The Use of Spatially Lagged Explanatory Variables for Modeling Neighborhood Amenities and Mobility in Older Adults

Tony H. Grubesic
Drexel University

Andrea L. Rosso
University of Pittsburgh

Abstract

As more researchers in the socioeconomic, planning, and health sciences embrace the use of spatial data for exploring the local context of study regions, the demand for alternative (not U.S. Census Bureau) databases is increasing. In particular, information pertaining to local amenities (for example, retail, recreation, and cultural resources) or disamenities (for example, crime and pollution) can provide important details about place. The purpose of this article is to provide a brief overview of a popular alternative data source for capturing local amenities in an urban environment: the Esri Business Analyst. This article also explains and illustrates an approach for incorporating these data into a spatial analysis. We specifically highlight the use of spatially lagged explanatory variables in general linear models. In the spirit of previous contributions to the SpAM series in Cityscape*, this article uses data and mirrors methods from a previously published study. In this case, we expand on the work of Rosso et al. (2013) and their recently completed analysis of neighborhood amenities and mobility for older adults in Philadelphia, Pennsylvania.*

Introduction

A growing body of literature in epidemiological and socioeconomic planning sciences focuses on the assessment of neighborhood influences on health outcomes (Diez Roux, 2007). The literature includes recent works pertaining to obesity (Berke et al., 2007; Booth et al., 2005), assaultive violence (Grubesic et al., 2013; Pridemore and Grubesic, 2012), and risky sexual behavior (Towe et al., 2010), among others. In part, the growth of neighborhood-related research is attributable to the explosion of spatial data and associated analytical methods available to researchers (Moore and Carpenter, 1999). High-resolution spatial data at the block group, block, or even household level greatly enhance public health studies (Gabrysch et al., 2011), improving context and often decreasing spatial uncertainty (Murray et al., 2014) when compared with more aggregate units such as census tracts or ZIP Codes (Grubesic and Matisziw, 2006). Alternative data sources are also gaining favor with many researchers attempting to break out of the traditional box of Census Bureau-based demographic and socioeconomic information. For example, the use of social media data to monitor influenza outbreaks (Corley et al., 2010) or the use of business establishment data to explore trends in broadband provision (Mack, 2014) are two cases in which alternative, somewhat unconventional data are being used to answer important, substantive, policy-related questions.

One area of community health research—the assessment of neighborhood influences on the mobility of older adults—has benefited from the interface of spatial analytical methods, geographic information systems, and alternative data sources (Rosso et al., 2013; Rosso et al., 2011). Although many of the more traditional studies in this domain (for example, Chaudhury et al., 2012; Patterson and Chapman, 2004) rely on census tracts to define neighborhoods, these measures and associated data remain somewhat coarse and fail to account for how the characteristics of proximal neighborhoods and their spatial effects (that is, interaction) may affect outcomes. The inclusion of spatial effects to help account for these complexities is now common in many disciplines, including economics, geography, ecology, and criminology (Florax and Nijkamp, 2003). The spatial effects have not been widely adopted, however, in public health research.

Accounting for spatial effects is often motivated by a combination of theoretical considerations (for example, understanding that neighborhoods are not islands and do not exist in isolation) and/or the peculiarities of the data used for empirical analysis (Anselin, 2002). The process of incorporating spatial effects, however, remains technically challenging for several reasons. First, different spatial models can create distinctly different spatial correlation patterns (Anselin, 2002). Therefore, a relatively deep understanding of how spatial weight matrices need to be constructed is needed for capturing the theorized spatial interaction (Anselin and Rey, 1991; Florax and Rey, 1995). Second, the use of a spatially lagged dependent variable (Wy) in regression models is often difficult to implement in public health research because individual study participants are frequently the unit of analysis in epidemiological studies. As a result, it is problematic to capture and model spatial contiguity in the dependent variable unless participants are specifically recruited to provide this contiguity. Third, models that capture spatial dependence often require specialized estimation methods (Anselin, 1988), most of which are not readily available in standard commercial statistical packages such as SPSS, NCSS, or SAS.

Given these challenges (and potential opportunities), the purpose of this article is twofold. First, we detail the utility of the Esri Business Analyst (hereafter, Business Analyst) data (Esri, 2010),

a popular alternative database for the ecological analysis of neighborhoods and their amenities. Second, we explain and illustrate the use of spatially lagged explanatory variables in general linear models, emphasizing their utility for exploring a range of neighborhood-related issues in the epidemiological and socioeconomic planning sciences. In the spirit of previous contributions to the SpAM series in *Cityscape*, this article uses data and mirrors methods from a previously published study. In this case, we expand on the work of Rosso et al. (2013) and their recently completed analysis of neighborhood amenities and mobility for older adults in Philadelphia, Pennsylvania.

Capturing Amenity Diversity in Neighborhoods

Many public health studies rely deeply on U.S. Census Bureau-based data for capturing the local, ecological conditions of neighborhoods (Krieger et al., 1997). Although this reliance on Census Bureau data is shared across many of the socioeconomic and planning sciences, it is important to note that these data are extremely limited in scope when considering the multifaceted composition of neighborhoods. As a result, analysts must use alternative data sources to capture information on neighborhood amenities such as retail establishments, local services, medical providers, and civic and community facilities.

Dun and Bradstreet (D&B) and infoUSA Inc. are two of the most widely available alternative databases for capturing the local ecological composition of neighborhood amenities (Powell et al., 2011), providing millions of data points for local businesses and services in the United States. In particular, the Business Analyst is a popular portal to the infoUSA Inc. data that are supplemented by information from other sources, such as federal and state business registries, local telephone directories, and information from the U.S. Postal Service, to cross-reference and enhance its local amenity data (Esri, 2011). Thus, the use of this supplemental information to create the Business Analyst database may offer an improvement to the raw infoUSA Inc. data. Previous validation work suggests that the Business Analyst includes approximately 51 percent of all business types (Hoehner and Schootman, 2010), but recent empirical studies on the concordance of the D&B and infoUSA Inc. data, focusing on retail food establishments (for example, food stores and restaurants), suggests that their validity is moderate, at best (Powell et al., 2011). Moreover, Powell et al. (2011) argue that these data should not be used as a substitute for "on-the-ground data collection" (Powell et al., 2011: 1130) unless additional efforts for verification, such as a telephone screening procedure, are made.

It is clear that no database is perfect. Secondary data on local establishments and amenities cannot be expected to reflect the dynamic business and economic environment with 100 percent accuracy, regardless of the supplementary data used for database development. Such environments have far too many changes to capture on a daily, weekly, and monthly basis. As a result, partial coverage and a lack of complete concordance are known limitations to these data. It can also be argued that these data remain valuable, however, even if they provide only a relatively conservative estimate of neighborhood amenities.

In a recent study of amenity diversity and its connections to mobility in older adults in the city of Philadelphia, Rosso et al. (2013) used the Business Analyst database to obtain a local ecological snapshot of multiple neighborhoods. Rosso et al. (2013) specifically leveraged the "diverse uses" criterion from the Leadership in Energy and Environmental Design Neighborhood Development

(LEED-ND) to define amenity diversity (USGBC, 2009), where the occurrence of any particular amenity type (up to two occurrences) was counted for each neighborhood.[1] These counts were then summed for each neighborhood across the 27 unique types of amenities used for analysis, which ranged from pharmacies to hardware stores and other retail and from medical clinics to post offices and public libraries.[2] The resulting scale of amenity diversity ranged from 0 to 54 and had a Cronbach's alpha value of 0.79, suggesting adequate consistency across multiple neighborhoods in Philadelphia.

This type of approach works for several reasons for neighborhood-level ecological analyses. First, the measure is structured to capture amenity diversity—nothing more, nothing less. It is not structured to provide a complete audit of establishments or amenities within a neighborhood. Second, because the measure focuses on amenity diversity, the use of a conservative database of establishments is actually beneficial for the resulting index. In effect, the lack of inflation in the Business Analyst suggests that when amenities are found, the likelihood of more existing in the neighborhood is high, even if they are unaccounted for in the database. Of course, the inverse is also true—where less common amenities remain obscured by the existing database—but these can be mitigated with alternative data sources as well. For example, Rosso et al. (2013) captured famers' market locations from an approved list of operations maintained by the city government of Philadelphia. Official city parks were captured in a similar way. In the end, analysts must structure their measures to reflect known uncertainties or limitations in the data—a process no different than using data from Census Bureau-based sources like the American Community Survey (Citro and Kalton, 2007).

Spatially Lagged Explanatory Variables

A second facet of the Rosso et al. (2013) work that provides some flexibility in capturing neighborhood interaction and enhancing the statistical legibility of the connections between the mobility of older adults and local amenities is the use of spatially lagged explanatory variables for use in generalized estimating equations (GEEs). GEEs are semiparametric regression techniques used to estimate parameters of a generalized linear model when the correlation between outcomes is unknown (Hardin, 2005). GEEs are popular for public health studies in which cohorts are distributed across multiple study areas (for example, neighborhoods) because GEEs are good at handling unmeasured dependence between outcomes (Lin et al., 1998). Spatially lagged explanatory variables (Wx) are used to capture the weighted sum of values for neighborhood i by using its local neighbors as weights. Specifically,

$$[Wx]_i = \sum_{j \neq i} w_{ij} x_j , \tag{1}$$

where the influence or weight of each link $i - j$ is expressed in the weight matrix. As detailed by Anselin (2002; 1988), these weights are often based on the geographic contiguity for each j, relative to the location of i, but the weights can easily be expressed via alternative conceptualizations

[1] In this study, census tracts were used as surrogates for neighborhoods in Philadelphia.

[2] For a complete list of amenities, see the original paper (Rosso et al., 2013).

such as *k* nearest neighbor or with distance-based matrices. Florax and Rey (1995) and Anselin and Rey (1991) provide some guidance on the proper specification of these weights matrices and also on errors attributable to a poor specification.

Spatially lagged explanatory variables are important tools to use for regression modeling, broadly defined. In fact, their potential use as cross-regressive terms stands in sharp contrast to the more widely used form of spatial regression modeling, where the dependent variable is lagged. Although space limitations prevent us from detailing the nuances of spatial reaction functions and their theoretical basis for dealing with spatial autocorrelation in linear regression models, readers are referred to Brueckner (2002) for more detail. In short, rather than creating a multiplier effect as with spatially lagged dependent variables, spatial cross-regressive terms can be used directly in a standard regression framework. With spatially lagged explanatory variables, variables can be spatially lagged, or not, depending on model context:

$$y = X\beta + WX\gamma + \varepsilon\,. \tag{2}$$

The range of the spatial cross-regressive terms spans from the very local, where only a few neighbors are included, to the global, where all neighbors (for all *i*) are included. This range is directly contingent upon the number of zero-restrictions ($w_{ij} = 0$) imposed for a study region or neighborhood (Anselin, 2002). Further, it is important to note that, unlike the more common spatially lagged regression models in which simultaneity makes the *Wy* variables endogenous, the spatial cross-regressive framework does not require any specialized estimation techniques. In other words, even ordinary least squares regressions would work with these data and not bias γ (Anselin, 2002).

Rosso et al. (2013) used a spatially lagged explanatory variable in a slightly different way for their analysis of neighborhood amenity diversity and adult mobility in Philadelphia. The lagged variable of amenity scores, which was defined with a queen's contiguity matrix, was specifically used as an interaction term for the GEEs. Interaction terms are often used in epidemiologic analyses to determine whether the association between an explanatory factor and the dependent variable is moderated by a third variable. Consider the following:

$$y = X_1\,\beta_1 + X_2\,\beta_2 + X_1\,X_2\,\beta_3 + \varepsilon, \tag{3}$$

where β_3 specifies the magnitude of the interaction. Rosso et al. (2013) used this interaction term to capture how amenity diversity for each tertile of the index census tract was moderated by amenity diversity of the surrounding census tracts.[3] This approach allowed for an objective assessment of whether the observed associations were specific to the characteristics of a participant's home census tract or were influenced by the characteristics of surrounding census tracts. Again, an advantage of this method is that it can be implemented in standard statistical packages.

Several GEE models with and without the spatially lagged explanatory variables and their associated interaction terms ultimately were compared by minimization of the penalized quasi-information criteria (QICu), which accounts for the number of parameters in the model (De Knegt et al., 2010).

[3] The LEED-ND guidelines provide cutoffs to define levels of diverse use. The Rosso et al. (2013) study divided these cutoffs into tertiles for analysis.

In analyses adjusted for individual- and neighborhood-level covariates, no association was observed between tertile of amenity diversity and mobility (that is, mean difference in mobility;[4] for example, when compared with the lowest tertile: mean difference in mobility at the middle tertile = -2.2, 95% CI: -6.5, 2.2 and the highest tertile = 3.2, 95% CI: -1.5, 7.9). When analyses were restricted to those individuals who reported the most time spent in their home neighborhood (see Rosso et al., 2013 for details), a significant association was observed for those living in census tracts in the highest tertile of amenity diversity compared with those in the lowest tertile (mean difference = 8.3; 95% CI: 0.1, 16.6) with approximately equal mobility for those in the middle compared with the lowest tertile (mean difference = -1.7 for middle; 95% CI: -10.0, 6.6).

Finally, no significant interactions were reported between tertile of amenity diversity at the index census tract and the spatially lagged explanatory variable (all $p > 0.2$). Inclusion of interaction terms also did not improve model fit.[5] Similar results were observed for analyses restricted to those individuals who spent the most time in their neighborhoods. These results indicate that the associations between mobility and the observed amenity diversity of a participant's home census tract were not greatly influenced by the amenity characteristics of neighboring census tracts. In part, the lack of influence from neighboring tracts may be explained by the moderate amount of concordance between amenity diversity at the index tract and the spatially lagged estimate of amenity diversity for neighboring tracts (44 percent of tracts were in the same tertile of amenity diversity as their spatially lagged counterpart; kappa = 0.2). Alternatively, this lack of influence may suggest that at least for associations between amenity diversity and mobility of older adults, measures at the participant's own census tract are sufficient to capture relevant neighborhood characteristics.

Discussion and Conclusion

Several points are worth further discussion. First, alternative data sources, such as the Business Analyst database, are useful sources of information to augment Census Bureau-based data. Although the data are not perfect, analysts who understand the limitations and take steps to mitigate known uncertainties will find these types of alternative datasets can provide more detail and depth for exploring neighborhoods and their ecological context. Second, the use of spatially lagged explanatory variables enables analysts to consider spatial effects between neighborhoods in a meaningful way. More important, this method can be accomplished without the use of specialized estimation techniques, making spatially lagged explanatory variables more readily integrated into many public health studies than other spatial regression techniques. The choice of spatial weights for developing these explanatory variables remains important, and analysts should take time to conduct some basic sensitivity analysis for evaluating which weights matrix best captures the theorized interaction.

Finally, where the Rosso et al. (2013) application is concerned, one limitation of using spatially lagged explanatory variables as interaction terms is that the model requires a sufficient sample size to detect interactions (Greenland, 1983). To be specific, because interaction relies on dividing

[4] Mobility was assessed by the Life-Space Assessment (Peel et al., 2005), which uses a scale of 0 to 104 points. Higher scores indicate higher mobility.

[5] QICu without interaction terms = 488, with interaction terms = 494; smaller QICu indicates better fit.

the study population into smaller subgroups, statistical power is lost. For many large-scale public health or socioeconomic and planning studies, this loss of statistical power may not be an issue. Statistical power must be considered on a case-by-case basis, however, by reviewing the distribution of study subjects within the various levels of the modifying variable (for example, tertiles). Note that interpretation of interaction terms can be difficult if the modifying variable is continuous.

In sum, the growing availability of alternative data sources, combined with the power of geographic information systems and associated analytical methods, provides a powerful foundation for advanced geographic reasoning at a highly localized level. Although the connection between neighborhood amenities and adult mobility is just one application, many more substantive domains exist where this fusion of data and methods, including the development of spatially lagged explanatory variables, would be useful. It is important to reiterate that care must be taken to understand the limitations of both the data and techniques being used for analysis, as uncertainties will remain. When applied rigorously, however, many opportunities arise to improve the efficacy of public policy and public health interventions with these methods.

Acknowledgments

The authors thank Elizabeth A. Mack for her helpful comments on the initial draft of this manuscript.

Authors

Tony H. Grubesic is the Director of the Center for Spatial Analytics and Geocomputation and a professor in the College of Computing & Informatics at Drexel University.

Andrea L. Rosso is a postdoctoral fellow at the Center for Aging and Population Health, Department of Epidemiology, Graduate School of Public Health at the University of Pittsburgh.

References

Anselin, Luc. 2002. "Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models," *Agricultural Economics* 27 (3): 247–267.

———. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht, the Netherlands: Springer.

Anselin, Luc, and Serge Rey. 1991. "Properties of Tests for Spatial Dependence in Linear Regression Models," *Geographical Analysis* 23 (2): 112–131.

Berke, Ethan M., Thomas D. Koepsell, Anne Vernez Moudon, Richard E. Hoskins, and Eric B. Larson. 2007. "Association of the Built Environment With Physical Activity and Obesity in Older Persons," *American Journal of Public Health* 97 (3): 486–492.

Booth, Katie M., Megan M. Pinkston, and Walker S. Carlos Poston. 2005. "Obesity and the Built Environment," *Journal of the American Dietetic Association* 105 (5): 110–117.

Chaudhury, Habib, Atiya Mahmood, Yvonne L. Michael, Michael Campo, and Kara Hay. 2012. "The Influence of Neighborhood Residential Density, Physical and Social Environments on Older Adults' Physical Activity: An Exploratory Study in Two Metropolitan Areas," *Journal of Aging Studies* 26 (1): 35–43.

Citro, Constance F., and Graham Kalton. 2007. *Using the American Community Survey: Benefits and Challenges*. Washington, DC: National Academies Press.

Corley, Courtney D., Diane J. Cook, Armin R. Mikler, and Karan P. Singh. 2010. "Text and Structural Data Mining of Influenza Mentions in Web and Social Media," *International Journal of Environmental Research and Public Health* 7 (2): 596–615.

De Knegt, Henrik J., Frank van Langevelde, Michael B. Coughenour, Andrew K. Skidmore, Willem F. de Boer, Ignas M.A. Heitkonig, Nichola M. Knox, Robert Slotow, Cornelis van der Waal, and Herbert H.T. Prins. 2010. "Spatial Autocorrelation and the Scaling of Species-Environment Relationships," *Ecology* 91 (8): 2455–2465.

Diez Roux, Anna V. 2007. "Neighborhoods and Health: Where Are We and Where Do We Go From Here?" *Revue d'Epidémiologie et de Santé Publique* 55 (1): 13–21.

Esri. 2011. *Methodology Statement: Esri Data—Business Locations and Business Summary*. Redlands, CA: Esri.

———. 2010. *Methodology Statement: Esri Data—Business Locations and Business Summary*. Redlands, CA: Esri.

Florax, Raymond J.G.M., and Peter Nijkamp. 2004. "Misspecification in Linear Spatial Regression Models." In *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard. San Diego: Academic Press.

Florax, Raymond J.G.M., and Sergio J. Rey. 1995. "The Impacts of Misspecified Spatial Interaction in Linear Regression Models." In *New Directions in Spatial Econometrics*, edited by Luc Anselin and Raymond J.G.M. Florax. Berlin, Germany: Springer: 111–135.

Gabrysch, Sabine, Simon Cousens, Jonathan Cox, and Oona M.R. Campbell. 2011. "The Influence of Distance and Level of Care on Delivery Place in Rural Zambia: A Study of Linked National Data in a Geographic Information System," *PLoS Medicine* 8 (1). DOI: 10.1371/journal.pmed.1000394.

Greenland, Sander. 1983. "Tests for Interaction in Epidemiologic Studies: A Review and a Study of Power," *Statistics in Medicine* 2 (2): 243–251.

Grubesic, Tony H., and Timothy C. Matisziw. 2006. "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data," *International Journal of Health Geographics* 5 (1): 58.

Grubesic, Tony H., William Alex Pridemore, Dominique A. Williams, and Loni Philip-Tabb. 2013. "Alcohol Outlet Density and Violence: The Role of Risky Retailers and Alcohol-Related Expenditures," *Alcohol and Alcoholism* 48 (5): 613–619.

Hardin, James W. 2005. *Generalized Estimating Equations (GEE)*. New York: John Wiley & Sons.

Hoehner, Christine M., and Mario Schootman. 2010. "Concordance of Commercial Data Sources for Neighborhood-Effects Studies," *Journal of Urban Health* 87 (4): 713–725.

Krieger, Nancy, David R. Williams, and Nancy E. Moss. 1997. "Measuring Social Class in U.S. Public Health Research: Concepts, Methodologies, and Guidelines," *Annual Review of Public Health* 18 (1): 341–378.

Lin, Danyu Y., Bruce M. Psaty, and Richard A. Kronmal. 1998. "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics* 54 (3): 948–963.

Mack, Elizabeth A. 2014. "Businesses and the Need for Speed: The Impact of Broadband Speed on Business Presence," *Telematics and Informatics* 31 (4): 617–627.

Moore, Dale A., and Tim E. Carpenter. 1999. "Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology," *Epidemiologic Reviews* 21 (2): 143–161.

Patterson, Patricia K., and Nancy J. Chapman. 2004. "Urban Form and Older Residents' Service Use, Walking, Driving, Quality of Life, and Neighborhood Satisfaction," *American Journal of Health Promotion* 19 (1): 45–52.

Peel, Claire, Patricia Sawyer Baker, David L. Roth, Cynthia J. Brown, Eric V. Bodner, and Richard M. Allman. 2005. "Assessing Mobility in Older Adults: The UAB Study of Aging Life-Space Assessment," *Physical Therapy* 85 (10): 1008–1019.

Powell, Lisa M., Euna Han, Shannon N. Zenk, Tamkeen Khan, Christopher M. Quinn, Kevin P. Gibbs, Oksana Pugach, Dianne C. Barkere, Elissa A. Resnicka, Jaana Myllyluomaf, and Frank J. Chaloupka. 2011. "Field Validation of Secondary Commercial Data Sources on the Retail Food Outlet Environment in the U.S.," *Health & Place* 17 (5): 1122–1131.

Pridemore, William Alex, and Tony H. Grubesic. 2012. "A Spatial Analysis of the Moderating Effects of Land Use on the Association Between Alcohol Outlet Density and Violence in Urban Areas," *Drug and Alcohol Review* 31 (4): 385–393.

Rosso, Andrea L., Amy H. Auchincloss, and Yvonne L. Michael. 2011. "The Urban Built Environment and Mobility in Older Adults: A Comprehensive Review," *Journal of Aging Research*. DOI: http://dx.doi.org/10.4061/2011/816106.

Rosso, Andrea L., Tony H. Grubesic, Amy H. Auchincloss, Loni Philip-Tabb, and Yvonne L. Michael. 2013. "Neighborhood Amenities and Mobility in Older Adults," *American Journal of Epidemiology* 178 (5): 617–637.

Towe, Vivian L., Frangiscos Sifakis, Renee M. Gindi, Susan G. Sherman, Colin Flynn, Heather Hauck, and David D. Celentano. 2010. "Prevalence of HIV Infection and Sexual Risk Behaviors Among Individuals Having Heterosexual Sex in Low-Income Neighborhoods in Baltimore, MD: The BESURE Study," *Journal of Acquired Immune Deficiency Syndromes* 53 (4): 522–528.

United States Green Building Council (USGBC). 2009. *Green Neighborhood Development: LEED Reference Guide for Neighborhood Development*. Washington, DC: U.S. Green Building Council.