Analysis of Web-Usage Statistics to Evaluate the Effectiveness of Online Maps for Community Outreach

Howard Veregin

AJ Wortley

ABSTRACT: The Wisconsin State Cartographer's Office (SCO) relies on Web technology to facilitate its community outreach mission of supplying geospatial information to professionals and the general public in the state. Of particular importance to this effort are the SCO's Web-mapping applications that provide online access to aerial photographs, geodetic control points, and PLSS (Public Land Survey System) corner points. This study employs Web-usage statistics to assess the effectiveness of these online applications in meeting the SCO's outreach goals. Specifically, the study focuses on the WHAIFinder (Wisconsin Historic Aerial Image Finder) application that serves up high-resolution scanned aerial photographs from the 1930s. The key questions addressed in the study include the following. Is the application reaching all areas of the state equally? Is it reaching communities in remote areas far from the physical photo collection? Are there regions of the state that are under-served? Is usage correlated to ease of access to the physical photo collection or the availability of a broadband Internet connection? The source of Web-usage data is Google Analytics, a free Google service that generates a host of usage statistics, including information on the locations of Web site visitors. The hypotheses and methods described in the study are aligned to work in Participatory GIS (PGIS), especially efforts to evaluate the effectiveness of online mapping applications for community engagement.

KEYWORDS: Web-based cartography; Web-usage; Data mining; Community outreach; Participatory GIS

Introduction

Since the 1990s the Wisconsin State Cartographer's Office (SCO) has relied on Web technology to help achieve its community outreach mission. As prescribed in state statute, the SCO's responsibilities include: cataloging geospatial data holdings throughout the state; coordination between local, state and federal mapping agencies; disseminating information about geospatial programs and activities; disseminating information about innovations in mapping and data collection methods; and publishing special maps and information ("University of Wisconsin System", 2012). These efforts are achieved in part through Web-based information delivery via the SCO Web site (www.sco.wisc.edu) and the Wisconsin Mapping Bulletin (www.sco.wisc.edu/news.html).

Equally important are the SCO's Web mapping applications, which evolved out of more traditional paper-based publications, maps and catalogs over the SCO's history (Veregin and Kennedy, 2012). These applications deliver geospatial data and metadata for aerial photographs, geodetic control, the Public Land Survey System, and other themes (www.sco.wisc.edu/online-tools/online-tools.html). This study focuses on the

WHAIFinder (Wisconsin Historic Aerial Image Finder) application, which provides online access to over 38,000 scanned aerial photographs of Wisconsin from the 1930s and 1940s (www.sco.wisc.edu/whaifinder/whaifinder.html). WHAIFinder was developed under a grant from the Baldwin Wisconsin Idea Endowment (www.provost.wisc.edu/baldwin.htm) awarded to the SCO, the UW-Madison Department of Geography's Robinson Map Library (www.geography.wisc.edu/maplib), and UW Digital Collections Center (uwdcc.library.wisc.edu) (Veregin et al., 2010). The goal of the project was to increase public access to a rare collection of 1930s-era historic aerial photographs covering the entire state. Desired outcomes of the project included increased use of the collection by making it available online, and increased ease of access by individuals physically far-removed from the paper collection at UW-Madison.

The WHAIFinder application, as well as the SCO's other online mapping applications, are technology-based methods that have the potential to reach a relatively large audience at low marginal cost. Once an application is developed and deployed, it can in theory be accessed by anyone with a broadband Internet connection at any time of day. In short, one of the biggest advantages of Web applications is that they have the potential to reach a larger audience at lower cost than traditional methods. This includes communities dispersed throughout the state, including remote areas far from physical data collections.

This study represents an effort to quantify whether these potential advantages are actually being realized. The study examines WHAIFinder Web-usage statistics to assess how effectively the application reaches its target community in the state. Specific questions include: Is there a correlation between WHAIFinder usage and overall population, as would be expected if the application was reaching all areas of the state equally? Are there areas of the state that are being under-served relative to population? Is usage concentrated in certain regions of the state? Does access to broadband Internet affect the relationship between visitors and population, as might be expected given the large download file size for WHAIFinder images? Does physical distance impact WHAIFinder usage, due to friction of distance effects noted by other researchers?

The study emphasizes mapping and analysis of spatial patterns in Web-usage statistics. This is in contrast to many other studies of Web usage, which have an e-commerce focus, e.g., using data mining to optimize Web site topology. The rationale for the study is the need to better understand how well the community is being served by the SCO's Web applications, with the ultimate goal of evaluating ROI (return on investment) associated with the cost of Web application development and maintenance. Such evaluations are especially important in today's environment of declining resource availability and increased demand for accountability.

Previous research

In a broad sense, this study is aligned with work in Participatory GIS (PGIS) on community-based online mapping. While there is no single definition of PGIS, it is often equated with the process of local participation and the integration of local knowledge in spatial decision-making. This includes activities such as participatory map-making that allow community groups and individuals to represent their views spatially. The PGIS literature has focused considerable attention on understanding the role of Web-based geospatial technology to empower citizens, and on analyzing and defining the concept of empowerment itself (e.g., Elwood, 2002; Corbett and Keller, 2005).

From a broader perspective, PGIS refers to a continuum of levels of citizen engagement, from basic information delivery on one end to collaboration and empowerment on the other. Schlossberg and Shuford (2005) describe different types of engagement in a matrix of possible public-participation outcomes. Likewise, Garau (2012) lists five levels of public participation, from informational to empowerment, and notes that the instruments used in a given PGIS initiative depend on the level of participation one wants to attain. These studies highlight the broad range of activities that fall under the rubric of PGIS. However, one common characteristic of all PGIS initiatives is that they often face the same set of challenges, including technical hurdles (e.g., lack of access to broadband Internet), conceptual limitations (e.g., lack of understanding of geospatial technology and concepts), and biases inherent in the technology itself (e.g., inability to use the technology to record diverse ways of understanding space).

The present study focuses on the first set of challenges, i.e., technical factors that affect Internet usage. As noted by Kyem and Saku (2009), lack of Internet connectivity reduces citizen participation opportunities. This is particularly acute for Web-based geospatial technology due to its data-intensive nature. Likewise, physical distance may drive down participation rates. It is commonly believed that the Internet has essentially eliminated friction-of-distance effects. However, empirical work challenges this idea. For example, Wang et al. (2003) investigated the correlation between physical distance and network performance for US higher education Web sites. They found distance to be positively correlated with latency (length of time for data to travel through the network), especially at distances below 1000 km. In short, distance is an important factor in determining access time for the Web.

Methods

Web usage statistics

The source of Web-usage data for this study is Google Analytics (www.google.com/analytics/). This service is offered for free for sites that generate up to 10 million hits per month. Google Analytics allows Web-site managers to analyze traffic volume and trends, identify visitor locations, and evaluate patterns of interaction. Google Analytics works by incorporating a snippet of JavaScript code in a Web page. The code activates Google Analytics tracking, which collects and sends usage statistics to the Google Analytics report site. Users then login to the site to view and download usage statistics. Google Analytics generates time-series data for a variety of metrics, including visitors, new and returning visits, referrals, pageviews, and visit duration. In addition, these statistics can be aggregated by geography at the country, state and city level.

The discussion below provides a brief explanation of various usage metrics and their definitions as they apply to the present study.

- Pageview. A pageview is calculated every time a page on the Web site loads. If a user visits a page, and then reloads the page, two pageviews will be counted. This is one of the more crude usage indicators, and is conceptually quite similar to the idea of a "hit". The main problem with the pageview statistic is that it can significantly overestimate actual usage.
- Visit. A visit (or session) is a period of interaction between a Web browser and a Web site. When the browser is closed or is inactive for a certain period of time, the visit has ended.
- Visitor. A visitor (or unique visitor) is essentially synonymous with a single user. A visitor is identified by a Google Analytics visitor cookie, based on a concatenation of a random visitor ID and a timestamp of the visitor's first visit, which creates a unique ID for that visitor.

While a visitor closely approximates a single user, there are differences. In particular, the count is hardware dependent, such that a single user would be counted as two separate visitors if the user accessed the Web site on different computers. Also, Google Analytics tabulates visitors uniquely over specific time periods and geographies. In other words, the count of visitors over a year-long time period may be smaller than the sum of the counts of visitors for each of the twelve months in that year, because Google Analytics is independently tabulating the number of visitors for the specific time frame requested. Likewise, the sum of visitors tabulated over all locations within a given time period may be larger than the count of visitors for that same time period, because Google Analytics is counting the number of visitors at each location independently.

Metrics describing traffic sources are based on the following distinctions:

- Direct traffic results from typing the URL of the Web site directly into a browser, or clicking on a bookmark to arrive at the site.
- Referring sites are other Web sites that send traffic to the Web site being tracked by Google Analytics (e.g., through an embedded link).
- Search engine traffic is traffic generated through a search engine, such as Google or Bing. This traffic can be organic (free) or cost-per-click.

Traffic sources identify how users are discovering and accessing the site. By cross-tabulating with other metrics, it is possible to assess which sources lead to the longer visit durations and other indicators of engaged usage.

Hypotheses

The specific hypotheses of this study are as follows.

1. Is there a correlation between WHAIFinder usage (number of visitors) and overall population? If the application is being accessed uniformly throughout the state, the expectation is that visitors and population will be positively correlated. A follow up question is whether there are areas of the state that are under-served.
2. Does access to broadband Internet affect this relationship? If population is weighted by level of access to broadband, does the correlation between visitors

and population increase? This effect is expected due to the large file size of WHAIFinder imagery, which puts lower access speeds at a disadvantage.

3. Does physical distance impact WHAIFinder usage, due to friction of distance effects? In the WHAIFinder context, users physically further removed from the data servers (on the UW-Madison campus) may experience slower responses, resulting in lower levels of satisfaction and use. Compounding this effect, it is possible that users closer to the physical collection of air photos (the Robinson Map Library at UW-Madison), who are familiar with the collection and have used it in the past, may be more knowledgeable about online access methods due to promotional materials available in the Robinson Map Library.

Google Analytics data

The main usage statistic employed in this study is the number of visitors, since the main interest of the study lies in correlating usage against population. Number of visitors is the best surrogate for actual number of unique users, which is related to population counts more closely than other statistics such as pageviews. To generate visitor data it was necessary to run a customized report in Google Analytics. The number of unique visitors was calculated for a date range of June 15, 2011 to June 14, 2012 (a single year) with location as a secondary variable. Google Analytics provides statistics on visit location down to the "city" level; information is derived by mapping visitors' IP addresses to geographic locations. For the WHAIFinder, Google Analytics identified almost 2500 cities for the selected year-long date range. Approximately 65 percent of these locations are outside of Wisconsin and have relatively small numbers of visitors (often just one). These were excluded from the analysis to focus attention on the state of Wisconsin.

The majority of the Wisconsin locations were identified as incorporated places (cities and villages). However the list also includes unincorporated places, as well towns (or townships) which have legal status as a local government entity in Wisconsin. After inspection it was determined that city-level statistics were not appropriate for analysis. In particular, it was observed that for some counties only one or two of the larger cities within the county were listed as the source of all visits. These cities often had a visitor count far exceeding what would be expected based on the size of the city. This is probably due to the way that IP addresses are assigned to geographic locations. Commercial products such as MaxMind's GeoIP® GeoLite database (www.maxmind.com/app/geolite) advertise an accuracy of 78 percent at the city level in the US, within a 40 kilometer radius. A circle with this radius is over 5000 square kilometers in area, which is much larger than the average Wisconsin county.

To dampen this effect, location information was generalized by computing the sum of visitors at the county level. Visitor counts for all cities, villages, townships, and unincorporated places within each county were summed to arrive at a final total visitor count for each of Wisconsin's seventy-two counties. Two sparsely-populated counties (Iron and Florence) reported no visitors, bringing the final count to seventy observations.

Broadband data

Broadband data were obtained from the National Broadband Map (NBM) initiative (www.broadbandmap.gov). NBM provides access to data and maps on broadband availability across the United States. It was created by the National Telecommunications and Information Administration (NTIA) in collaboration with the Federal Communications Commission (FCC) and the US states and territories. The NBM is part of NTIA's State Broadband Initiative (SBI) that provides funding to state entities or non-profits to integrate and map statewide broadband data ("State Broadband Initiative", 2012). In Wisconsin the SBI is known as LinkWISCONSIN and is administered through the Public Service Commission of Wisconsin ("The LinkWISCONSIN Initiative", 2012).

Data collection methodologies for the NBM are complex, with much of the necessary information coming from broadband service providers. A variety of broadband statistics are tabulated, including maximum upload and download speeds. Broadband speed refers to the amount of data that a user can download or upload per second. It is usually measured in kilobits per second (Kbps) or megabits per second (Mbps). Broadband speed determines how quickly a Web site will load or how much time it will take to download a file. The advertised maximum speed represents the maximum speed that the network can support; observed speed is typically slower than the network's theoretical maximum.

For the NBM, broadband is defined as a high-speed, "always-on" connection to the Internet, providing two-way data transmission (i.e., upload and download) with advertised speeds of at least 768 Kbps for download and 200 Kbps for upload. The NBM differentiates between wireless and wireline technologies; this study focuses on only the latter. NBM data are tabulated by census block (as well as by road segment and address point for larger blocks). The Census block is the smallest unit of geography for which the US Census Bureau tabulates population data. Many blocks, particularly in metropolitan areas, have multiple broadband speeds available due to the presence of more than one service provider. For the purposes of this study the main impact of broadband speed is likely to be the time it takes to download a scanned WHAIFinder image. The highest resolution images (600 dpi) are approximately 30 Megabytes (MB) in size. At a download speed of 768 Kbps such an image would take over 5 minutes to download.

For this study, data were obtained for download speeds of 3 Mbps or faster. While 768 Kbps defines the lower limit of broadband access, 3 Mbps is more in line with the FCC's National Broadband Plan as a desired base level of access. Moreover, access to 768 Kbps download speeds is not a differentiating factor, as it is available to 95 percent or more of residents in all Wisconsin counties. The percentage of county population with access to advertised maximum download speeds of 3 Mbps or faster was obtained using the NBM analysis tool (www.broadbandmap.gov/rank). Percentages were then converted to total population counts by county.

Other data

Population data for Wisconsin counties were obtained from the Wisconsin Legislative Technology Services Bureau (LTSB) which processes US Census data for use by the Legislature and its service agencies (legis.wisconsin.gov/ltsb). To test distance effects, driving distance between Madison and the location of each county seat was computed. Driving time was thought to best represent the actual distance that a visitor to the Robinson Map Library would need to travel to utilize the physical collection. County seats were selected given their importance as the locations of local government, as well as generally being large population centers within each county.

Results

Basic Statistics

The WHAIFinder application generated 21,109 pageviews and 16,497 visits over the one year time frame of the study. The number of unique visitors is 9725, or approximately 27 visitors per day, or about 0.17% of Wisconsin's population. Breaking down visitors by source reveals that 68.2% of visitors are referrals, i.e., they arrive at the application from a link on another site. Nearly half of these visitors (41.2%) are referred by the WHAIFinder landing page on the SCO site (sco.wisc.edu), while another sizeable group (22.1%) are referred from the Robinson Map Library site (on geography.wisc.edu). Direct visitors, who enter the WHAIFinder URL directly into their browser or use a bookmark, are also very common (27.8% of visitors). The low referral rates for search engines (4.0%) suggests that most visitors are already aware of the WHAIFinder application.

Spatial patterns of usage

Figure 1 shows visitor counts summed for each Wisconsin county. The highest number of visitors occurs in Dane county, which contains the state capital (Madison) as well as the UW-Madison campus. Figure 2 shows visitor density, i.e., visitor count divided by county population. The data are expressed in visitors per 1000 population. Some interesting spatial patterns are evident, including high densities in Dane county, the far north of the state, and the La Crosse and Eau Claire areas. Lower densities occur around Milwaukee and the eastern edge of the state. However, Moran's I statistic for this distribution (I = 0.00045) based on county adjacency is not significant at the 95% level.

Regression analysis: Population

To assess the relationship between visitors and population, bivariate regression analysis was performed using population as the independent variable and number of visitors as the dependent variable. The best results were obtained with a double-log regression. Results are shown in Table 1.

The overall model is statistically significant, and the slope coefficient is significantly different from zero. The fact that the slope coefficient is positive indicates that the number of visitors increases with population at the county level. Since the regression model is based on the natural logs of visitors and population, it can be expressed as

$$\text{Visitors} = 0.00193\ \text{Population}^{0.937}$$

This equation shows that the number of visitors increases with population but a declining rate, since the population exponent of 0.937 is less than one. This finding is analogous to the allometric relationships that have been observed between population and other variables for urban settlements (see Veregin and Tobler, 1997).

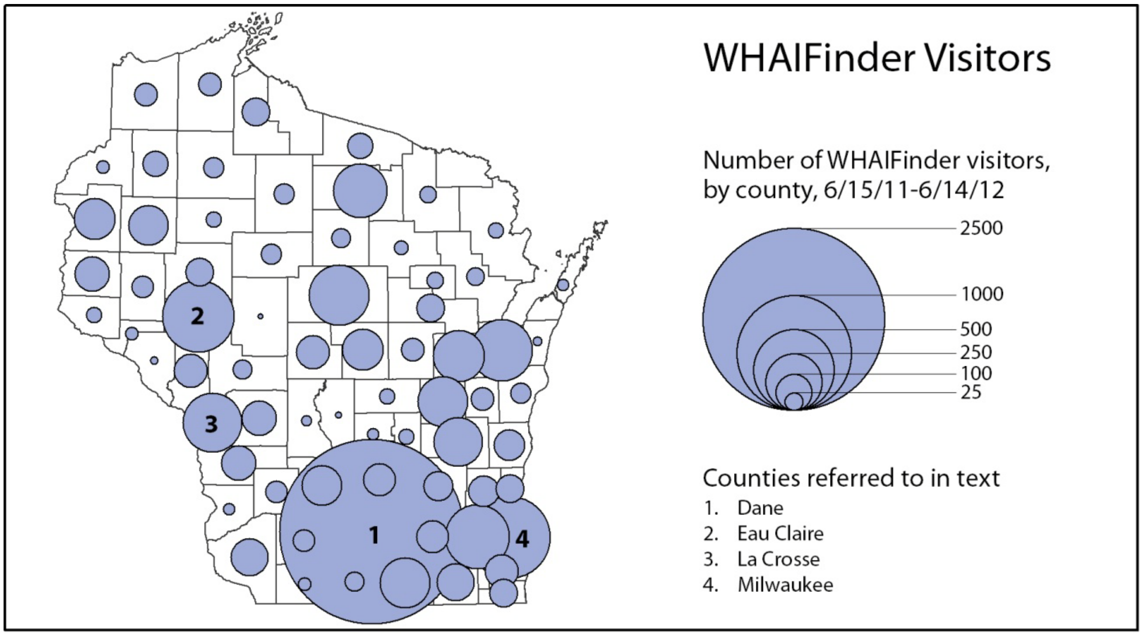

Figure 1: WHAIFinder visitors by county.

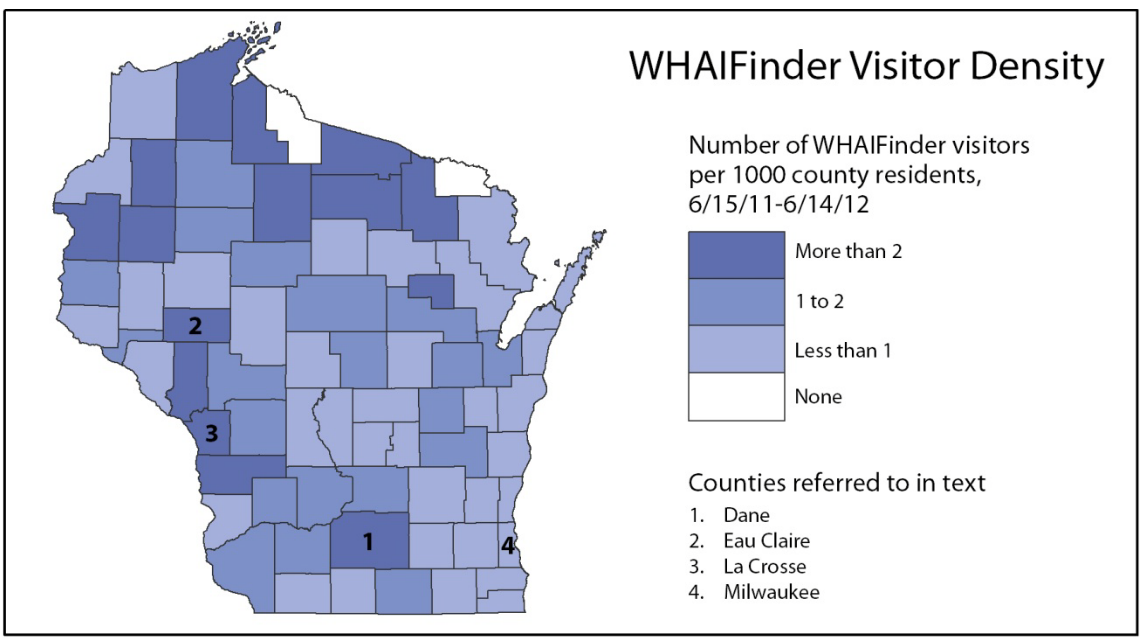

Figure 2: WHAIFinder visitor density by county.

Table 1: Regression results, WHAIFinder visitors vs. county population.

Coefficient	*Estimate*	*t-value*	*P* > |*t*|
Intercept	-6.249	-5.807	< 0.001
Slope	0.937	9.361	< 0.001
$R^2 = 0.563$ N = 70 F = 87.636 P < 0.001			

Regression residuals were examined to determine if any patterns existed. Figure 3 is a map of these residuals. The main features of the map are as follows:

1. In general, under-represented counties (fewer visitors than expected based on population) and over-represented counties (more visitors than expected) are distributed throughout the state fairly randomly. A test of spatial autocorrelation in the residuals was performed using the Moran *I* statistic, resulting in an observed value of -0.000329 (not significant at the 95% level).
2. Some patterns are visible nevertheless. Most notable is Dane county, which has the highest over-representation of visitors. This is probably because Dane county contains many state agency offices and the UW-Madison campus.
3. Two other highly over-represented counties are Eau Claire and LaCrosse, which also have large UW campuses. The higher than anticipated use in these counties may be due to greater awareness of the application due to classroom use, or the presence of a link to the application on a county Web site. Visitor source statistics show that Eau Claire county's Web site was responsible for referring 422 visitors (over 6 percent of all referrals).
4. Many of over-represented counties are in the northern parts of the state, far removed from the physical collection of data. This may reflect the greater benefits to using the online application rather than the physical collection at UW-Madison, given the long driving times to Madison for these counties.
5. There is a pattern of under-representation (fewer visitors than expected) along the eastern edge of the state, especially in Milwaukee county. It is unclear what this trend represents, although it has been suggested that southeastern Wisconsin is somewhat independent of the rest of the state in terms of geospatial activity (Day and Ghose, 2012). In addition, several of the counties in this region have made county-wide mosaics of the 1930s imagery available on their Web sites, which may be a more attractive data source than the WHAIFinder for users.

Regression analysis: Broadband access

To assess the effects of broadband access, county population figures were weighted by the percentage of county population with access to broadband Internet with speeds of at least 3 Mbps. The best results were obtained for the double-log regression (Table 2).

The overall model is statistically significant and the slope coefficient is significantly different from zero. Access to download speeds of 3 Mbps or more is associated with higher visitor rates. These results support expectations. Note that the regression results are very similar, but marginally better in terms of predictive power, to the results for total population unweighted by broadband access. The regression model can be expressed as

$$\text{Visitors} = 0.00210\ \text{Access}^{0.934}$$

where “access” is the population of each county with access to average advertised download speeds of 3 Mbps or more. The equation shows that the number of visitors increases with access, but at a decreasing rate.

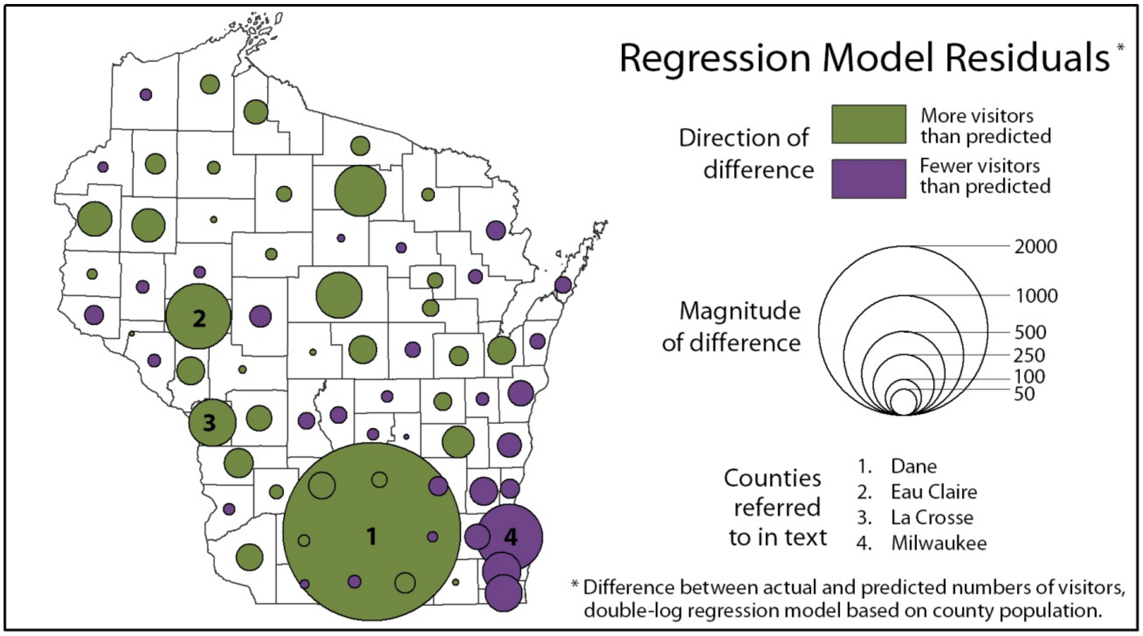

Figure 3: Residuals from regression of WHAIFinder visitors vs. county population.

Table 2: Regression results, WHAIFinder visitors vs. population with access to > 3 Mbps broadband.

Coefficient	*Estimate*	*t-value*	*P > |t|*
Intercept	-6.164	-6.092	< 0.001
Slope	0.934	9.874	< 0.001
$R^2 = 0.589$ N = 70 F = 97.500 P < 0.001			

Regression residuals were examined to determine if any patterns existed. The map is quite similar to that in Figure 3. A test of spatial autocorrelation in the residuals was performed using the Moran I statistic, resulting in an observed value of -0.000299 (not significant at the 95% level).

Regression analysis: Distance

To assess the effects of distance, county visitor density was regressed on distance to Madison. Best results are shown in Table 3. Results do not show evidence of a friction-of-distance effect; the slope coefficient and the model as a whole are not highly significant, and predictive power is very low.

Table 3: Regression results, WHAIFinder visitor density vs. distance to Madison.

Coefficient	*Estimate*	*t-value*	$P > \|t\|$
Intercept	9.29×10^{-4}	3.304	0.00152
Slope	3.01×10^{-6}	1.707	0.0923
$R^2 = 0.0411$ N = 70 F = 2.916 P = 0.0923			

Conclusion

The results of this study indicate that there is a strong positive relationship between WHAIFinder visitors and county population, lending support to the idea that the application is reaching all areas of the state equally. There is some indication of under-utilization along the eastern edge of the state, but at this time the cause is not known. There are some isolated pockets of higher than anticipated usage throughout the state, perhaps associated with greater awareness of the WHAIFinder application due to classroom use or the presence of a link on a county or city Web site. There is little evidence that either broadband access or physical distance exerts an effect on this basic relationship. Weighting county population by access to broadband speeds of 3 Mbps or greater marginally improves the predictive power of the regression model, while driving distance does not exhibit a significant correlation with visitor density.

Web usage statistics offer a cost-effective way for organizations like the SCO to assess the effectiveness of data and information delivery. Such assessments are especially important in an era of shrinking budgets and tightening resources, in order to assess ROI and identify where information outreach is successful and where it can be improved. Analysis of Web site usage statistics allows for quantification, statistical analysis and mapping, rather than just anecdotal evaluation. As such the information can be used to set outreach and information dissemination goals (e.g., reaching a certain percentage of the community) and help assess whether such goals are being met (Turner, 2010).

The analysis discussed in this study represents a small fraction of what can be done with additional usage statistics and other analysis methods. For example, assessment of the quality of referrals (as evinced by visit duration) for different referral sources could be used to identify those sources that drive the most high-quality traffic. Such analyses may be useful to researchers in PGIS and other fields who are interested in assessing the effectiveness of geospatial technology for community access and engagement.

References

Corbett, J. M. and Keller, P. C. (2005) An analytical framework to examine empowerment associated with participatory Geographic Information Systems (PGIS). *Cartographica*, 40, 4, pp. 91-102.

Day, P. and Ghose, R. (2012) E-planning through the Wisconsin Land Information Program: The contexts of power, politics and scale. *International Journal of E-Planning Research*, 1, 1, pp. 75-89.

Elwood, S. (2002) GIS use in community planning: A multidimensional analysis of empowerment. *Environment and Planning A*, 34. 5, pp. 905-22.

Garau, C. (2012) Focus on citizens: Public engagement with online and face-to-face participation – a case study. *Future Internet*, 4, 2, pp. 592-606.

Kyem, P. A. K. and Saku, J. C. (2009) Web-based GIS and the future of participatory GIS applications within local and indigenous communities. *The Electronic Journal on Information Systems in Developing Countries*, 38, 7, pp. 1-16.

The LinkWISCONSIN Initiative. Retrieved July 2012 from http://www.link.wisconsin.gov/lwi/default.aspx

Schlossberg, M. and Shuford, E. (2005) Delineating "public" and "participation" in PPGIS. *URISA Journal*, 16, 2, pp. 15-26.

State Broadband Initiative. Retrieved July 2012 from http://www2.ntia.doc.gov/sbdd

Turner, S. J. (2010) Website statistics 2.0: Using Google Analytics to measure library website effectiveness. *Technical Services Quarterly*, 27, 3, pp. 261-278.

University of Wisconsin System. In Wisconsin Statutes & Annotations. Retrieved July 2012 from https://docs.legis.wisconsin.gov/statutes/statutes/36

Veregin, H., Gorman, P. C., Stoltenberg, J., Wortley, AJ and Bricknell, M. (2010) An open source web application for historic air photo display and distribution in Wisconsin. *Proceedings, AutoCarto 2010*.

Veregin, H. and Kennedy, T. (2012) Online information dissemination at the Wisconsin State Cartographer's Office using map services and APIs. In M. Peterson (Ed.), *Online Maps with APIs and Mapservices*. (forthcoming)

Veregin, H. and Tobler, W. R. (1997) Allometric relationships in the structure of street-level databases. *Computers, Environment and Urban Systems*, 21, 3/4, pp. 277-90.

Wang, Y., Lai, P. and Sui, D. (2003) Mapping the Internet using GIS: The death of distance hypothesis revisited. *Journal of Geographical Systems*, 5, pp. 381-405.

Howard Veregin, Wisconsin State Cartographer, Department of Geography, University of Wisconsin-Madison, 550 N. Park St., Madison, WI 53706. Email veregin@wisc.edu

AJ Wortley, Senior Outreach Specialist, Wisconsin State Cartographer's Office, Department of Geography, University of Wisconsin-Madison, 550 N. Park St., Madison, WI 53706. Email lwortley@wisc.edu