Geolinguistics: The Incorporation of Geographic Information Systems and Science

Shawn Hoch
Children's Health Services Research
Indiana University School of Medicine
Indianapolis, IN 46202
E-mail: shoch@indiana.edu

James J. Hayes
Department of Geography
California State University Northridge
Northridge, CA 91330
E-mail: james.hayes@csun.edu

ABSTRACT

Modern geographic information systems (GIS) and its incorporated spatial analysis tools allow sophisticated and efficient analysis of spatial data by researchers in many fields. Although the field of linguistics has long been of interest to geographers and spatial variation of language to linguists, researchers have made little use of the power of GIS and GIScience theory to address hypotheses regarding spatial variation of language and correlated physical and social variables. Discussion of modern GIS tools for spatial analysis, quantitative analysis, and cartography in geolinguistics has been largely absent from the literature. Linguists have applied GIS technology in language atlases, including recent on-line atlases; however, analytic and data processing capabilities are seldom discussed. Following a review of geolinguistics work incorporating GIS, this article discusses potentially useful GIS tools and techniques for geolinguistics. The article concludes with reflection on the future role of GIS in geolinguistic thought and practice.

Key Words: dialectometry, geolinguistics, GIScience, GISystems, linguistic geography

INTRODUCTION

Geolinguistics is an interdisciplinary field that often incorporates language maps depicting spatial patterns of language location or the results of processes that lead to language change. Accordingly, GIS is well-suited for geolinguistic studies, although researchers have yet to fully explore the potential for data management and analysis tools incorporated in GIS software. In a review of the literature on geolinguistics, we found few studies either employing GIS or discussing methodology for doing so (Lee and Kretzschmar 1993; Williams and Van der Merwe 1996; Goebl 2006). Also, we found few studies that acknowledge early advances in the use of GIS to examine language variation (Pederson 1993;

Kretzschmar and Schneider 1996; Kretzschmar 2003). Although researchers have developed GIS methods for spatial language data analysis, they do not often cite the history and progress of this development in the geolinguistics literature.

Linguists have produced extensive cartographic work, most notably in the form of linguistic atlases (Kurath et al. 1939-1943; Pederson et al. 1986; Labov, Ash, and Boberg 2006). GIS has undoubtedly played an increasing role in spatial data analysis and cartographic methods for linguistic data in recent geolinguistics research; however, we suggest that a more open discussion of, and focus on, the role of GIS in geolinguistics would further benefit spatial linguistics and GIScience. Geolinguistics is poised to adapt GIS and the fundamentals of geography and cartography to address both well-developed and new questions within the field.

Despite early definitions of geolinguistics as inherently interdisciplinary (Van der Merwe 1992) or even as a subdiscipline of geography (Williams 1988), there remains great potential for mutually enriching collaboration between geolinguists and GIScience practitioners. Lee and Kretzschmar (1993) described infrequent contribution of geographic expertise to linguistics research beyond the purposes of cartographic support, noting the absence of quantitative spatial analysis methods in previous work of linguistic geographers. Their call for collaboration was elaborated with examples and discussion of the use of GIS to analyze data from the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS) database (Lee and Kretzschmar 1993). Williams (1996) also described the relationship between linguistics and geography as slow to develop, pointing to their differing academic cultures. It appears that these calls have been largely unanswered as evidenced by the paucity of subsequent research. Since these publications, GIS has developed substantially in quantitative and visual spatial analysis, as well as in its further democratization. Given these advances, we see an opportunity for geography to re-engage this historically important, but currently quiet, area of geographic inquiry.

The aim of this article is to highlight discussion in the literature that *does* specifically address GIS methodology used in geolinguistic research and map making, and to reflect on the relationship between theory and method in geolinguistics and GIScience. We do not seek to present a comprehensive overview of the use or function of GIS in geolinguistics research, but rather to highlight cartographic products, research articles, and books which have explicitly discussed the role of GIS in their production. We begin by reviewing some early applications of spatial data analysis in the field, many of which took place in the formative stages of both GIScience and contemporary geolinguistics. In this section, we also address the linguistic atlas as a traditional product of cartographic methods in geolinguistics and we note its advances towards incorporation of GIS. After reviewing recent applications and ongoing projects, we aim to invigorate the discussion initiated by Lee and Kretzschmar (1993) by suggesting GIS tools potentially useful to the geolinguist.[1]

GEOLINGUISTICS: FOUNDATIONS OF SPATIAL ANALYSIS OF LANGUAGE

Early theoretical studies indicated the field of geolinguistics is rich with questions and challenges that can be approached with GIS. Breton described the process through which geographic thought becomes a tool for linguists: "In analyzing the distribution in space and in society of the facts of language, the linguist employs the methods of geography: cartography and the establishment of correlations and causalities between spatial phenomena" (1991, 19). Breton's model indicated that linguists have engaged geographic thought throughout the development of geolinguistics, especially those interested in dialectology, phonology, word choice, and the more overarching areas of language change, contact, function, history, and policy.

Given the long-established ties between linguistics and geography, what potential questions in geolinguistics can geographic information systems and science address? Mackey (1988) began to pose questions of geolinguistics which find potential solutions in GIS, asking the reader to consider the meaning of language boundaries in cartographic representation. Do borders represent transitions between languages or dialects? Do they represent zones of conflict or thriving multilingualism? Ormeling (1992) suggested that boundaries should represent the course along which the largest number of sociodemographic and physical characteristics diverge. Kretzschmar (1992) and Davis (2000) framed much use of isoglosses, boundaries delineating diverging linguistic features, as conceptual models rather than statistically reliable figures. Mackey (1988) also pointed out that language mapping should take into account the various functions and sociological aspects of language such as education and commerce. Through such questions, Macauley (1985), Mackey (1988) and others began early conversations on geolinguistic analyses such as language border measurement before the tools to conduct them were readily available outside of GIS specialist circles.

EARLY GIS APPLICATIONS: REALIZED BENEFITS OF COMPUTERIZED LINGUISTIC DATA

How have GIS applications traditionally assisted in geolinguistic research when used? What were the immediate appeals of computerized linguistic data? Though the examples are few, evidence suggests the introduction of computer technology for storage of survey data and production of linguistic atlases beginning in the mid-1970s. Researchers during this period commonly cited benefits of data storage and transport (Pederson 1986; Alvar 1991; Nerbonne and Kretzschmar 2003) and mapping on the fly (Pederson 1988; Kretzschmar 1996). Alvar (1991) composed a collection of writings on linguistic atlas projects and on the field in general. He highlighted the *Atlas Lingüístico y Etnográfico de la Provincia de Santander* (Linguistic and Ethnographic Atlas of the Province of Santander) as an example of an "automated" linguistic atlas and extolled the advantages of a computerized versus manually drawn and reproduced atlas. Alvar (1991) described the database developed for this atlas as a highly useful product of the project facilitating mapping on-demand and the preparation of indices used in interpretation of linguistic atlases. The end result was a leap forward in time- and cost-effectiveness of atlas design and reproduction.

Thomas (1980) presented an early example of GIS used to measure spatial autocorrelation in computerized data from a linguistic survey, describing how he placed numerical values representing Welsh word usage in appropriate regions on a base map of Wales. He then used a specialized grid overlain on mapped survey sites to reveal "site clusters" based on the rate at which survey results coincided with those of neighboring units. In his explanation of the process, Thomas expressed the need for a more advanced spatial analysis than his "relative geographical disposition of sites": "Ideally, enquiry sites would have been located in the cells of a regular geometrical grid superimposed on a geographical map, with the closeness of its mesh adjusted according to population density and the frequency of settlements" (13). Here Thomas alluded to the advantages of GIS raster analysis and vector grid capabilities that would be readily available to a language mapping project today.

Throughout the 1980s, linguists heralded the increasing availability of desktop computing as a benefit to geolinguistic work in attribute storage and recall (Pederson 1986, 1988) and in providing easily generated maps as research tools (Pederson 1988; Alvar 1991). These were early indicators of the vital roles of some basic GIS functions in addressing significant limitations in managing and displaying large linguistic survey datasets. However, linguistic techniques benefiting from computation, in dialectology in particular, were still hampered in their development and acceptance due to limitations of the technology available (Kirk and Kretzschmar 1992; Nerbonne and Kretzschmar 2006). Moreover, early examples of work using GIS had to endure the transition from hard copy cartography to digitized base maps (Kirk and Kretzschmar 1992). In spite of these limitations, Pederson's resourceful efforts represent early advances in geolinguistic visualization of survey data with multiple variables and quantitative measurement of word frequency. Citing inspiration by Thomas (1980), Pederson's work towards computerized storage and display of data from the *Linguistic Atlas of the Gulf States* (Pederson et al. 1986) calls to mind some essential tools of GIS that would not be widely commercially available in a graphical user interface until nearly a decade later. In establishing the visual arrangement of ASCII characters representing informant positions and responses (e.g., uses "soda" or does not; represented as "+" or "-," respectively), Pederson (1986, 1988) placed the characters as close as possible to the known locations on a base map, essentially manually geocoding the informant locations. He also employed a sequence of ASCII characters at the geocoded locations displaying several sociolinguistic attributes of informants or multiple phonemic or lexical variants (e.g., race/education/income represented as the string R-E-I) at one time. This innovation allowed storage and display of multiple linguistic attributes, albeit limited in the latter by the readability of strings of multiple characters.

The linguistic atlas has proved a vital tool and product of geolinguistics since the earliest stages of the field and has provided a stage for the incorporation of GIS. French linguist Jules Gilliéron is considered the pioneer of the linguistic atlas, having coauthored the *Atlas Linguistique de la France* (1902-10). Henceforth, linguists have produced thematic language maps and atlases of various regions. The atlas has traditionally been the starting point for research and progress in the formation of geolinguistics as a field.

An ongoing linguistic atlas project in which GIS has played a prominent role is the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) (McDavid and O'Cain 1980). Origins of the current LAMSAS project can be found in some of the earliest large-scale linguistic mapping efforts in the United States (Kurath et al. 1939-1943). Schneider and Kretzschmar (1989) began to report data organization of LAMSAS enabling computerized statistical testing and the creation of a grid optimized to contain equal numbers of respondents in a cell for the purposes of analyzing linguistic variation and regional characteristics. In the following years, their work with LAMSAS continued towards geographical analysis, commenting on the use of MapInfo in which they mapped coordinates of atlas informants (Kretzschmar and Schneider 1996). They observed the modifiable areal unit problem (MAUP) arising from their grid of irregularly shaped polygons. As Gotway and Young (2002) suggested, further exploration of GIS tools and geovisualization could help analyses in projects such as LAMSAS address the modifiable areal unit / change of support problems.

One of the most frequently referenced collections of language data, and a widely consulted source in the formation of other atlases, is the Ethnologue (Gordon 2005). The project was initiated and is maintained by SIL International, an organization originally concerned with biblical translations in minority languages. For over fifty years, the Ethnologue has appeared in numerous editions primarily as an authoritative directory of living languages, the locations of their speakers, and basic speaker population statistics. For nearly as long as it has been published, it has also included maps of countries and linguistic regions. With recent editions available online (http://www.ethnologue.com), it has added basic data exploration capabilities insofar as the user can call up maps of countries and regions by clicking on their respective links. SIL has also collaborated with a vector data resource called the *World Language Mapping System* (WLMS), making WLMS boundaries and attribute data comprising the Ethnologue available for purchase in GIS-ready formats.

RECENT APPLICATIONS AND PROJECTS INCORPORATING GIS

Some recent applications of GIS in linguistics have begun to work towards greater ease in data exploration. Whereas early linguistic atlases offered little in the way of data exploration, an example of a more fully and intentionally interactive linguistic atlas is the Modern Language Association (MLA) Language Map. Designed using ESRI's ArcIMS, the MLA Language Map compiles vast amounts of U.S. Census data (MLA 2009). The user is able to produce and manipulate thematic maps by choosing various language distributions. One can also vary the region being mapped (U.S. or individual states), and the enumeration units (counties or zip codes). The user also has access to the data tables providing languages spoken and numbers of speakers by state and county. Clearly, the MLA Language Map offers a degree of flexibility in language data representation that would be beneficial for users of all large-scale language data projects such as the Ethnologue.

One of the most recent disseminations of data from the aforementioned LAMSAS is maintained online as part of the "Linguistic Atlas Projects" (www.lap.uga.edu). The site hosts survey data from this and several other atlas projects developed in the U.S. in the early- to mid-20th century. It currently allows the user to browse the survey areas by state, with survey locations geocoded and linked to informant descriptors and responses. The site has been accessible in other versions since the mid-1990s and represents a long-standing resource for visualization of some of the most influential work in the field.

The work of Van der Merwe arguably set the stage for a subsequent generation of interactive linguistic atlases such as those discussed above while also providing a role for GIS beyond visualization. In his analyses of Cape Town (Van der Merwe 1993), he began by establishing enumeration units based on neighborhood subdivisions throughout the area and compiling multiple years of South African census data for these units. He frequently used spatial measures of central tendency to display center of gravity shifts in English, Afrikaans, and Xhosa throughout the area.

Williams and Van der Merwe (1996) went on to combine their experience in spatial language data analysis with theories concerning informed language policy in linguistically complex urban environments. The authors described the overall goal of their work as the compilation of comprehensive, dynamic, and up-to-date geolinguistic data to assist in sound decisions in education, urban planning, and language policy. They argued that overly simplistic data and a general lack of interdisciplinary geolinguistic work had left national-level planning at a loss with little accurate data on changing language use, resulting in planning and policy that was out of touch with urban realities. They pointed out that notions of national-level language patterns prior to the early 1990s simply omitted the linguistic complexities of urban South Africa, which comprised over half of the population. They offered GIS-based analysis at the localized urban level as an answer to the problems associated with a coarser regional perspective.

GIS also played a key role in a study of language use and the state of bilingualism. McGuirk (2004) explored the roles of several demographic data, their associations with language use, and implications for the future of bilingualism in Miami-Dade County, Florida. Of chief concern was the issue of language maintenance in a country that has historically assimilated immigrant cultures such that multilingualism represents only a provisional phase in the process, eventually resulting in increasingly monolingual (English-speaking) generations. Williams (1988) addressed the role of place in settings where speakers must navigate socially constructed rules of using more than one language. This importance is reflected in one of McGuirk's central research questions: “What sociolinguistic characteristics…make Spanish-English bilingualism and Spanish language vitality unique within the Miami-Dade County social and geographic context?” (McGuirk 2004, 8)

McGuirk began his geolinguistic analysis by aggregating census tracts based on established neighborhoods such as Little Havana, mapping these units and linking census data using manifold.net’s Manifold System 5.50. After performing multiple regressions, he used University of Illinois Spatial Analysis Laboratory’s GeoDa to produce choropleth maps displaying the same units with a color scheme based on the Moran Local Indicator of Spatial Association statistic (Anselin 1995). He then compared the results to those from San Diego County, California, a community with a comparably large immigrant Hispanic population. In his conclusion, McGuirk (2004) noted how these geographic analyses helped to confirm the sociolinguistic uniqueness of the target area in that Spanish speakers in Miami-Dade County were not clustered in socioeconomically deprived areas as found elsewhere.

Some recently presented work offers a rare example of the results of a vibrant relationship between geographers and linguists, and in particular demonstrates how GIScience can advance language mapping techniques. Using LAMSAS data, Thill et al. (2008) applied a self-organizing map (SOM) algorithm to assign informants to geospatial clusters, exploring the emerging patterns and comparing them with U.S. dialect regions defined by Kurath (1949) decades earlier which had been untested by empirical studies until recently. SOM algorithms offer a form of exploratory data analysis which reduces the dimensionality of a spatially referenced dataset and re-displays the data in a desired number of classes. Although the authors noted that SOM techniques are far from straightforward and must be painstakingly tailored to unique datasets, their work with this tool, bridging advanced geographical analyses and linguistics, was in itself a commendable step forward for geolinguistics. In the following section, we continue to discuss methods and possibilities that spatial science can offer geolinguistics given similar collaboration and forward-looking techniques.

SPATIAL THEORY AND METHODOLOGY APPLIED IN GEOLINGUISTICS

While the latest technological innovations of geolinguistic study are making use of the data handling and display capability of GIS, there is infrequent evidence of adoption of the analysis and cartographic functionality offered through modern GIS tools. Early examples of spatial language analysis exist, but there are limited signs that researchers have carried them forward with the advance of tools and methods. Relative to the body of geolinguistics literature, there are few published examples of how quantitative spatial methods can be applied to geolinguistic research questions. Understanding the role of space and distance and their relationships to other variables is a key component to understanding any phenomenon that plays out over a geographic area. Explicitly considering their effects can reveal important relationships that affect linguistic processes (Nerbonne and Heeringa 2007).

There are four “broad areas” of geographic information analysis that are relevant for geolinguistic research: spatial data manipulation, spatial data analysis, spatial statistical analysis, and spatial modeling (O’Sullivan and Unwin 2003). Familiarity with the theory and methodological limitations of each is critical to its use and here we see great potential for interface between geography and linguistic science. GIS and the geographical approach offer geolinguistics researchers many possibilities for advancing and reexamining theory, hypotheses, and data visualization. GIS and GIScience can offer an articulation of spatial theory as a framework for approaching hypotheses in linguistics research. In addition, GIS can simply make much research in geolinguistics faster and easier.

Much language mapping still uses choropleth maps; however, traditional choropleth mapping has two distinct disadvantages for dialect mapping. First, discrete polygon boundaries (usually political boundaries) are incompatible with modern geolinguistic theory (Mackey 1988; Dahl and Veselinova 2005). The boundaries (isoglosses) between areas of language usage (mapping units) are not discrete, but rather are features defined by gradual changes in a number of variables including dialect, ethnicity, and location (Girard and Larmouth 1993). Linguistic boundaries are therefore more like the boundaries of climatic regions or forest types (Mark and Csillag (1989). Second, each area on the map is required to belong to one and only one class, but many points on a linguistic choropleth map will share some affinity with nearby classes and areas. Several geospatial techniques have been developed to address these cartographic boundary problems.

Points on any classification map of language variation will have some probability of belonging to multiple classes. Assigning points and areas outright to discrete classes therefore increases both spatial and attribute error in the map. Traditional choropleth and isoline mapping techniques ignore both the nature of the dialect boundary and the complex multi-attribute nature of dialect space, yet some generalization and error is necessary to make the map of use.

Techniques drawing on the cartographic communication model and emphasizing the need for balance between precision mapping and usability of choropleth maps can help address cartographic issues in language mapping. A suggested solution for maintaining spatial and attribute accuracy while accommodating spatial gradation is the graded area-class map (Kronenfeld 2005). Area-class maps do not have predefined boundaries, but boundaries based on the spatial variation in the attribute of interest itself and the probability distribution that points within the mapped area belong to a designated class (Mark and Csillag 1989). Probability surfaces of class membership (Mark and Csillag 1989) and fuzzy set membership functions (Girard and Larmouth 1993) have been used to better describe and locate class (attribute) and map (spatial) boundaries by noting variations in the rate of dialect change across space. Graded area-class maps share this basic approach to identifying class membership, but rather than drawing discrete boundaries of a rigid classification, use gradation of lightness or hue to indicate changes across space based on a multidimensional attribute space (Kronenfeld 2005). Kronenfeld (2007) introduced the idea of the categorical gradient field, implemented with categorical data in vector or TIN data models to represent transition between areas of more certain class membership (Fig. 1). Using polygon and TIN data can be an advantage with linguistic

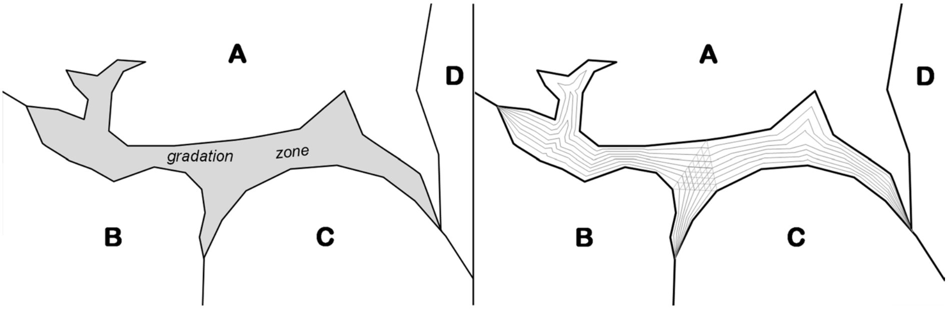

Figure 1. Illustration representing a transitional zone between four categorical classes. The gradation zone indicates an area where probability of membership is > 0 for more than one class (left). Probability of class membership may be mapped as a categorical gradient field to indicate how membership affinity varies across the transition (right).

data which are often aggregated into areal units rather than points or grids. This approach could be very useful for geolinguists when classes are determined from multiple measures of class membership and exhibit spatial gradation.

Discussions of quantitative analysis of linguistic data have been careful to include sampling bias and statistical independence (Guy 1993; Kretzschmar and Schneider 1996); however, the literature does not consistently consider spatial dependence, a potential constraint on achieving unbiased and independent samples. Thomas (1980) and McGuirk (2002) applied an understanding of spatial autocorrelation to linguistic data, but infrequent discussion of this phenomenon in geolinguistics suggests that researchers may not widely recognize its effects or are just barely exploring them. GIS makes the spatial analysis techniques related to spatial dependence more accessible than ever. Mapped linguistic similarity indices and "dialect kernels" are examples of methods for detecting "spatiolinguistic correlation" using geovisualization software ("Visual DialectoMetry") developed expressly for analyzing linguistic data (Goebl 2006). Yet, these visualization techniques can be taken further to quantitative exploration of spatial relationships.

Geostatistical methods such as semivariance (or semivariogram) analysis can be useful for better understanding linguistic variation related to spatial dependence, revealing important information about rates of change across space, language variability as a function of distance between samples, random variability in data, inter-sample distances necessary to achieve independent samples, and uncertainty in interpolated values. Semivariance analysis models variability as a function of the distance between sampling points (Burrough and McDonnell 1998). Such a model provides information on the relationship between distance and the intensity of spatial dependence between sampling locations, and the distance at which samples are independent (Rossi et al. 1992).

Kriging, a method of spatial interpolation based on semivariance analysis, may also be of use to geolinguistics. Kriging recognizes that spatial variables are too stochastic to be mapped using deterministic interpolation methods. Such variables are better represented as regionalized and having some systematic component such as a mean, but also a stochastic, spatially autocorrelated component and a random "noise" component (Burrough and McDonnell 1998). Also, kriging has the advantage of providing error estimates for the interpolated values at any point on the map. The "quantitative maps" of linguistic data described by Guy (1993) would lend themselves well to this type of analysis, potentially revealing underlying relationships and the role of space in shaping the observed patterns.

A hypothetical example of kriging is illustrated in Figure 2. The points in Figure 2A represent the centroids of neighborhoods within a city. Frequencies of informant attributes (e.g., race, educational attainment) and linguistic features are associated with each point in the spatial database. Figure 2B is an example of a prediction map created by kriging for one attribute from a survey. One can follow the same procedure for additional variables and compare the characteristics of the variograms and prediction maps to assess postulated relationships (e.g., are ethnolinguistic features coincident with patterns of segregation in the city?)

Point pattern analysis (PPA) is another quantitative approach for point data that can allow inference of patterns in linguistic phenomena (Lee and Kretzschmar 1993). Second-order nearest-neighbor PPA statistics such as Ripley's K (Ripley 1976, 1988; Dale 1999) can help identify spatial patterns that are more "clumped" or "dispersed" than a random spatial process. This approach also provides information on the scale of clumping or dispersion. The idea behind this technique is to examine a neighborhood of a given size (radius) around every point and determine if the points in that neighborhood are more or less dense than expected. The

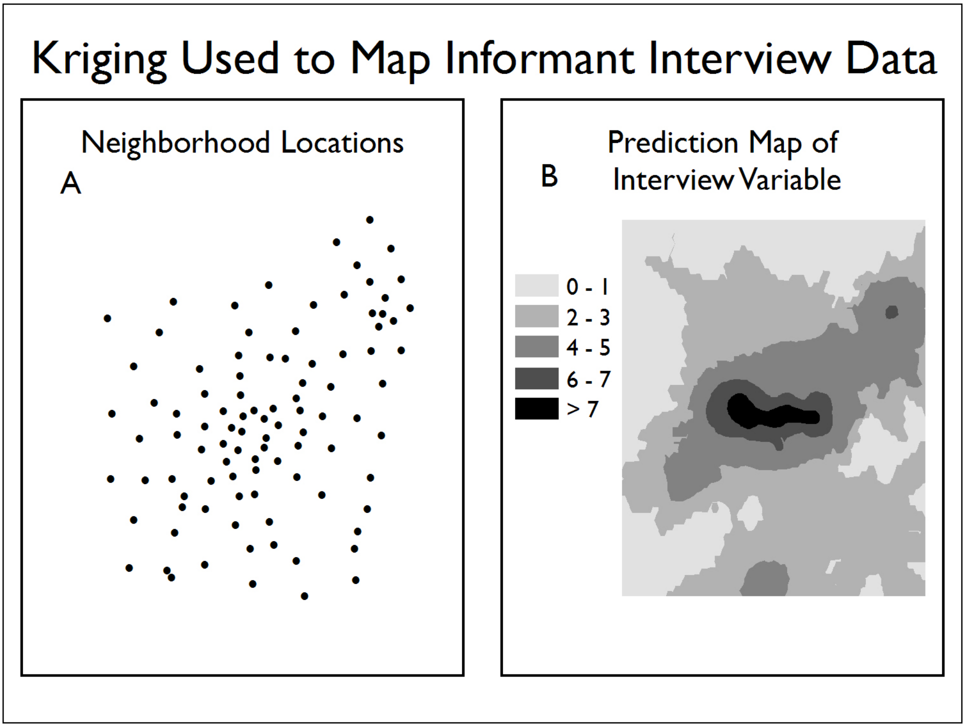

Figure 2. Example of kriging to create a continuous map from survey point data. (A) Neighborhood centroids where informants live. (B) Spatial variation of one variable from informant data estimated using kriging.

Ripley's K statistic can be used to examine spatial distributions of points for departure from complete spatial randomness (CSR) (Haase 1995). An edge-corrected transformation of Ripley's K is the *L(t)* transformation (Haase 1995). The *L(t)* statistic is calculated for all data using a given distance *t*, then repeated with sequentially larger values of *t*. Positive deviations from 0 indicate aggregation of points (clumping), while negative deviations indicate uniform dispersion. Monte Carlo simulations can be used to generate confidence envelopes of "significant" CSR deviation. Figure 3A illustrates a case of univariate application of the *L(t)* function. In the example showing hypothetical data points across the Indianapolis metropolitan area, data values within the spatially random envelope limits (dashed lines) indicate that points at those distances are distributed randomly. Data values outside of the envelope are clumped if above the envelope, dispersed if below. This example indicates that the points are clumped more than expected and that the clumping is especially pronounced at a neighborhood size of about seven kilometers in diameter.

Bivariate PPA can be of use in geolinguistic data analysis for comparing spatial distributions of, for example, two alternate pronunciations. In areas where dialects or languages intermix, bivariate PPA could be useful in determining whether the two occur together randomly, if they tend to cluster, or if they are spatially segregated. Figures 3B and 3C are examples of bivariate Ripley's K analysis. In Figure 3B the two types of points are segregated from one another at neighborhood sizes of about 12 kilometers (roughly the size of most individual clumps), but at

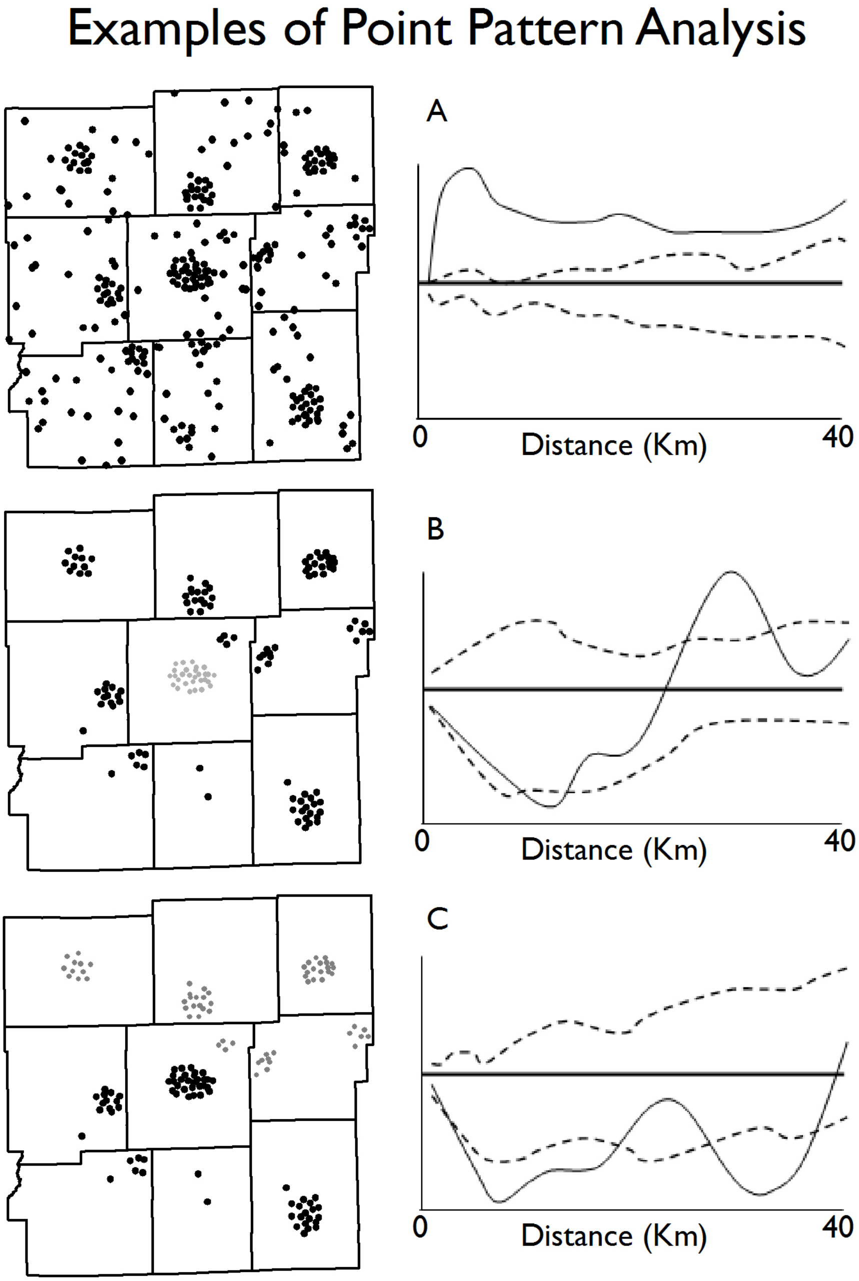

Figure 3. Three examples of PPA applied to hypothetical data. (A) Univariate example exhibiting spatial clustering. (B) Bivariate example exhibiting segregation of two responses, switching to random association before forming a cluster of aggregation. (C) Bivariate example exhibiting segregation at two different spatial scales.

a neighborhood size of about 30 kilometers the two are more clustered than a random distribution. The two types of points are randomly arranged with respect to one another at other distances. Figure 3C illustrates a case where the two types of points are segregated at short distances, randomly intermixed at intermediate distances, and segregated again at larger distances. Segregation at the shorter distances reflects the small individual clumps of points, while at the longer distances reflects the segregation of the points in the southeast half of the map from the points in the northeast.

Inferential spatial statistics can be useful for objective assessment of a spatial hypothesis, but it will not always be necessary or appropriate. Visual analysis of map data by an experienced geolinguist may be all that is necessary in some cases to identify and interpret observed spatial patterns. This can at least lead to development of new hypotheses.

GIS offers geolinguistics a range of possibilities for visualization of geographic relationships, allowing creation and comparison of multiple alternative maps with ease once the data are collected and organized. Map overlay and tools to examine spatial relationships among variables are easily accessible in most current GIS software. For geolinguistics, map overlay is likely to be concerned with the spatial coincidence among language and other variables; maps of these variables can be quickly created and compared to base language maps. Buffering is a common type of overlay technique that could be useful in examining the occurrence of language within a specified distance around particular cultural, political, or physical features. With GIS one can quickly and easily create multiple buffer maps for numerous variables to explore potential relationships. Other common overlay operations include containment, proximity, adjacency, and Boolean AND/OR and TRUE/FALSE operations. All of these visual analysis alternatives are available with little cost and effort once data entry is complete, and provide a means for efficient quantitative summaries of spatial characteristics.

CONCLUSION

This review has explored a broad scope of early and recent GIS applications in linguistics including linguistic atlases, lexical and phonological surveys, and a sociolinguistic analysis. As GIS continues to find a place in geolinguistics, some questions remain that concern GIS applications from perspectives of both parent disciplines. On a practical level, we note that geolinguists must also develop distinct approaches to storage, analysis, and display of linguistic data from the feature level, as in phonetic variation, to the largest scales found in the Ethnologue (2005) and other catalogues of modern spoken languages. This will require the expertise of linguists and the data processing and visualization skills of both disciplines.

Pederson (1995) and Kretzschmar (2006) both addressed the distinction between deductive and inductive approaches in geolinguistics. Pederson (1995) described deductive research as well-established within linguistics since Gilliéron's work, and his conceptual models depicted the process of deductive language structure investigation as beginning with known classes, then examining components of these classes whereas inductive processes imply the inverse (Pederson 1995). Kretzschmar (2006) also narrated the entrenchment of deductive approaches within American dialectology and linguistic thought more generally. Given that, as Kretzschmar explained, both deductive and inductive approaches have long-standing roles in dialectology, how might the option affect GIS applications and outcomes?

Kretzschmar (2006) also posed a related and more poignant question for geolinguistics researchers using GIS in whether future work will focus more intently on the "science" of problems and hypotheses systematically approached through technology rather than the "art" inherent in the experimentation and development of computational methods thus far. Over a decade ago, geographers posed similar questions regarding the science of the use of GISystems in general. Wright, Goodchild, and Proctor (1997) suggested "If [GIS is a tool]…significance derives strictly from the progresses made on the substantive research problem." They then offered that GIS as a *science* "is concerned with the analysis of the fundamental issues raised by the *use* of GIS in geography or in other disciplines" (Wright, Goodchild, and Proctor 1997). Moving forward, GIS can become an integral part of the science of geolinguistics while also answering Lee and Kretzschmar's (1993) call for interdisciplinary collaboration and advanced techniques.

Possibly the most salient issue for future consideration is how to facilitate the overall progress of GIS in geolinguistics. Literature on recent projects often includes comments indicating a lack of awareness of previous GIS applications in the work of other active geolinguists (see Rivero, Llull, and Merlo 2002). This indicates not only a need for reviews of the literature such as included here, but also for a vigorous discussion of methodology that seems to be lacking. This is critical in order for ongoing and future projects to benefit from and build on earlier work in both geography and linguistics. As GIS and geolinguistics become more conversant, the overall role of GIS in major publications and products of the field might become more tangible, allowing for further exploration, criticism, and progress.

NOTES

1. Though not the focus in this paper, we note that GIScience has begun discussion of incorporating spatial information ontologies, informed by sociolinguistics, into GIS interfaces to account for differing conceptions of geographic concepts across languages (Mori 2002). We focus here on the placement of GIS in geolinguistics, not the reverse; although the two are interrelated, the direction of the relationship remains important.

ACKNOWLEDGEMENTS

Thanks to Barry Kronenfeld for the use of the diagrams in Figure 1. We also thank Steven Schnell and two anonymous reviewers for comments and suggestions that greatly improved the manuscript.

REFERENCES

Alvar, M. 1991. *Estudios de Geografía Lingüística (Studies of Linguistic Geography).* Madrid: Colección Filologica Paraninfo.

Anselin, L. 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis* 27: 93-115.

Breton, R. J.-L. 1991. *Geolinguistics: Language Dynamics and Ethnolinguistic Geography.* Translated by H.F. Schiffman. Ottawa: University of Ottawa Press.

Burrough, P.A. and R.A. McDonnell. 1998. *Principles of Geographic Information Systems.* New York: Oxford University Press.

Dahl, Ö. and L. Veselinova. 2005. Language Map Server. *ArcUser* [http://proceedings.esri.com/library/userconf/proc05/papers/pap2425.pdf].

Dale, M.R.T. 1999. *Spatial Pattern Analysis in Plant Ecology*. Cambridge: Cambridge University Press.

Davis, L.M. 2000. The Reliability of Dialect Boundaries. *American Speech,* 75(3): 257-9.

Gilliéron, D. and E. Edmont. 1902-10. *Atlas Linguistique de la France*. Paris: Champion.

Girard, D. and D. Larmouth. 1993. Some Applications of Mathematical and Statistical Models in Dialect Geography. In *American Dialect Research,* edited by D. Preston, Amsterdam: John Benjamins, pp. 107-132.

Goebl, H. 2006. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing,* 21(4): 411-435.

Gordon, R.G. 2005. *Ethnologue: Languages of the World*. Dallas: SIL International.

Gotway, C. and L. Young. 2002. Combining Incompatible Spatial Data. *Journal of the*

American Statistical Association, 97: 632-648.

Guy, G. 1993. The Quantitative Analysis of Linguistic Data. In *American Dialect Research*, edited by D. Preston. Amsterdam: John Benjamins, pp. 223-250.

Haase, P. 1995. Spatial Pattern Analysis in Ecology Based on Ripley's K-function: Introduction and Methods of Edge Correction. *Journal of Vegetation Science* 6(4): 575-582.

Kirk, J.M. and W.A. Kretzschmar. 1992. Interactive Linguistic Mapping of Dialect Features. *Literary and Linguistic Computing,* 7(3): 168-75.

Kretzschmar, W.A. 1992. Isoglosses and Predictive Modeling. *American Speech,* 67(3): 227-249.

_____. 1996. Quantitative Areal Analysis of Dialect Features. *Language Variation and Change,* 8: 13-39.

_____. 2003. Mapping Southern English. *American Speech,* 78(2): 130-149.

_____. 2006. Art and Science in Computational Dialectology. *Literary and Linguistic Computing,* 21: 399-410.

Kretzschmar, W.A. and E. Schneider. 1996. *Introduction to Quantitative Analysis of Linguistic Survey Data*. Thousand Oaks: SAGE Publications.

Kronenfeld, B. 2005. Gradation as a Communication Device in Area-Class Maps. *Cartography and Geographic Information Science,* 32: 231-241.

Kronenfeld, B. 2007. Triangulation of Gradient Polygons: A Spatial Data Model for Categorical Fields. In *Spatial Information Theory*, edited by S. Winter, M. Duckham, L. Kulik, and B. Kuipers. New York: Springer, pp. 421-37.

Kurath, H. 1949. *A Word Geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

Kurath, H., M. Hansen, B. Bloch, and J. Bloch. 1939-43. *Linguistic Atlas of New England*. 3 vols. Providence: Brown University Press for American Council of Learned Societies.

Labov, W., S. Ash, and C. Boberg. 2006. *The Atlas of North American English*. Berlin: Mouton de Gruyter.

Lee, J. and W.A. Kretzschmar. 1993. Spatial Analysis of Linguistic Data with GIS Functions. *International Journal of Geographical Information Science,* 7(6): 541-560.

Macauley, R.K.S. 1985. Linguistic Maps: Visual Aid or Abstract Art? In *Studies in Linguistic Geography*, edited by J.M. Kirk, S. Sanderson, and J.D.A. Widdowson. London: Croom Helm, pp. 172-86.

Mackey, W.F. 1988. Geolinguistics: Its Scope and Principles. In *Language in Geographic Context*, edited by C.H. Williams. Clevedon: Multilingual Matters Ltd., pp. 20-46.

Mark, D.M. and F. Csillag. 1989. The Nature of Boundaries on 'Area-Class' Maps. *Cartographica,* 26: 65-78.

McDavid, R. and R. O'Cain. 1980. *Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

McGuirk, D.G. 2004. An Ethnolinguistic Analysis of Hispanics in Miami-Dade County. Ph.D. diss., Florida International University.

Modern Language Association. 2009. *The Modern Language Association Language Map: A Map of Languages in the United States*. [http://www.mla.org/map_main].

Mori, M. 2002. Semantic Analysis of Spatial Expressions in Japanese. Ph.D. diss., State University of New York at Buffalo.

Nerbonne, J. and W. Heeringa. 2007. Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In *Roots: Linguistics in Search of its Evidential Base*, edited by S. Featherston and W. Sternefeld. New York: Walter de Gruyter, pp. 267-318.

Nerbonne, J. and W.A. Kretzschmar. 2003. Introducing Computational Techniques in Dialectometry. *Computers and the Humanities,* 37: 245-255.

_____. 2006. Progress in Dialectometry: Toward Explanation. *Literary and Linguistic Computing,* 21(4): 387-397.

O'Sullivan, D. and D.J. Unwin. 2002. *Geo-*

graphic Information Analysis. Hoboken: Wiley.

Ormeling, F. 1992. Methods and Possibilities for Mapping by Onomasticians. *Discussion Papers in Geolinguistics,* 19-21: 50-67.

Pederson, L. 1986. A Graphic Plotter Grid. *Journal of English Linguistics,* 19: 25-41.

_____. 1988. Electronic Matrix Maps. *Journal of English Linguistics,* 21: 149-174.

_____. 1993. An Approach to Linguistic Geography. In *American Dialect Research*, edited by D. Preston. Amsterdam: John Benjamins, pp. 31-92.

_____. 1995. Elements of Word Geography. *Journal of English Linguistics,* 25(1): 33-46.

Pederson, L., S. McDaniel, G. Bailey, and M. Basset. 1986. *Linguistic Atlas of the Gulf States*. Athens: University of Georgia Press.

Ripley, B.D. 1976. The Second Order Analysis of Stationary Point Processes. *Journal of Applied Probability* 13: 255-266.

_____. 1988. Statistical Inference for Spatial Processes. Cambridge: Cambridge University Press.

Rivero, A., G. Llull, and G.D. Merlo. 2002. Mapping the Spatial Distribution of Language. *ArcUser* [http://www.esri.com/news/arcuser/1002/linguistics.html].

Rossi, R.E., D.J. Mulla, A.G. Journel, and E.H. Franz. 1992. Geostatistical Tools for Modeling and Interpreting Ecological Spatial Dependence. *Ecological Monographs,* 62: 277-314.

Schneider, E. and W.A. Kretzschmar. 1989. LAMSAS Goes SASsy: Statistical Methods and Linguistic Atlas Data. *Journal of English Linguistics* 22: 129-136.

Thill, J.-C., W.A. Kretzschmar, I. Casas, and X. Yao. 2008. Detecting Geographic Associations in English Dialect Features in North America within a Visual Data Mining Environment Integrating Self-Organizing Maps. In *Self-Organising Maps: Applications in Geographic Information Science*, edited by P. Agarwal and A. Skupin. London: Wiley, pp. 87-105.

Thomas, A.R. 1980. *Areal Analysis of Dialect Data by Computer: A Welsh Example.* Cardiff: University of Wales Press.

Van der Merwe, I.J. 1992. A Conceptual Home for Geolinguistics: Implications for Language Mapping in South Africa. *Discussion Papers in Geolinguistics* 19-21: 33-49.

_____. 1993. The Urban Geolinguistics of Cape Town. *GeoJournal,* 31: 409-417.

Williams, C.H. 1988. An Introduction to Geolinguistics. In *Language in Geographic Context,* edited by C.H. Williams. Clevedon: Multilingual Matters Ltd., pp. 1-19.

_____. 1996. Geography and Contact Linguistics. In *Contact Linguistics: An International Handbook of Contemporary Research*, edited by H. Goebl, P.H. Nelde, Z. Stary, and W. Wolck. New York: Walter de Gruyter, pp. 63-75.

Williams, C.H. and Van der Merwe, I.J. 1996. Mapping the Multilingual City: A Research Agenda for Urban Geolinguistics. *Journal of Multilingual and Multicultural Development,* 17: 49-66.

Wright, D.J., M.F. Goodchild, and J.D. Proctor. 1997. GIS: Tool or Science? Demystifying the Persistent Ambiguity of GIS as “Tool” versus “Science”. *Annals of the Association of American Geographers,* 87(2): 346-362.