Data Shop

Data Shop, a department of Cityscape, *presents short articles or notes on the uses of data in housing and urban research. Through this department, PD&R introduces readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis is on sources and methods that analysts can use in their own work. Researchers often run into knotty data problems involving data interpretation or manipulation that must be solved before a project can proceed, but they seldom get to focus in detail on the solutions to such problems. If you have an idea for an applied, data-centric note of no more than 3,000 words, please send a one-paragraph abstract to david.a.vandenbroucke@hud.gov for consideration.*

A Note on Data Preparation Procedures for a Nationwide Analysis of Urban Form and Settlement Patterns

Robert N. Renner
U.S. Department of Housing and Urban Development

Selma Lewis
University of Maryland

John I. Carruthers
U.S. Department of Housing and Urban Development; University of Maryland

Gerrit-Jan Knaap
University of Maryland

This note is a companion to U.S. Department of Housing and Urban Development Working Papers REP 08-03 and REP 09-03 and, as such, it reiterates some of the discussion contained in those two papers. It expressly describes work that is in progress, which may, and indeed likely will, evolve as the projects move forward. The opinions expressed in this note are those of the authors and do not necessarily reflect the opinions of the U.S. Department of Housing and Urban Development or the U.S. government at large.

Abstract

This note outlines in detail the process of preparing data used in studying patterns of urbanization across the United States, using spatial hazard models—a class of durational models often employed in analyzing lifecycles. The note provides a brief overview of spatial hazard models and their application in the analysis of urbanization patterns and continues to describe the collection and processing of settlement point patterns needed for the analysis. Analyzed at the census block group level, data come from (1) a nationwide count of housing units at the census block level in 2006, which the Census Bureau provided to the U.S. Department of Housing and Urban Development; and (2) Census Summary File 3 from the 2000 Census of the population.

Introduction

The purpose of this note is to detail the process of preparing data needed for studying patterns of urbanization across the United States via spatial hazard models (for applications, see Carruthers et al., 2008, and Lewis et al., 2008). Spatial hazard models are geographic adaptations of proportional hazard models—a class of duration, or failure time—models normally used for analyzing lifecycles (Cleves, Gould, and Guitierrez, 2004; Kiefer, 1988; Lawless, 2002). Duration or failure analysis involves the modeling of time-to-event data. As a simple example, time-to-event modeling could be the rate or time a lightbulb lasts. The idea of adapting hazard models to the spatial realm originates from Odland and Ellis (1992), who were the first to use the method for point pattern analysis, a method of spatial analysis that has long been used to study both natural and social phenomena (Boots and Getis, 1988; Diggle, 2003). More recently, Waldorf (2003) elaborated on the mathematical logic of using hazard models for this purpose and laid out a detailed conceptual framework for bringing the models to bear on spatial point patterns. Since 1988, the approach has been applied to an array of spatial phenomena, ranging from the reach of market areas (Esparza and Krmenec, 1996) to the adoption of agricultural technology (Pellegrini and Reader, 1996) to the spread of disease (Reader, 2000). Although still uncommon, spatial hazard models appear to represent a promising—and, in the view of the present authors, an exciting—new way of analyzing the kind of spatial processes and outcomes that are commonly of interest in the social sciences and public policy fields. The next section, which is a condensed version of the background discussion in Carruthers et al. (2008), outlines the nature of the modeling framework and the data needed for it.

Spatial Hazard Models

Distance, like time, is a nonnegative random variable that terminates at a given point conditional on the probability of having made it to that point in the first place. That is, a particular distance between geographically referenced points either extends or terminates based on the nature of the spatial process that gave rise to the distribution of the referenced points in the first place. At the core of this conceptualization is a so-called "spatial hazard function" that mathematically describes the conditional probability of distance terminating (for details, see Waldorf, 2003). For example, intuitively, one might expect the hazard function for the spacing of settlements to exhibit positive dependence, or a hazard of terminating that increases with distance, and the hazard function for the spread of an illness to exhibit negative dependence, or a hazard of terminating that decreases with distance. A proportional hazard model of a particular point-generating process is achieved by choosing an appropriate statistical distribution for the baseline hazard, plus a set of independent explanatory variables that accelerate and/or decelerate, as the case may be, the rate at which distances between spatial points, say d_{ij}, terminate. The Weibull distribution is the most widely used distribution in survival analysis, and it works particularly well because it gives the hazard a flexible shape (Lawless, 2002):

$$h(d_{ij} \mid X) = h_0(d_{ij}) \cdot \exp(X \cdot \Phi). \qquad (1)$$

In this spatial hazard model, the hazard function consists of two components: (1) a baseline hazard, $h_0(d_{ij}) = \lambda d_{ij}^{\lambda-1}$, described by λ, a shape parameter, which gives the rate at which the distances between spatial points terminate when $X = 0$; and (2) an exponential scale parameter, Φ, which accelerates or decelerates the baseline hazard, depending on how the independent factors in the vector X influence the termination rate. Both the shape and scale parameters must be estimated via maximum likelihood.

In most forms of spatial analysis, the major challenge is to structure the experimental setting in a way that lines up not only with theory, but also with the logic of the particular analytical method itself. Regarding the analysis presented herein, urban economic theory (Fujita, 1987) clearly indicates that the hazard function for distance separating the spatial points that make up urban areas, whether structures, small-area population centers, or something else, should exhibit positive spatial dependence—but that the hazard decelerates with distance from the interior of the region (Carruthers et al., 2008). Based on this theoretical framework, a Weibull-distributed spatial hazard model of urbanization takes the following form:

$$h(d_{ij} \mid X_{ik}) = h_0(d_{ij}) \cdot \exp(\phi_{d_c} \cdot x_{d_{ic}} + X_{ik} \cdot \Phi_k). \qquad (2)$$

Here, $h(d_{ij} \mid X_{ik})$ indicates that the baseline hazard for distance between nearest neighbors *i* and *j*, $h_0(d_{ij})$, is scaled by X_{ik}, a vector of *k* independent variables, including $x_{d_{ic}}$, the distance from *i* to the regional center; and Φ_k (including ϕ_{d_c}) measures the influence the vector of additional independent variables has on the conditional probability of distance between nearest neighbors terminating. The two hypotheses at the heart of this model, both of which flow directly from urban economic theory, are (1) the conditional probability of distance between nearest neighbors terminating increases with distance and (2) the probability of terminating decelerates with distance from the interior of the region. The two variables essential for estimating this model are d_{ij} and $x_{d_{ic}}$, the distance from *i* to its nearest neighbor and the distance from *i* to the regional center, respectively. The remainder of this note is dedicated to explaining the process of generating those variables.

Data Collection

The modeling framework previously described has been applied in two empirical analyses of urban form and settlement patterns in the United States (Carruthers et al., 2008; Lewis et al., 2008). The analysis discussed herein by Carruthers et al. (2008) examines the viability of using spatial hazard models to study urban form via an analysis of point patterns in the nation's 25 largest core-based statistical areas (CBSAs). Where applicable, CBSA divisions are used in place of the greater CBSA, so the model considers a total of 43 distinct areas. The unit of analysis is the 2000 Census Block Group and the data come from only two sources: (1) a nationwide count of housing units at the census block level in 2006, which the Census Bureau provided to the U.S. Department of Housing and Urban Development; and (2) Census Summary File (SF) 3 from the 2000 Census of the population.

Spatial Point Patterns

Spatial point patterns representing the urban fabric of the 43 metro areas being analyzed and relevant distance measurements were developed via a six-step process using a variety of software tools, including ESRI's ArcInfo GIS and GeoDa—a program developed for spatial analysis and computation (Anselin, 2003; Anselin, Syabri, and Kho, 2006).

In the first step, the 2006 housing unit count was joined to each of the 8,205,582 blocks in the United States (territories were excluded). Hardware—that is, PC computing power—constraints required the data to be split into sections to make the data processing possible. The blocks were split into 10 separate, subnational regions and a shapefile was created for each one. The geographic centroid of each census block was used to generate a mean center for all 208,643 block groups in the United States. This so-called *mean center* is defined as the population-weighted average Cartesian $\{x, y\}$ coordinate of all the block centroids in a given block group. The mean center was calculated using the mean center tool in the ArcGIS Spatial Statistics Toolbox. The 2006 housing unit count was used as the "weight" field and the block group identification number was used as the "case" field, which groups each block into the correct block group and evaluates that block group accordingly. The output resulted in 10 new shapefiles, one for each subnational region, containing the points that specify what can be thought of as the "center of gravity" of each block group in the country. At the completion of this step, the 10 regional shapefiles were then merged into a single national shapefile.

In the second step, the same mean center routine was run to generate housing unit-weighted centers for the 939 CBSAs in the country. This step differed from the first in that the CBSAs themselves were used as case fields to separate mean center calculations; therefore, the points produced indicated the regions' housing unit-weighted centers of gravity. It is important to note that the CBSAs' mean centers differ from the traditional central business districts and, therefore, may end up in locations not consistent with the areas of the highest residential density, which are normally just outside the central business district. The CBSA mean centers indicating centers of gravity are desirable because the object of spatial hazard analysis is to examine the distribution of spatial point patterns relative to their geographic center, as opposed to a central business district. By way of contrast, density gradients—another, more common econometric method of analyzing regional land use—emanate from the central business district/s (see McDonald, 1988, for a review).

In the third step, each block group's mean center point was assigned to its nearest CBSA mean center point via the ArcMap's *near* routine. The near routine measures the distance from each point in the "input features" file—block group mean centers—to the nearest point in the "near features" file—CBSA mean centers. This process yielded the block group's mean center distance in meters to the respective regional center and the relevant regional center's identity. It is important to note that a block group's nearest CBSA may not be the CBSA that it formally belongs to.

In the fourth step—which accomplished the same distance measures, but for the nearest neighbor block groups—GeoDa had to be used because the ArcGIS Toolbox apparently has no routine that will identify a feature's nearest neighbor and calculate a distance to that feature; that is, the ArcGIS Toolbox cannot identify nearest neighbors and measure distance between nearest neighbors within the same shapefile. To get around this deficiency, GeoDa's spatial weights-creating function, which locates k nearest neighbors, was used instead. Specifically, the block group point file was imported into GeoDa, which was used to generate a spatial weights matrix that assigned each block group point a single nearest neighbor. The output file from this operation is a text file with a GWT file extension and, for each observation, the file lists (1) its own identification number, (2) the nearest neighbor's identification number, and (3) the distance separating the two points. The GWT was imported into a Microsoft Access database, where it was transformed into a database file format and then exported and, ultimately, joined with the block group mean center shapefile. To be clear, the result of these machinations was an output file identifying nearest neighbors and the distances between them. Although admittedly ad hoc, this process proved to be a good workaround for the ArcGIS Toolbox's inability to match nearest neighbors within the same shapefile.

In the fifth step, rays connecting each block group to its CBSA and nearest neighbor were drawn using an ESRI user-written extension, Desire Line—a tool that creates a line between a point of origin and a point of destination. These rays also give the distance measures, which were used to validate the nearest neighbor routines run in ArcGIS and GeoDa. The results of this step are shown in exhibit 1—a map of CBSAs and their spheres of influence, including the 43 regions that are the focus of the analysis shown in dark gray—and in exhibit 2— a map of spatial point patterns in the Chicago, Dallas, Los Angeles, and New York regions. In the latter exhibit, both the rays connecting block groups to their regional center and the rays connecting nearest neighbor block groups are visible. The outcome of the preceding steps described was essentially two variables: distance to the nearest CBSA mean center and distance to the nearest block group mean center for each of the block groups.

In the sixth and final step, additional variables involved in the analysis, which were added easily by linking data sets, consist of median household income, median age of housing, and average duration of journey to work. These data were obtained from the Census SF 3 and joined with the previously described variables based on the common block group ID. In the proportional hazard model, distance to the nearest neighbor is modeled using the distance to the nearest CBSA and other census variables as the control variables.

Exhibit 1

Core-Based Statistical Areas and Their Spheres of Influence

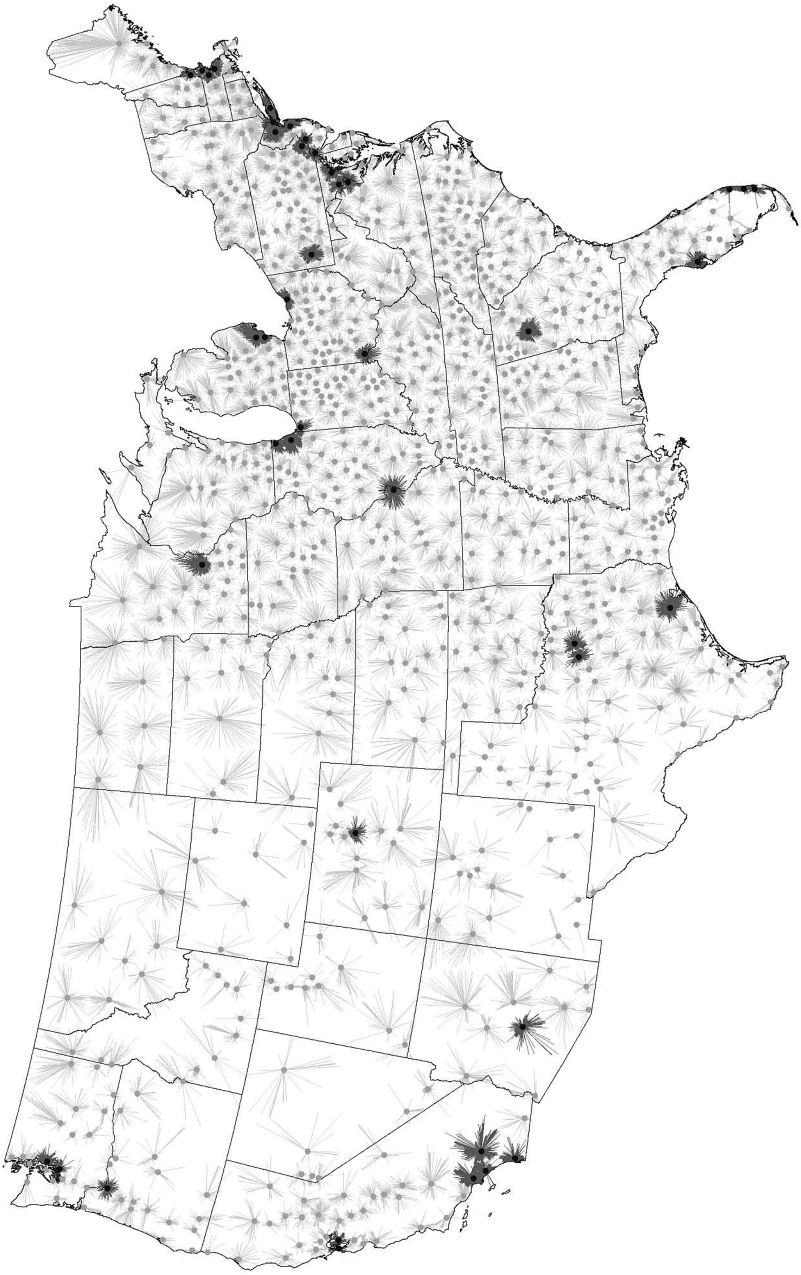
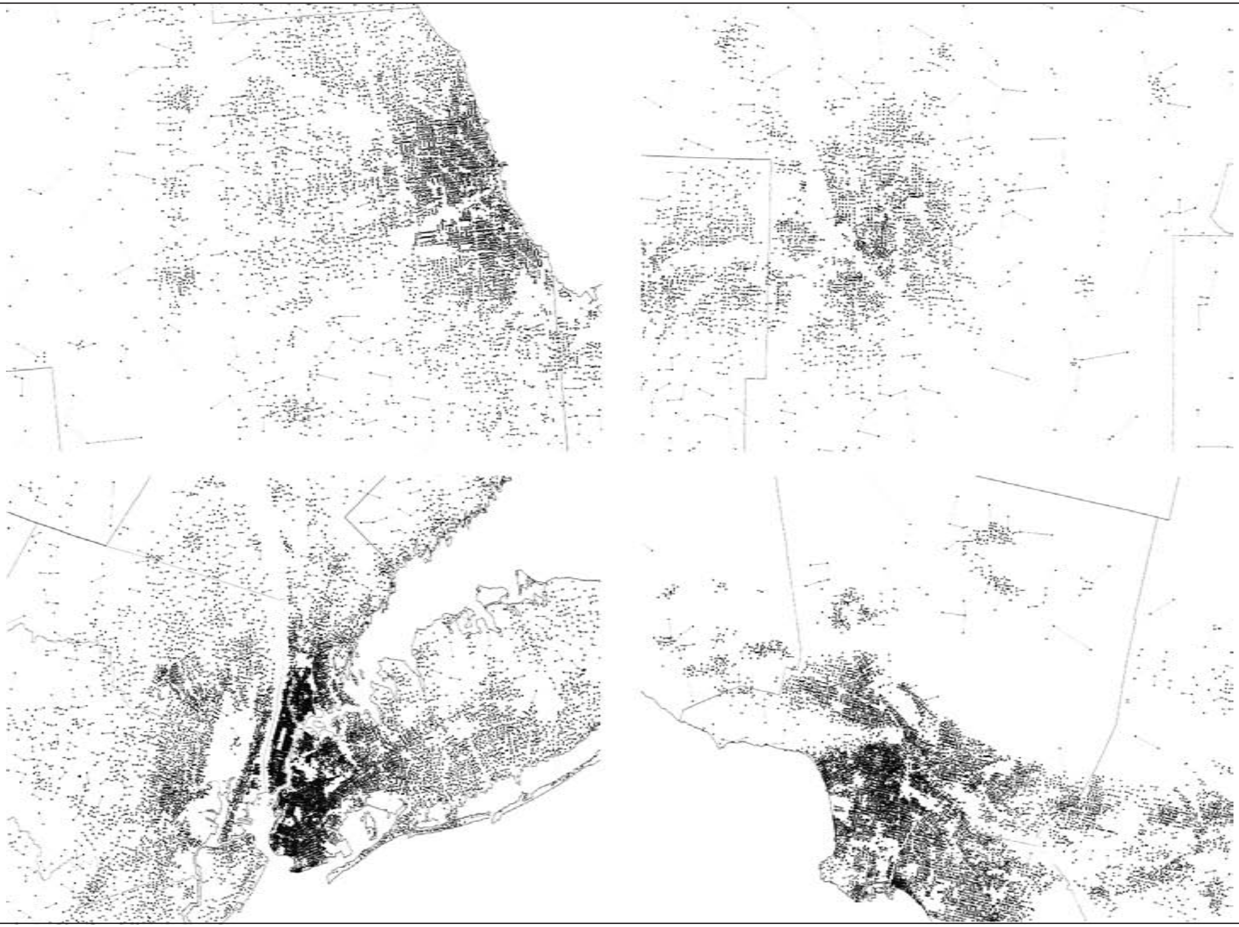


Exhibit 2

Spatial Point Patterns in the (clockwise from upper left) Chicago, Dallas, Los Angeles, and New York Regions



Summary and Conclusion

This note has detailed the process of constructing a spatial data set for studying urban form and settlement patterns via spatial hazard models. As noted, a key trick in spatial analysis is to structure the experimental setting in a way that lines up not only with theory, but also with the logic of the particular analytical method itself. The strategy presented in this note is one example of how such a challenge has been overcome.

Authors

Robert N. Renner is a social science analyst at the U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

Selma Lewis is a PhD research assistant at the University of Maryland, National Center for Smart Growth Research and Education.

John I. Carruthers, corresponding author, is an economist at the U.S. Department of Housing and Urban Development, Office of Policy Development and Research, and is center affiliate at the University of Maryland, National Center for Smart Growth Research and Education.

Gerrit-Jan Knaap is a professor and director at the University of Maryland, National Center for Smart Growth Research and Education.

References

Anselin, Luc. 2003. *GeoDa 0.9 User's Guide*. Urbana-Champaign, IL: University of Illinois, Spatial Analysis Laboratory.

Anselin, Luc, Ibnu Syabri, and Youngihn Kho. 2006. "GeoDa: An Introduction to Spatial Data Analysis," *Geographical Analysis* 38: 5–22.

Boots, Barry N., and Arthur Getis. 1988. *Point Pattern Analysis*. Newbury Park, CA: Sage Publications.

Carruthers, John I., Selma Lewis, Gerrit-Jan Knaap, and Robert N. Renner. 2008. Coming Undone: A Spatial Hazard Analysis of Urban Form in American Regions. Working paper REP 08-03. Washington, DC: U.S. Department of Housing and Urban Development.

Cleves, Mario, William W. Gould, and Robert G. Guitierrez. 2005. *An Introduction to Survival Analysis Using Stata*. College Station, TX: Stata Press.

Diggle, Peter J. 2003. *Statistical Analysis of Spatial Point Patterns*. New York: Hodder Arnold.

Esparza, Adrian, and Andrew J. Krmenec. 1996. "Business Services in the Space Economy: A Model of Spatial Interaction," *Papers in Regional Science* 73: 55–72.

Fujita, Masahisa. 1987. *Urban Economic Theory: Land Use and City Size*. Cambridge, UK: Cambridge University Press.

Kiefer, Nicholas M. 1988. “Economic Duration Data and Hazard Functions,” *Journal of Economic Literature* 26: 646–679.

Lawless, Jerald F. 2002. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience.

Lewis, Selma, John I. Carruthers, Gerrit J. Knaap, and Robert N. Renner. 2008. The American Way of Land Use: A Spatial Hazard Analysis. Working paper REP 09-03. Washington, DC: U.S. Department of Housing and Urban Development.

McDonald, John F. 1988. “Econometric Studies of Urban Population Density: A Survey,” *Journal of Urban Economics* 26: 361–385.

Odland, John, and Mark Ellis. 1992. “Variations in the Spatial Pattern of Settlement Locations: An Analysis Based on Proportional Hazards Models,” *Geographical Analysis* 24: 97–109.

Pellegrini, Pasquale A., and Steven Reader. 1996. “Duration Modeling of Spatial Point Patterns,” *Geographical Analysis* 28: 217–243.

Reader, Steven. 2000. “Using Survival Analysis To Study Spatial Point Patterns in Geographical Epidemiology,” *Social Science and Medicine* 50: 985–1000.

Waldorf, Brigitte S. 2003. “Spatial Point Patterns in a Longitudinal Framework,” *International Regional Science Review* 26: 269–288.