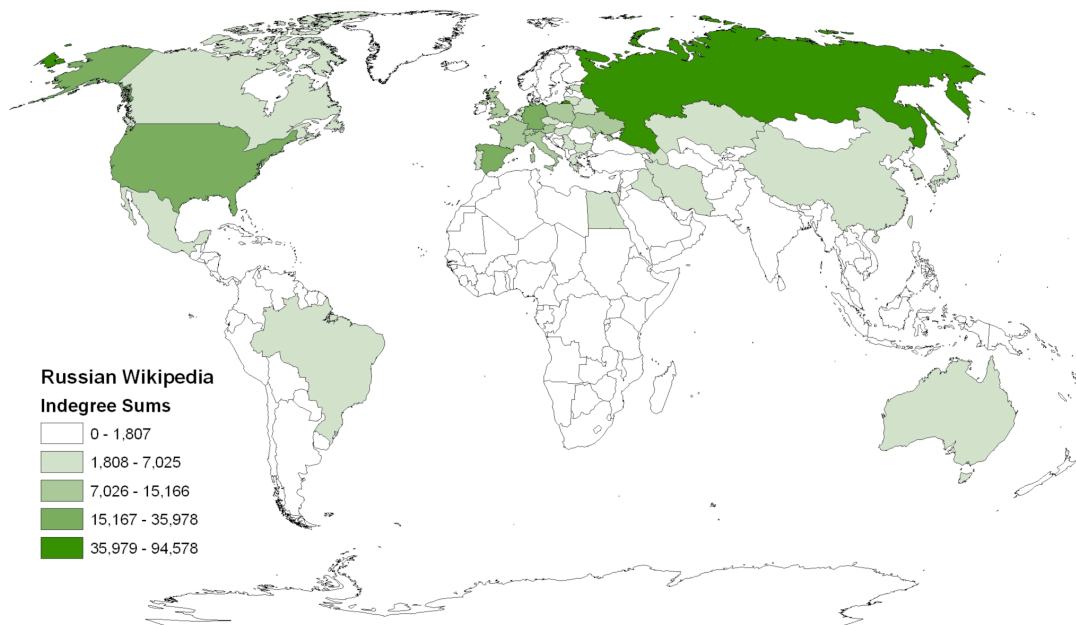


Figure 4: Number of links to articles about places in each country in the Russian Wikipedia.

Figure 4 shows the rather extreme nature of our results: Russia is the destination of the most links in the Russian Wikipedia (by far). This was repeated across nearly all 15 languages. In order to truthfully convey the results of our study in map form (Figure 4 appeared in our paper), we made absolutely sure that our data classification strategy accurately represented our findings. A cartographic novice or an expert manipulator could easily exploit the map's legend to naively or unscrupulously alter the reader's impression of the data, especially given the lesser-known units of "inlinks". It is also possible through naïveté to produce maps that are simply very difficult for the reader to interpret. Before producing a **choropleth** (i.e. colored-polygon) map, it is important that the researcher be familiar with the standard methods of data classification (e.g. **quantile**, **natural breaks**, etc.). Many websites²³ provide good tutorials on this topic. However, consulting a GIS or cartography textbook, (e.g. (Slocum, McMaster, Kessler, & Howard, 2009) or reading the entertaining "How to Lie With Maps" (Monmonier, 1996) is of course a more complete solution.

Hopefully, through this case study the reader has gained a greater understanding of how geography can enable exciting virtual communities research. Readers should also be able to repeat many of the steps above in their own work.

NEXT STEPS: WHERE TO GO FROM HERE

In this chapter, we have covered what we believe to be the minimal information required to begin examining virtual communities with a **geographic lens**. However, this chapter is by no means a replacement for a solid GIS course series. The majority of major universities (and many community colleges) will have at least one GIS course available. There are also online courses

²³ Statistics Canada provides an excellent overview at:
http://atlas.gc.ca/sitefrancais/english/learningresources/cartocorner/map_content_carto_symbolology.html

offered by universities such as Pennsylvania State²⁴, which is well known in GIScience circles, and GIS software companies²⁵. Finally, a growing number of universities including Harvard, UC Berkeley and UC Santa Barbara offer geographic analysis consultation centers in the vein of academic statistics consulting.

ACKNOWLEDGEMENTS

We wish to offer special thanks to Dr. Martin Raubal (UC Santa Barbara, Geography), Dr. Emilee Rader (Northwestern University, Technology and Social Behavior), and Dr. Holly Barcus (Macalester College, Geography) for their invaluable comments and suggestions. We also thank the anonymous reviewers of this chapter for their feedback.

REFERENCES

- Adar, E., Skinner, M., & Weld, D. S. (2009). *Information Arbitrage Across Multi-lingual Wikipedia*. Paper presented at the WSDM '09: Second ACM International Conference on Web Search and Data Mining.
- Boggs, S. W. (1947). Cartohypnosis. *The Scientific Monthly*, 64(6), 469-476.
- Burrough, P. A., & McDonnell, R. A. (1998). *Principles of Geographical Information Systems*. New York, NY: Oxford University Press.
- Callahan, E., & Herring, S. C. (2009). *Cultural Bias in Wikipedia Content on Famous Persons*. Paper presented at the AoIR 10.0: Internet Research 10.0.
- de Smith, M. J., Goodchild, M. F., & Longley, P. A. (2009). Data Exploration and Spatial Statistics *Geospatial Analysis* (3 ed.). Leicester, UK: Matador.
- Goodchild, M. (2001). *A Geographer Looks at Spatial Information Theory*. Paper presented at the COSIT '01: Conference on Spatial Information Theory 2001.
- Goodchild, M., & Hill, L. L. (2008). Introduction to Digital Gazetteer Research. *International Journal of Geographical Information Science*, 22(10), 1039-1044.
- Goodchild, M., Yuan, M., & Cova, T. J. (2007). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21(3), 239-260.
- Goodchild, M. F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69(4), 211-221.
- Hecht, B., & Gergle, D. (2009). *Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories*. Paper presented at the Communities and Technologies 2009: Fourth International Conference on Communities and Technologies, University Park, PA, USA.
- Hecht, B., & Gergle, D. (2010). *On The "Localness" Of User-Generated Content*. Paper presented at the CSCW 2010: 2010 ACM Conference on Computer Supported Cooperative Work.
- Hecht, B., & Moxley, E. (2009). *Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge*. Paper presented at the COSIT '09: 9th International Conference on Spatial Information Theory.
- Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints *Research and Advanced Technology for Digital Libraries*. Berlin / Heidelberg, Germany: Springer.

²⁴ <http://www.worldcampus.psu.edu/GISCertificate.shtml>

²⁵ <http://www.gis.com/education/online.html>. These educational opportunities are provided by ESRI, which sells the famous, powerful, and rather expensive ArcGIS software.

- Larsen, J., Axhausen, K., & Urry, J. (2006). Geographies of Social Networks: Meetings, Travel, and Communications. *Mobilities*, 1(2), 261-283.
- Larsen, J., Urry, J., & Axhausen, K. (2006). *Mobilities, Networks, Geographies*. Aldershot, England: Ashgate.
- Lieberman, M. D., & Lin, J. (2009). *You are where you edit: Locating Wikipedia users through edit histories*. Paper presented at the ICWSM '09: 3rd International Conference on Weblogs and Social Media.
- Longley, P., Goodchild, M., Maguire, D., & Rhind, D. (2005a). Georeferencing *Geographic information Systems and Science* (Second ed.).
- Longley, P., Goodchild, M., Maguire, D., & Rhind, D. (2005b). Introduction *Geographic Information Systems and Science* (2nd ed., pp. 4-33). West Sussex, England: John Wiley & Sons, Ltd.
- Monmonier, M. (1996). *How to Lie with Maps* (Second ed.). Chicago, IL: University of Chicago Press.
- Pasley, R., Clough, P., Purves, R. S., & Twaroch, F. A. (2008). *Mapping geographic coverage of the web*. Paper presented at the ACMGIS '08: 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.
- Priedhorsky, R., & Terveen, L. G. (2008). *The computational geowiki: what, why, and how*. Paper presented at the CSCW '08: The 2008 ACM Conference on Computer Supported Cooperative Work.
- Pultar, E., & Raubal, M. (2009). Progressive Tourism: Integrating Social, Transportation, and Data Networks. In N. Sharda (Ed.), *Tourism Informatics: Visual Travel Recommender Systems, Social Communities, and User Interface Design* Hershey, PA: IGI Global.
- Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2009). *Thematic Cartography and Geovisualization* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Snyder, J. P. (1987). *Map Projections - A Working Manual*. Washington, D.C.: U.S. Geological Survey.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.
- Yahoo! Developer Network (2009). Yahoo! Placemaker Beta Beta. Retrieved July 20, 2009, from <http://developer.yahoo.com/geo/placemaker/>
- Yanai, K., Yaegashi, K., & Qiu, B. (2009). *Detecting Cultural Differences using Consumer-Generated Geotagged Photos*. Paper presented at the LocWeb '09: Second International Workshop on Location and the Web in Conjunction with CHI 2009.