https://doi.org/10.48009/2_iis_2009_226-230

RELATIONAL DATA MODELING TO ENHANCE GIS-BASED VISUAL INFORMATION SYSTEMS

Matthew A. North, Washington & Jefferson College, mnorth@washjeff.edu
Samuel B. Fee, Washington & Jefferson College, sfee@washjeff.edu
Jacqueline M. Bytnar, Washington & Jefferson College, bytnarjm@washjeff.edu

ABSTRACT

This paper uses a geographic information system (GIS) case study on the subject of poverty detection and visualization to illustrate how relational data modeling can be used to enhance the data set upon which the GIS is based. Regular expressions, followed by one-to-one joins performed directly within the GIS, significantly increase the richness of the data, and subsequently, the visual representations of those data. Correlations are drawn between poverty levels and other spatial attributes including population density, transportation infrastructure, and other potentially influential variables. Use of the methodology applied in this case study for business or governmental decision support is then discussed in order to illustrate the utility of the paper's topic to address a wide range of organizational problems and questions.

Keywords: Geographic Information Systems (GIS), Decision Support Systems, Relational Modeling, Data Visualization.

INTRODUCTION

Geographic Information Systems are becoming an increasingly useful tool for decision support, with both corporate and non-profit applications. As more organizations adopt GIS technology, the library of data layers available for public use grows significantly [4]. These layers however are usually prepared with a specific purpose, and are therefore unlikely to contain all data attributes or variables which may be needed to address issues facing other consumers of the layers [9]. In fact, it is likely that the data set upon which any GIS is built may be enriched by additional data not present in the data table of the GIS layers being used [11].

Using basic regular expressions and one-to-one relational joins, it is possible to add available external data to existing GIS layers within ESRI's ArcGIS Desktop software application. For this case study we use U. S. Census Bureau (USCB) poverty data to illustrate how existing GIS layers can become a foundation to which additional data sets can be appended, allowing users to create a richer GIS without starting the process of data collection and georeferencing from scratch [3, 5].

Statement of the Problem

Most data gathered within any given organization are not gathered with later use in a GIS as a primary objective [6]. Often, data are simply gathered as a result of normal transactional business operations (such as records of purchases at a cash register), or collected to support specific, standing management information systems. This does not mean however that the data are not useful for ad hoc data analysis or application to other systems, such as a GIS [8, 10]. Beneficial data points and variables often do not find their way into GIS systems used by the organization, where the evaluation of such data could improve decision making using the array of analytical tools available in the GIS software. In order to take advantage of the power hidden in these disparate data sets, a method for linking external data into existing GIS systems is needed to enable broader analysis of organizational data [2].

Literature Review

Beginning with the 1990 census, and dramatically enhanced in the 2000 census, the USCB began implementing some of their data into publicly available GIS layers [1, 7]. The expansion of GIS data services offered by the USCB has increased exponentially in the current decade [6]. Although this new development has improved census-related decision making and research, the USCB cannot possibly anticipate, nor incorporate all of the variables and observations that might be relevant to the broad range of consumers that the Bureau serves.

Fortunately for GIS analysts, the basic principles of data modeling are applicable in most GIS software packages, include ESRI's ArcGIS Desktop, the dominant software player in the market today [1]. This includes the ability to create links to external data sources via delimited text files or open database connectivity (ODBC), and to create joins between attribute tables within GIS layers and database tables

in adjoining databases [2]. This ability significantly enables the expansion of spatial analysis, limited only by the researcher's time, ability, and access to data which are interesting to a given problem [12].

METHODOLOGY

As stated in the abstract and introduction, the topic selected for this paper is spatial analysis of poverty levels by census tract. The primary layer selected as the basis for this case study is the Census Tract Cartographic Boundary [3]. Although the USCB collects poverty rates for each census tract represented in this layer, the actual poverty rate attribute for each tract is not included in the layer—it is only available as a separate downloadable text file [13]. Each census tract's poverty level is listed in the text file, identified by a 12-digit code comprised of state, county, tract and block IDs. Table 1 outlines the composition of this key:

Table 1. 12 Digit Composition of Census Tract Codes

- Digits 1-2 = State code
- Digits 3-5 = County code
- Digits 6-11 = Census Tract code
- Digit 12 = Blockgroup code

Substring parsing of the census tract code using regular expressions in Microsoft Excel allows for the creation of a five attribute data set, with census tract code (digits 6-11) as the primary key in the data set. Use of LEFT, MID and RIGHT functions made extraction of the various portions of the 12 digit codes relatively simple to accomplish. Table 2 depicts poverty rates by census tract before and after parsing.

Table 2. Census Tract Codes before and after parsing

Pre-parsed data: 01001020100 , 0.1268156425

Post-parsing: 01 001 02010 0, 12.68%

As described in table 1, the first two digits of the census tract code represent the state. For this case study, state code 42, Pennsylvania, was chosen.

With the data parsed, we are prepared to use a relational join to connect the external data source to the data already contained in the GIS layer. Since each census tract ID exists only once in both the attribute table of the Cartographic Boundary shape file and in the Excel spreadsheet, the data sources can be merged using a one-to-one relational join, resulting in a unified attribute table which can then be visually represented and analyzed in the GIS.

Using the Jenk's Natural Breaks algorithm available within the GIS, shading ramps are employed to indicate poverty levels by darkening the area of each census tract—as poverty increases, the tract becomes darker [1]. Additional GIS layers are added to the map to evaluate correlation with other spatial features. Proximity to police, fire or public works facilities; transportation infrastructure; residential, commercial and industrial zones, etc. are all feasible additions to the GIS in order to evaluate and plan responses to impoverished areas.

RESULTS

In the 2000 census, the state of Pennsylvania was represented by 3,135 tracts. The size of these tracts varied tremendously based on population density. In the relatively sparsely populated regions of central Pennsylvania, tracts are larger, while the more urban areas around Pittsburgh, Philadelphia, Harrisburg and Wilkes-Barre/Scranton are home to smaller physical census tracts.

In the 2000 census, a family of four would have had to earn $17,603 or less to be considered to live below the poverty level [4]. Eighteen tracts in the state enjoyed zero percent poverty levels in the 2000 census, while 47 tracts in the state suffered from poverty levels above 50%. Fifteen tracts did not report their poverty levels. The average statewide poverty level was approximately 12%.

Figure 1. Poverty Levels by Census Tract, Pennsylvania, 2000 U.S. Census.

In order to better visualize the reality of these figures, the Jenks Natural Breaks algorithm was applied to the unified attribute table, with color representations selected to define poverty levels. Figure 1 depicts all

3,135 census tracts according to their relative poverty levels ranging from low (lightest shade), to extreme (darkest shade).

The visualization of these statistics is enabled by the relationship created after the data were parsed. While the census tract number defines each tract individually, the additional attribute defining each tract's poverty level allows for graphical representation by shade [11]. It is interesting to note when evaluating figure 1, that no one specific region of the state emerged as a predominately dark (impoverished) area. In contrast, several areas, including suburban Pittsburgh, suburban Philadelphia, and the area surrounding Centre County (the location of Penn State, the state's flagship university) are all predominately light in color. While it may seem obvious that such areas would not likely suffer from poverty, the data represented in the map substantiate this expectation, rather than assume it.

Interestingly, when aggregated to the county level within the GIS, a previously unseen pattern of poverty emerges within the state. Figure 2 depicts this phenomenon.

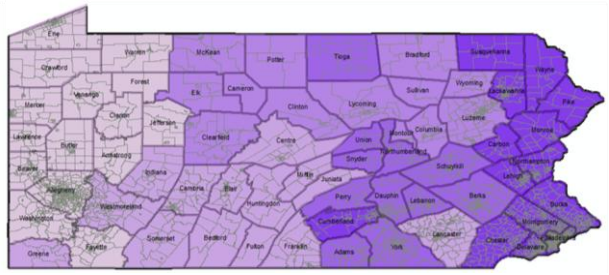

Figure 2. Poverty Levels by County, Pennsylvania, 2000 U.S. Census.

This view of the same combined data paints a different picture of poverty in the state when displayed by county. This view can be further extended by adding further data layer to the GIS in order to seek correlations between poverty and other variables. For example, poverty is sometimes stereotypically linked with ethnicity. Figure 3 depicts the same data as displayed in Figure 2, with an overlay of pie charts showing each county's ethnic breakdown. According to the United States Census Bureau the diversity in ethnicity throughout the state of Pennsylvania is not extremely variable. The state of Pennsylvania is predominantly white. This fact makes it hard to conclude whether or not ethnicity is a factor in determining causes of poverty.

In this visual representation, as with the poverty rate data itself, an external data set needed to be joined relationally with the attribute table of the GIS layer. In this instance however, ethnic breakdowns were joined based on the county code, rather than the census tract number.

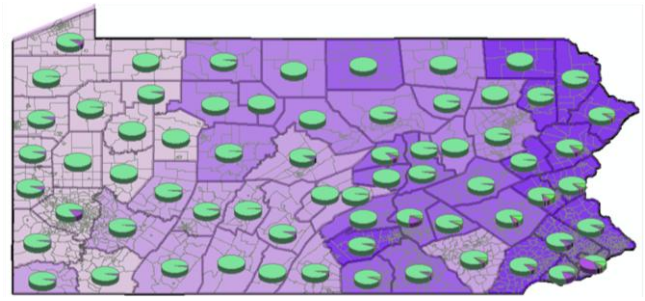

Figure 3. Poverty by Ethnicity by County, Pennsylvania, 2000 U.S. Census.

One final example of the power of relational data joins within the analytically rich environment of the GIS is depicted in Figure 4. In this example, coincidence of poverty intersecting urbanization is sought. A GIS layer displaying Pennsylvania cities is placed over the same poverty data displayed in Figures 2 and 3. Although Pennsylvania is home to two relatively large cities by American standards, Philadelphia and Pittsburgh, it is largely a rural state. Assumptions could be made in either direction: Pennsylvania's poor live in it two most urban regions; or, Pennsylvania's poor live dispersed across its hilly and rural landscape.

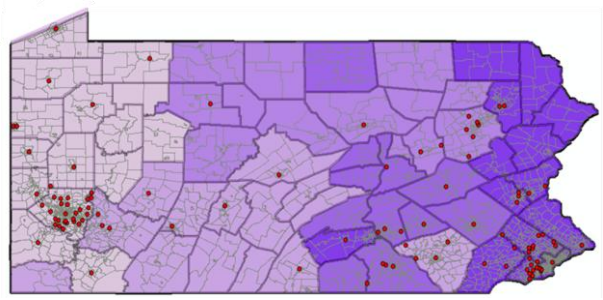

Figure 4. Poverty Levels by County Near Cities, Pennsylvania, 2000 U.S. Census.

Although persuasive arguments for both aforementioned assumptions could easily be made, the GIS tells a different story. In the extreme southeast of the state, as represented in Figure 4, a cluster of red dots represent the cities in the Philadelphia metropolitan area. The counties behind these cities are tinted a dark shade, representing higher rates of poverty. By contrast, the cluster on the western side of the state shows the cities around

the Pittburgh region. The primarily lightly colored counties behind this cluster of cities indicate much lower levels of poverty. Finally, in the northeast of the state, few cities are found, but the counties here are also tinted dark for their higher levels of poverty. Thus, any argument laying the blame for poverty at either the feet of urbanization or rural communities is at best, short sighted.

Future Study

The methodology shown in this paper is applicable to any data where a baseline GIS layer's attribute table can be joined through a common key to a database table, spreadsheet, or other data set containing that same attribute. Expansion of the concepts illustrated in this case study could include creating one-to-many joins on database tables or dataset where multiple observations relate to a single observation in the baseline GIS layer's attribute table.

A more likely scenario for extension of this work however lies in the opportunity for researchers to apply the methodology of relationally joining any data set to existing data in a GIS by way of a common key, exploiting the database's (or data set's) superior capacity for data storage with the GIS's powerful analytical capabilities.

CONCLUSIONS

It is clear from the examples presented in this case study that external data, joined relationally into existing GIS data, enrich the analytical environment of the GIS, and subsequently, the quality of any results gleaned from it. Often times, organizations have access to data which could be of benefit if incorporated into a GIS.

The use of regular expressions, even simple ones such as LEFT, MID, and RIGHT, as used in this case study, can create necessary attributes which can serve as relational keys in order to incorporate external data into an existing GIS. This data can then in turn be used in analyzing and evaluating spatial data in ways which were not previously possible with only the existing attribute table data in the GIS layers themselves.

Organizations using GIS for analytics, decision making or operations management would be wise to consider the concepts and methodologies proposed in this paper to enrich, as much as possible, the data available for analysis in their GIS systems. The basic methods proposed here could result in the discovery of unexpected patterns in data or the support (or dismissal) or preconceived assumptions. The examples of poverty levels in the state of Pennsylvania, with their correlations (or lack thereof) with other data variables relevant to the state (such as ethnicity or population density), illustrate how parsing and joining external data into a GIS can enhance that system's utility to end users.

REFERENCES

1. Barnes, S. (2003). Paired with ESRI. *Geospatial Solutions, 13*(4), 12-13.
2. Baza, I., Geymenb, A. & Er, S. M. (2009). Development and Application of GIS-based Analysis/Synthesis Modeling Techniques. *Advances in Engineering Software, 40*(2), 128-140.
3. Census 2000: Census Tract Cartographic Boundary Files - U.S. Census Bureau. (2000). *Retrieved on March 18, 2009 from:* http://www.census.gov/geo/www/cob/tr2000.html.
4. Census 2000: Census Tract Cartographic Definitions - U.S. Census Bureau. (2000). *Retrieved on March 18, 2009 from:* http://www.census.gov/hhes/www/poverty/definitions.html
5. Census 2000: Poverty Thresholds 2000 - U.S. Census Bureau. (2000). *Retrieved on March 18, 2009 from:* http://www.census.gov/hhes/www/poverty/threshld/thresh00.html
6. Engelhardt, J. & Barnes, S. (2004). ESRI Conference and the Language of Geography. *Geospatial Solutions, 14*(9), 12-14.
7. Fonseca, B. (2003). IBM, ESRI Target Geospatial Data. *eWeek, 20*(48), 28.
8. Keenan, P. B. (1998). Spatial Decision Support Systems for Vehicle Routing. *Decision Support Systems, 22*(1), 65-71.
9. Matty, J. M. (2003). GIS Data Made Available. *Rocks & Minerals, 78*(3), 153.
10. Ravallion, M. (2001). Growth, Inequality and Poverty: Looking Beyond Averages. World Development, 29(11), 1803-1815.
11. Shuo-sheng, W., Le, W. & Xiaomin, Q. (2008). Incorporating GIS Building Data and Census Housing Statistics for Sub-Block-Level Population Estimation. *Professional Geographer, 60*(1), 121-135.
12. Stevens, D., Dragicevica, S. & Rothleyb, K. (2007). iCity: A GIS–CA modelling tool for urban planning and decision making.

Environmental Modelling & Software, 22(6), 761-773.

13. U. S. Census Tract Poverty Data. (2000). *Retrieved on March 8, 2009 from:* http://www.hsph.harvard.edu/thegeocodingproject/webpage/monograph/povdata.htm.