GeoJournal **61:** 53–67, 2004.

Spatial statistics for urban analysis: A review of techniques with examples

Antonio Páez* and Darren M. Scott
*Centre for Spatial Analysis, School of Geography and Geology, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1 Canada; *Author for correspondence (E-mail: paezha@mcmaster.ca)*

Key words: Geographic Information Systems, modifiable areal unit problem, spatial association, spatial heterogeneity, spatial statistics, urban analysis

Abstract

Traditionally, urban analysis has been quick to adopt and benefit from developments in technology (e.g., microcomputer, GIS) and techniques (e.g., statistics, mathematical programming). This has not been the case, however, with newer methods of spatial analysis – in particular, spatial statistics. Only recently has this situation started to change. This paper documents the confluence of spatial statistics and urban analysis by first reviewing developments in spatial statistics, and then presenting examples of recent applications in urban analysis. The developments reviewed fall under the rubric of global and local forms of spatial analysis, and cover three major technical issues: spatial association, spatial heterogeneity and the modifiable areal unit problem. The examples highlight the relevance and usefulness of the techniques reviewed for urban transportation and land-use applications. The paper concludes with conjectures concerning future developments at the intersection of spatial statistics and urban analysis.

Introduction

The analysis of urban systems is a discipline that boasts a long history using the computer for modeling and simulation. As recounted by Harris (1985), urban simulation, since its inception in the 1950s, was enabled and stimulated by developments in digital computing. The influence of the computer, moreover, was both direct and indirect through the role that digital computing played in making possible the present scope of statistical and other methods used in research (e.g., mathematical programming). The history of urban analysis is thus intertwined with that of the computer, on the one hand, and the development of statistical and mathematical methods on the other. A recent expression of the long tradition in urban analysis of rapidly embracing technological developments can be observed in the adoption of geographic information systems (GIS). In the last 15 years or so, a substantial body of research that explores the role and potential of GIS to support various forms of urban analysis and planning has accumulated (e.g., Levine and Landis, 1989; Harris and Batty, 1993; Innes and Simpson, 1993). Indeed, as numerous recent examples show (including a special issue of this journal), today GIS is routinely used in many aspects of urban research (see, inter alia, Vichiensan *et al.*, 2001; Okunuki, 2001; Du, 2001; Abed and Kaysi, 2003; Liu and Zhu, 2004; Barredo *et al.*, 2004).

The advantages of using GIS for urban analysis are many. GIS, as a database management tool, offers forward data mapping functions for displaying geographical information, and backward data retrieval functions for 'querying' maps (Levine and Landis, 1989). These 'front end' and 'back end' operations allow analysts and planners to better manage, display and communicate information (Miller, 1999). The power of these functions, moreover, is augmented by techniques for interactive data modeling (e.g., cartographic analysis, data conversion routines), which can enhance urban transportation analysis (e.g., Wang and Cheng, 2001; Stanilov, 2003) and land-use analysis (e.g., Landis and Zhang, 1998a,b). In some cases, using GIS for urban analysis may even enable a researcher to break 'free' from the 'tyranny of zones' (Spiekermann and Wegener, 2000). However, despite its considerable advantages, GIS in and by itself does not 'free' the analyst from the necessity of coping with the nuances, complexities and subtle relationships inherent in the spatial data commonly used in urban research.

Regarding the complexities of modeling spatial data, a recent paper by Fotheringham (2000) poses the question of whether the adoption of GIS has represented a step forwards or a step backwards for spatial modeling. In answering the question, he asserts that, to date, most GIS-based modeling represents a step backwards – although commercial GIS packages now incorporate spatial models, in addition to cartographic analysis techniques, these models tend to be outdated and far away from the research frontier. In the case of urban analysis, the adoption of GIS has given the strong impression that the discipline embraces space in modeling. It has been noted, however, that the statistical models underlying conventional urban analysis remain, for the most part, aspatial (Landis and Zhang, 2000), and thus ignore important issues such as the failure of most conventional statistics to adequately summarize locational information (the sufficiency criterion; Griffith, 1988), the lack of independence and inherent stickiness of spatial data (e.g., spatial association), differential effects in spatial processes (e.g., heterogeneity), and the implications of shape and representation in spatial analysis. It is thus surprising, given the seriousness of these issues (e.g., Anselin and Griffith, 1988), that compared to the swift adoption of GIS, urban analysts have been slower to embrace technical developments in spatial statistics. This is, however, starting to change.

The objective of this paper is to document the confluence between urban analysis and spatial statistics, which is now becoming apparent. To this end, a number of developments in spatial statistics are reviewed, and recent examples showing how these techniques can be put to work in urban analysis are presented. Three major technical issues are covered (i.e., spatial association, heterogeneity and the modifiable areal unit problem) from the perspective of global and local forms of spatial analysis (i.e., methods that seek to identify, explore and model large scale relationships, or rules, and variation at the local scale, or exceptions; see Fotheringham and Brunsdon, 1999). In order to highlight the relevance and usefulness for urban analysis of the techniques reviewed, examples are drawn from the recent literature to illustrate how spatial statistics can fruitfully assist in the task of analyzing processes in both urban transportation and urban land use. Although the potential of spatially explicit methods remains yet to be realized by many urban analysts and modelers, the examples presented in this paper indicate an incipient and promising trend towards the application of increasingly sophisticated spatial statistical methods for urban analysis.

Urban processes and spatial processes

Urban analysis has been defined as the use of multidisciplinary knowledge and skills with the objective of solving urban problems (Pacione, 1990). Although valid as a statement of purpose, a more targeted definition of urban analysis is required to delineate the scope of this review, and to differentiate it from other thematic areas that could benefit from the use of spatial analysis, but are beyond the scope of the present paper (e.g., health research (e.g., Robinson, 2000; Jerrett *et al.*, 2003), urban crime (e.g., Craglia *et al.*, 2000; Craglia *et al.*, 2001; Ceccato *et al.*, 2002) and political-historical processes (e.g., Flint, 2002)).

The primary focus in this review is on the analysis of urban systems. The term 'urban systems analysis' can be applied to the study of an individual city, conceptualized as a collection of various interrelated components. Typically, these components include an activity sub-system that determines a city's land-use configuration, a transportation sub-system and the interactions between these components (Black, 1981; De la Barra, 1989; Kanaroglou and Scott, 2002). More recently, urban analysts have paid increasing attention to the urban environment and the effect of land-use and transportation activities on it (a comprehensive list of urban sub-systems and their interactions is provided by Moeckel *et al.*, 2003). Processes of interest in urban analysis include different types of construction (residential, industrial, transportation infrastructure), economic and demographic change, mobility (travel, residential choice and freight) and environment-related processes, such as energy consumption, emissions and noise. Methodological issues include the definition of units of analysis, which could be aggregated (e.g., Traffic Analysis Zones or TAZs) or disaggregated (e.g., individuals or households).

A characteristic of most urban processes is the fact that they are intrinsically spatial and, moreover, space-dependent. Consider, for example, the following description of a process observed in Britain and the United States in the mid-1980s (Hall, 1986):

> "Starting with the biggest cities and spreading gradually down to the smaller ones, progressively the growth of cities tends to slow down. Thus, eventually, they actually start to lose people to their suburbs. A little later, they start to lose out to smaller towns in their immediate vicinity, and then to more distant places. Eventually, they start to contract; the growth passes to the smaller towns and to the countryside."

Urban growth and decline, according to this depiction, are spatially-conditioned processes, and the outcome at one location is partially affected by events at other locations. The spatial concepts of contiguity (e.g., cities start to lose people to their suburbs), proximity (e.g., population is lost to smaller towns in the vicinity) and/or connectivity (e.g., the process starts with the biggest cities and gradually spreads down) are important elements of these processes. Several spatial processes relevant to urban analysis have been identified in the literature on spatial data analysis (see, for example, Haining, 1990; Krieger, 1991; and Landis and Zhang, 2000):

- Spatial diffusion. Diffusion is the gradual adoption of a new attribute by a fixed population. Typically, the likelihood of adoption is influenced by distance to previous adopters or place in a hierarchy, with the consequence that diffusion results in geographical patterns. The process could occasionally be physical in nature (e.g., land development).
- Exchange and transfer. Urban economies are bound together by commodity exchange and income transfer. The spatial structure of some social and economic variables may be a reflection of spillover effects, which usually are evident from the data in the form of spatial association. Agglomeration economies, on the other hand, may produce differential effects over space.
- Spatial interaction. Interaction, as a spatial process, has received considerable attention, especially in the transportation literature. Although this process is usually conceptualized in terms of physical movement of people or commodities, information flows can link events at spatially dispersed locations.
- Segmentation. The partition of a formerly homogeneous region into two or more sub-regions each with clearly unique characteristics is an example of segmentation. In an urban context, spatial segmentation may relate to agglomeration economies, industrial, commer-

cial and residential stratification, and racial or social self-selection, among other processes.

The existence of temporal dynamics in urban processes has long been a topic of interest in urban analysis (e.g., Wegener *et al.*, 1986). As hinted by the brief description of the above spatial processes, the existence of what may be termed 'spatial effects' is also important. In fact, spatial processes that translate as clustering and/or dispersion, or as systematic variability across space, violate basic assumptions of independence and homogeneity implicit in conventional statistical analysis. Violation of these assumptions, in turn, leads to information loss, biased and/or inefficient parameters and the possibility of seriously flawed conclusions and policy prescriptions (Griffith and Layne, 1999). Given these potential pitfalls, it is not surprising that spatial effects have often been regarded as nuisances. Increasingly, however, they are perceived rather as opportunities to obtain deeper insights regarding the processes under study. Regardless of the view espoused, there are strong methodological and conceptual arguments favoring the adoption of spatial statistical methods for the analysis of urban processes. The reminder of this paper reviews some major technical issues in spatial statistics and presents substantive examples of application in urban analysis.

Spatial processes: Some major analytical issues

Three major analytical issues, namely spatial association, heterogeneity and the modifiable areal unit problem (MAUP), are reviewed in this section, along with techniques to deal with them.

Spatial association

One of the major concerns in the analysis of spatial data, association is the tendency of variables to display some degree of systematic spatial variation. In urban studies, this often means that high variable values are found near other high values and low values appear in geographical proximity. Sometimes, however, the effect may have a negative quality when the ordering reflects systematic dissimilarity among neighboring observations. Spatial association may be caused by a variety of spatial processes, including, among others, interaction, exchange and transfer, and diffusion and dispersion. It can also result from missing variables and unobservable measurement errors in multivariate analysis.

An important feature of spatial association is a form of serial arrangement similar to the one in the analysis of time series. This feature invalidates the assumption of independence, and compromises the applicability of conventional statistics, which may lead to biased and inconsistent estimates. In what follows, a general approach to modeling spatial association is presented. This is followed by reviews of two exploratory techniques for the local analysis of spatial association, namely Getis and Ord's distance-based statistic (Getis and Ord, 1993; Ord and Getis, 1995) and Anselin's (1995) local decomposition of a global statistic of spatial association.

Spatially autoregressive models

A well-known technique in spatial statistics is the model with spatially autoregressive components. This approach to modeling spatial association is primarily an outcome of developments in geography (e.g., Cliff and Ord, 1981; Griffith, 1988; Haining, 1990), a field often concerned with the analysis of areal units (e.g., census tracts) or network data (e.g., nodes in a network). Although hardly a new addition to the spatial analysis toolbox (a version of it was first proposed by Whittle in 1954), it was not until relatively recently that this model started to find substantial application in urban analysis. A widely used specification is the one proposed by Anselin (1988), which can be written in matrix form as follows:

$$\boldsymbol{Y} = \rho \mathbf{W}\mathbf{Y} + \mathbf{X}\beta + \varepsilon, \tag{1}$$

$$\varepsilon = \lambda \mathbf{W}\varepsilon + \mu. \tag{2}$$

This model is a generalization of the linear regression model. The elements of the model are a vector $\boldsymbol{Y}(n \times 1)$ of objective variable observations, a matrix $\mathbf{X}(n \times K)$ of independent observations that include the usual constant and a vector $\beta(1 \times K)$ of parameters corresponding to K independent variables. Scalars ρ and λ are parameters of spatial association corresponding to the objective variable and the error terms ε, respectively, while μ are independent and possibly homogeneous error terms.

An important element of the model is the spatial lag operator $\mathbf{W}$, which is simply a matrix $(n \times n)$ containing weights w_{ij} that describe the degree of spatial relatedness (i.e., contiguity, proximity and/or connectivity) between units of analysis i and j. Matrix $\mathbf{W}$ can be defined in different ways. One basic definition is based on physical contiguity with binary weights that assign a weight of 1 to pairs of zones sharing a border and 0 otherwise. In addition to, or in place of contiguity, connectivity can be given in terms of travel between pairs of origins and destinations, inventories of material flows or flows of goods (Peeters and Ruhigira, 2004, discuss use of the 'transportation problem' to measure proximity). Alternatively, proximity can be defined in terms of distance or various accessibility measures, such as travel time or generalized costs. Technical details concerning the matrix of spatial weights are found in Cliff and Ord (1981).

In the above modeling framework, spatial association may appear under two different guises. Spatial association of the objective variable (i.e., the association between y_i and spatially-related observations y_j) can sometimes be interpreted as a spatial economic externality (e.g., Murdoch *et al.*, 1997). Spatial error autocorrelation, on the other hand, is basically a statistical nuisance and is best described as a proxy for missing variables that follow a meaningful spatial pattern. Regardless of the form of association, the model seeks to explain the covariation between $\boldsymbol{Y}$ and $\mathbf{X}$, ideally based on some substantive theory of urban processes.

A commonly adopted two-stage modeling strategy begins by estimating a simple aspatial regression model that is then tested by means of a statistic of spatial association (e.g., Moran's I; see Cliff and Ord, 1981) or a model-based statistic (e.g., the Lagrange Multiplier tests of Anselin, 1988).

Failure at this stage to find significant spatial association leads to the conclusion that the simple model accounts for all (i.e., spatial and other) observed variability, and the analyst would not need to estimate spatial models (see, for example, the analysis by Briggs *et al.*, 1997). Indication of spatial association, on the other hand, would be followed by a second stage where a spatial model is estimated. In this process, different definitions of matrix **W** could be seen as alternative hypotheses regarding the process of interest. Detailed discussion regarding estimation and inference of this type of model can be found in Anselin (1988).

Local statistics of spatial association

Originating as part of a recent trend in spatial analysis that emphasizes the study of local spatial effects, two local statistics of spatial association have received considerable attention in the last 10 years: the Getis and Ord family of $G_i(d)$ statistics (Getis and Ord, 1993; Ord and Getis, 1995, 2001) and Anselin's LISA (Local Indicators of Spatial Association; 1995). Both statistics are designed to summarize the level of local spatial association, and are thus useful in detecting places with unusual concentrations of high or low values (i.e., 'hot' or 'cold' spots). They accomplish this in very different ways.

The $G_i(d)$ is a distance-based statistic that measures the proportion of a variable found within a given radius of a point, respective to the total sum of the variable in the study region. The statistic is defined for location i as follows:

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{i=1}^{n} x_i}, \tag{3}$$

where x_j is the value of the observation at j, $w_{ij}(d)$ is the ij element of a binary **W** matrix with ones for all sites within distance d of location i and zeros elsewhere (including the diagonal), and n is the number of observations. Matrix **W** in this case formalizes the concept of proximity in geographical space. The mean and the variance of this statistic can be obtained from a randomization process, and used to derive a standard statistic (i.e., a Z-score). It is assumed, for inferential purposes, that the statistic is approximately normally distributed. This is reasonable when n (the number of points in the sample) is large and the distance d in equation (3) is not too small (i.e., as to encompass no observations within the radius) or too large (i.e., as to encompass all the observations in the study area). If the (absolute) value of the standardized statistic is greater than the cutoff value at a pre-specified level of significance, then positive or negative spatial association is said to exist. Positive values of the statistic are interpreted as a spatial agglomeration of relatively high values (more than would be expected by chance), while negative values represent relatively low values clustered together.

The $G_i(d)$ statistic gives the proportion of a variable (the sum of values) within distance d of location i, and thus intuitively provides a measure of the concentration (or lack thereof) of values around a given location. The statistic is useful to reveal spatially homogeneous locations in terms of the concentration of high or low values. By being distance-based, it provides a very flexible way of studying local spatial association that works with positive variables that have a natural origin. Flexibility, on the other hand, carries a price, as standard rules to guide the selection of distance d for analysis remain somewhat underdeveloped.

A different approach to the local analysis of spatial association is decomposing a global statistic of spatial association. Take, for example, the well-known statistic Moran's I:

$$\begin{aligned} I &= \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j wij(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2} \\ &= \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum \sum w_{ij}\hat{x}_i\hat{x}_j}{\sum_i \hat{x}_i^2} \end{aligned} \tag{4}$$

where $\overline{x}$ is the mean of variable vector $\boldsymbol{x}$ and $\hat{x}_{i,j}$ are deviations from the mean. In Moran's I, spatial association is measured as the covariance between values at spatially-related locations (w_{ij} are the elements of a spatial lag operator **W**), standardized by measures of total variation ($\sum_i \hat{x}_i^2$) and connectivity in the system ($\sum_i \sum_j w_{ij}$). When values are interrelated in meaningful spatial patterns, similar values (in deviations from the mean) are found at neighboring locations (i.e., positive spatial autocorrelation), but when dissimilar values are found at neighboring locations, negative spatial association is said to ensue. Zero association implies a spatially random set of observations.

The local version of Moran's I is given by the following expression (Anselin, 1995):

$$I_i = \hat{x} \sum_j w_{ij}\hat{x}_j. \tag{5}$$

The sum of (5) over all i is proportional to the global statistic in (4). As with the $G_i(d)$ statistic, it is possible to derive the mean and the variance of I_i based on a randomization procedure, and inference can be carried out by obtaining a normalized statistic $Z(I_i)$.

Interpretation of the LISA statistic is less intuitive than interpretation of the $G_i(d)$ statistic. In general, there are four patterns of local spatial association:

- High-high association. When the value of x_i is above the mean and the values of x_j at 'neighboring' zones (i.e., $\sum_j w_{ij}\hat{x}_j$) are generally above the mean, the statistic is positive;
- Low-low association. When both values are below the mean, the statistic is positive;
- High-low association. When the value at i is above the mean and the values at neighboring zones are, in general, below the mean, this gives a negative I_i; and
- Low-high association. When the value at i is below the mean and the weighted average is above the mean, I_i is negative.

These patterns are not readily apparent from the sign of the I_i statistic alone, but they can be gleaned from a Moran's scatterplot tool (Griffith and Layne, 1999; pp. 15–18). The combination of LISA and the scatterplot tool provides detailed information on different types of spatial association at the local level. LISA statistics can also be aggregated to produce a global measure of spatial association, and thus can be used to identify influential locations in spatial association analysis. In addition, the definition of matrix **W** makes it ideal to work with area data, although in other situations this definition could impose a fairly rigid spatial structure on the observations.

Spatial heterogeneity

Urban processes often exhibit patterns of spatial heterogeneity – that is, they do not always operate in exactly the same way over space. Spatial heterogeneity is frequently thought to result from large-scale regional effects or administrative subdivisions that delimitate the reach of some process (e.g., zoning). Heterogeneity may lead to biased parameter estimation, misleading significance levels and/or sub-optimal forecasts. Thus, it should receive serious attention in urban analysis. Among different potential issues with spatial data identified by Griffith and Layne (1999), simulations suggest that heterogeneity is potentially the most damaging effect (pp. 71–73). Heterogeneity is commonly seen as a statistical nuisance. Recently, however, with an increased interest in the role of context and locality in spatial and urban analysis, the role of spatial heterogeneity is being re-examined, and the methods used to detect and model it are receiving renewed attention.

In statistical terms, spatial heterogeneity can be represented as structural variation in the definition of the variance or as systematic variation in the mean of the process. Three models are reviewed below that deal with spatial heterogeneity, namely switching regressions, multilevel models and geographically weighted regression or GWR. Of these methods, the first two constitute a compromise between global-local modeling. GWR on the other hand is a local form of spatial analysis.

Switching regressions

A method originally developed by Quandt (1958), switching regressions operate on the principle of classifying a dataset into a number of mutually exclusive and collectively exhaustive regimes. Switching regressions pose a solution to the problem of heterogeneity that is particularly suitable in situations where data can be classified into a small number of regimes. Assuming that a dataset can be divided in two classes (e.g., north and south in a city, or CBD and suburbs), regression data could be re-arranged to obtain the following expression (in matrix form):

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}. \tag{6}$$

In the above, two vectors of dependent variables $\boldsymbol{Y}_1 (n_1 \times 1)$ and $\boldsymbol{Y}_2$ ($\mathrm{n}_2 \times 1$) contain observations that correspond to two different spatial regimes ($n_\#$ is the number of observations in each regime). In a similar way, the set of independent variables is rearranged to match the locations represented by vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, resulting in a matrix of dimensions $(n_1 + n_2) \times 2k$. Vectors β_1 and β_2 include the parameters corresponding to each of the two regimes, and μ_1 and μ_2 are the error terms of the model. Redefining the variables, this specification is identical to the linear regression model:

$$\boldsymbol{Y}^+ = \mathbf{X}^+ \beta^+ + \varepsilon^+. \tag{7}$$

Unlike the common regression model that assumes homogeneity, the structure of the covariance in a switching regression is defined as:

$$E[\mu^+ \mu^{+\prime}] = \Omega = \begin{bmatrix} \sigma_1^2 \mathbf{I}_1 & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_2 \end{bmatrix} \tag{8}$$

to give location-specific variance parameters σ_1^2 and σ_2^2 ($\mathbf{I}_1$ and $\mathbf{I}_2$ are identity matrices of size n_1 and n_2, respectively). The method can be easily generalized to more than two spatial regimes, with the total number constrained by the incidental parameter problem (i.e., the number of parameters must be less than the number of observations). In practical terms, applications must keep the total number of parameters, including variance parameters, within manageable limits.

The switching regression paradigm is a fairly simple modeling technique that bridges the gap between global and local analysis. Instead of estimating a set of blanket parameters that apply equally to all the study area, parameters are specific to each regime. Parameters are thus heterogeneous between regimes (inter-class heterogeneity), but are homogenous with a given regime (intra-class homogeneity).

Multilevel models

Multilevel models have been proposed in geographical research as a way to model spatial heterogeneity (Jones, 1991; Duncan and Jones, 2000). The method, also known as hierarchical modeling, operates on the principle of expanding the parameters of a model using random components as part of the expansion (Jones and Bullen, 1994), a concept that the following simple bi-level model illustrates:

$$y_{ij} = \alpha + x_{ij}\beta + \varepsilon_{ij}. \tag{9}$$

The subscripts indicate individual observations i (=1, 2,..., n) in the bottom level, and *groups* of observations in the top level ($j = 1, 2, \ldots, q$; $q \ll n$). Nests in the second level could be districts, schools, age groups or any other suitable classification. One way in which the parameters of equation (9) can be expanded is:

$$\alpha_j = \alpha_0 + \varepsilon_j^\alpha. \tag{10}$$

As result of this expansion, a parameter α_j, specific to class j in the second level is obtained. This parameter consists of two parts: a global element α_0, and a random term specific to category j. The terminal model becomes:

$$y_{ij} = \alpha_0 + x_{ij}\beta + (\varepsilon_{ij} + \varepsilon_j^\alpha). \tag{11}$$

All the non-systematic variation in the model is captured by the two error terms in parentheses. If these terms are normally distributed, the model can be fully specified by the set of parameters α_0 and β and one variance parameter for each set of random terms.

Equations (9) to (11) are a simple example of a bi-level expansion. However, more sophisticated expansions are possible. For example, deterministic elements can be added that use additional explanatory variables to produce linear or quadratic expansions (see Duncan and Jones, 2000). Similar to the switching regression framework, the parameters in a multilevel model vary between classes or zones, but are global or fixed within a given class or zone.

Geographically weighted regression

A different method to model spatial heterogeneity is geographically weighted regression (GWR), a local form of analysis proposed by Brunsdon *et al.* (1996) as a simple way of modeling complex spatial variation. The method is based on the idea of assigning weights to individual observations according to geographical distance from a location termed the focal point o. Underlying this idea is the well-known geographical concept of distance-decay, thus implying that the relative importance of particular observations decreases with distance from the focal point. In practical terms, the weighting scheme places a moving window over a spatially distributed set of observations, the influence of which is progressively discounted with distance from the center of the window. This scheme has the effect of producing sub-samples of data around specific points in space.

The GWR model takes the following form (see Brunsdon *et al.*, 1999):

$$y_o = \sum_{k=1}^{K} x_{ok}\beta_k(c_o) + \varepsilon_o, \tag{12}$$

where sub-index o indicates a focal point that needs not correspond to an actual observation of y. The error terms in the above expression (see McMillen, 1996, p. 105; Brunsdon *et al.*, 1999, p. 502) assume the usual conditions of independence and constant variance. The difference with the well-known linear regression model is that $\beta_k(c_o)$ are unspecified functions, not necessarily linear, of the geographical coordinates c_o of the focal point. Brunsdon *et al.* (1999) note that if the function $\beta_k(c.)$ is reduced to a constant, the ensuing model is identical to the ordinary least squares specification. In other cases, the analyst can evaluate the unspecified function at a given point in space to give local estimators of the function.

The moving window used to locally estimate the function $\beta_k(c_o)$ is defined by a monotonically decreasing function, or kernel, such as:

$$g = \exp(-\gamma d_{oi}^2), \tag{13}$$

where d_{oi} is the distance between the location of the focal point and observation i. A number of alternative kernel functions exist. It is commonly agreed, however, that selection of parameter γ (which controls the size of the window) and the number of observations in the sub-sample is more important than selecting a kernel function (McMillen, 2001, p. 470). There are at least two possible strategies for the selection of γ. One is based on a procedure known as cross-validation, essentially a goodness-of-fit criterion that minimizes the total sum of squared errors between the observed values and the values predicted by the model (e.g., Brunsdon et al., 1996). A second approach, which is adopted by McMillen (2001), adopts an arbitrary value of to give a predetermined window size.

GWR is often interpreted as a smoother (a predecessor is locally weighted regression, a scatterplot smoother proposed in statistics by Cleveland, 1979). Given this, GWR can be used to approximate to a very high level of accuracy the observed variable surface, a feature that makes it very attractive for various aspects of urban analysis.

Modifiable areal units

Urban analysis is often conducted using 'aggregate' geographical data – that is, data reported for pre-defined areal units such as census tracts and traffic analysis zones. A well-known spatial analytical issue that can influence the results of such studies is known as the modifiable areal unit problem (MAUP). The seriousness of this issue is clearly demonstrated by the wide variety of techniques whose results are known to be affected by it. These techniques include correlation analysis (Gehlke and Biehl, 1934; Blalock, 1964; Openshaw and Taylor, 1979), regression analysis (Fotheringham and Wong, 1991; Amrhein and Flowerdew, 1992; Amrhein, 1995; Okabe and Tagashira, 1996; Tagashira and Okabe, 2002), spatial interaction modeling (Openshaw, 1977b; Batty and Sikdar, 1982a,b,c,d; Putman and Chung, 1989), location-allocation modeling (Goodchild, 1979; Fotheringham *et al.*, 1995; Hodgson *et al.*, 1997; Murray and Gottsegen, 1997) and discrete choice modeling (Guo and Bhat, 2004) – all of which are commonly used in urban analysis. Furthermore, the MAUP has been shown to affect indices derived from areal data such as the segregation index D (Wong *et al.*, 1999) and the excess commute (Horner and Murray, 2002). Indices such as these are also used in urban analysis.

The MAUP arises in urban analysis due to the fact that an infinite number of zoning systems could be constructed to subdivide a city into smaller areal units. Of course, this implies that the data reported for areal units will differ between zoning systems. A direct consequence of this is that results will vary for studies using the same analytical technique, but different zoning systems for the same city (see, for example, Wong *et al.*, 1999; Horner and Murray, 2002). In other words, the outcome from a study using data reported for areal units of a specific zoning system is merely one manifestation from a range of possible outcomes.

The MAUP has two components: the scale effect and the zoning effect. The scale effect is a consequence of spatial aggregation – that is, results from the same analytical technique tend to vary according to the level of spatial resolution. Horner and Murray (2002), for instance, have shown that estimates of excess commuting decrease with increasing spatial aggregation. Other studies have documented increases in the values of correlation coefficients as spatial resolution becomes coarser (Gehlke and Biehl, 1934; Blalock, 1964; Openshaw and Taylor, 1979). The zoning effect, on the other hand, results from the multitude of zoning schemes that could be constructed and used at any given scale. Several studies have documented this effect (e.g., Fotheringham and Wong, 1991; Fotheringham *et al.*, 1995; Wong *et al.*, 1999; Horner and Murray, 2002). Basically, they show that for a given level of spatial resolution there exists a range of possible outcomes from an analytical technique owing to changes in spatial partitioning.

Clearly, the only way true frame-independence can be achieved in urban analysis – meaning that the results from the analysis do not depend on the zoning system used (Fotheringham, 1989; Tobler, 1989) – is to use individual-level data. Such data differ from areal data in that their locations are represented by geographical coordinates (e.g., latitude and longitude), not zones. This implies that the objects or phenomena under scrutiny are visualized as points in space. Trips, individuals, households, firms and land parcels are but a few objects in a city for which individual-level data are routinely collected by various public agencies. Such data are, however, difficult to obtain to support basic research (Fotheringham *et al.*, 2000) for one obvious reason – confidentiality must be preserved at all costs. This implies that if the data are released for public consumption, locational precision is often sacrificed. At best, locations are represented by zones. In some instances, however, locations are removed altogether. Given that we now live in a 'geocodable' world fueled by advances in spatial data capture technologies (e.g., global positioning systems and high-resolution satellites), which not only allow for realistic representations of urban environments and the recording of human spatial behavior, but also the exploitation of such knowledge for commercial purposes (e.g., for location-based services), the issue of privacy is likely to be revisited in the coming decade. In the event of such a possibility, which may preclude individual-level data from being released to researchers as is generally the case today, and in instances where such data simply do not exist, other methods are available for mitigating MAUP effects in urban analysis.

The simplest of these strategies is to employ the smallest areal units available (i.e., base zones) in an effort to report the sensitivity of analytical results to scale and zoning effects (Wong, 1996; Fotheringham et al., 2000). If it can be demonstrated that the results are stable over a wide range of zoning systems, then the analyst can express confidently that the results are meaningful and not simply artifacts of the way the data are partitioned in space (Fotheringham *et al.*, 2000). Obviously, this strategy requires that new zoning systems be created by merging base zones. Several researchers (Fotheringham and Wong, 1991; Murray and Gottsegen, 1997; Horner and Murray, 2002) have tackled this problem by using the Theissen region approach. This method randomly selects a user-specified number of 'seed' zones around which Theissen polygons are generated. Base zones are then merged with their closest seed zone to create new zones. Using this approach, multiple zoning systems (e.g., 100) are created for each of various scales of analysis, which are determined by the number of seed zones specified by the analyst. Although this procedure has been used repeatedly to address MAUP effects, its handling of the scale effect is to some extent arbitrary – that is, scale is synonymous with the number of areal units to be formed. It is assumed that as the number of seed units decreases, the size of the areal units will increase. While this may be true in terms of an average, it is clear that considerable heterogeneity will exist among areal units in terms of their size. A simple example illustrates this point. Assume that a city is partitioned initially into 200 base zones. New zoning systems, consisting of 175 and 150 zones, are to be formed. This implies that, at most, 25 zones in the first system will increase in size, while 150 will remain unchanged. For the second system, the numbers are 50 and 100, respectively. Most likely, the number of zones increasing in size will be fewer than the numbers given due to the randomness of seed selection – that is, the analyst has no control over the spatial distribution of the seed zones.

Spatial scale is, however, treated explicitly in a method developed by Wong (1999). Unlike the Theissen region approach, base zones are represented by their centroids, not their boundaries. Space itself is partitioned into regular square cells of a given size (i.e., grid cells). When more than one centroid falls into a cell, the data associated with the centroids are aggregated to form new data for the cell. To assess the scale effect of the MAUP, the size of the grid cells is changed systematically. To assess the zoning effect at any given scale, the grid system is positioned randomly many times in space. Its exact placement is determined by first moving it along two orthogonal axes in geographical space, such as latitude and longitude, and then rotating it. This method can be easily implemented using a GIS.

A more complex strategy for handling MAUP effects is to design optimal zoning systems. The idea here is to create zoning systems that are less arbitrary. In other words, zoning systems should be constructed to capture the underlying processes that are being investigated. Arguably, the most comprehensive and well-known work on this subject is that by Openshaw and his colleagues (Openshaw and Baxter, 1977; Openshaw, 1977a,b, 1978a,b; Openshaw and Rao, 1995). The automatic zoning procedure developed in this line of work solves a combinatorial optimization problem that classifies N areas into M regions (where $N > M$) so as to maximize some function of the zoning system subject to spatial contiguity constraints.

Arguably, the most complex of all strategies proposed to date for addressing the MAUP is that by Steel and Holt (1996), Holt *et al.* (1996) and Tranmer and Steel (1998). In essence, their method adjusts the aggregate-level variance-covariance matrix (i.e., that derived from zonal data) to better approximate that pertaining to the individual-level, which is unknown. The adjusted matrix can then be used in various statistical techniques, such as correlation analysis and regression analysis, to correct for aggregation bias. To work, the strategy requires a set of 'grouping' variables, which must be measured for individuals and must be related in some way to the process being analyzed at the aggregate level. These grouping variables account for one of two sources of aggregation bias identified by the researchers – the other being a residual bias conditional on the variables. Despite its intuitive appeal, the strategy does, however, suffer from two limitations. First, it requires individual-level data to estimate a variance-covariance matrix for the grouping variables. As discussed above, such data are not always available. Second, the strategy remedies the scale effect of the MAUP only.

Despite the promise of the methods described herein for mitigating MAUP effects in urban analysis, a general solution to the issue has yet to be found. This leaves room for researchers, such as King (1997), to seek solutions to what has been regarded as the most stubborn problem in spatial science to date.

Spatial statistics and urban analysis

Spatial analysis of transportation data and travel behavior

The analysis of transportation data and the study of travel behavior are necessary components of urban transportation planning. As noted by Ortúzar and Willumsen (2001), travel demand is characterized, among other things, as occurring over space. Indeed, space is treated explicitly by some methods used in transportation analysis. For instance, the gravity model is often used to distribute trips between origins and destinations in a city. The issues of spatial association, spatial heterogeneity and the MAUP, on the other hand, are not widely known to transportation analysts. For example, a survey of transportation modeling and planning books (e.g., Meyer and Miller, 2001; Ortúzar and Willumsen, 2001) and a recent perspective on GIS for transportation (Thill, 2000) shows that these issues have yet to permeate the transportation literature. In fact, it is only in a book by geographers (Miller and Shaw, 2001) that the explicit spatial analysis of transportation data is discussed. There is, however, hope that this situation is changing as there are now a few examples suggesting that the neglect of spatial issues in the analysis and modeling of transportation processes is finally being addressed.

Two approaches have been used by researchers to analyze transportation data and study travel behavior: aggregate analysis and disaggregate analysis. Aggregate analysis deals with group relationships. Outcomes of interest (e.g., number of trips per zone, proportion of trips by mode, etc.) are seen as resulting from a myriad of individual decisions. Aggregation and lack of data describing spatial structure can, however, give rise to spatial association and heterogeneity problems, not to mention MAUP effects.

The first two issues are recognized by Bolduc *et al.* (1995) who analyzed travel flows and modal split. Flow models are usually written in terms of the number of trips between pairs of zones (or transformations thereof), and a set of explanatory variables that include network attributes (e.g., in-vehicle travel time, cost) and socio-economic characteristics of the points of origin and destination (e.g. population, income, employment) As noted by Bolduc *et al.* (1995), the variables used to characterize a flow between zones and pertain to that market only, while the variables of competing markets are usually missing. Furthermore, other relevant variables describing the geography of the region (e.g., size of zones, length of common borders, etc.) are also typically ignored. Omission of these variables is significant because it can lead to spatial association, which in this case could be interpreted as a statistical nuisance (i.e., spatial error autocorrelation).

In order to address this situation, Bolduc *et al.* (1995) propose a model that blends an error components specification with spatial error autocorrelation (also see Bolduc *et al.*, 1989, 1992). The error components model is similar in principle to the multilevel model, except that errors reflect different sources of uncertainty as opposed to random parametric variation between spatial classes. Errors in the travel flow model are broken down into three elements that attempt to capture unmeasured effects at the origin, unmeasured effects at the destination and non-systematic network variability. Spatial error autocorrelation measures the degree of linear dependency between the error at the origin (destination/network element) and the corresponding spatial neighbors, with neighborhood defined in proximity terms using a distance-decay function. Application of the model to a case study using empirical modal split data from Winnipeg (Canada) suggests that misleading significance levels could result from ignoring spatial autocorrelation. A car-related cost variable, for example, becomes insignificant when autocorrelation is accounted for, and other variables are revealed to be generic and not specific to a given alternative (i.e., car or bus). Another significant finding is the indication, based on estimation results, that spatial error autocorrelation is caused by missing network variables, but not by unmeasured effects at the origins or destinations. Overall, the spatial model is found to give a better fit to the data compared to non-spatial models. An experiment using synthetic data further suggests that substantial information gains can be obtained when the proper error structure is used in estimation.

Aggregate models are, for a number of reasons, including data availability, the state-of-the-practice. More recently, an alternative approach known as activity analysis has received increased attention as a way to study disaggregate (i.e., individual or household) behavior, as opposed to the aggregate behavior of groups of people. Travel in this framework is seen as a means to reach activity locations – a conceptually superior formulation that can lead to a more refined understanding of travel behavior. Moving from an aggregate, trip-based framework to a more holistic, activity-based study of travel behavior suggests numerous challenges. Individual movement, for example, tends to follow complex trajectories that are influenced by factors in a large number of interacting dimensions, including, among others, space, time, frequency and duration of trips, the distribution of activities and social interaction. Much of this complexity must be reduced to make the problem tractable. However, before reducing these dimensions for formal analysis, Kwan (2000) argues that visualization and exploration of movement patterns can prove useful to reveal space-time interaction structures. In turn, these structures can lead to more focused analysis and therefore more realistic travel models.

Using activity diary data from Portland (Oregon), Kwan (2000) explores the complexity of travel behavior data using visualization techniques. Exploration of the spatial distribution of activities (e.g., work, home and non-work activities) and activity duration poses the considerable challenge of presenting and making intelligible tens of thousands of data points. In this case, visualization is aided by the application of kernel estimation (a univariate, moving-windows form of analysis similar in principle to GWR) to produce activity-density surfaces. The contribution of spatial analysis is to summarize and render legible a large amount of data – that is, transform data into information. The resulting surfaces represent the spatial intensities of activity distributions (the figures can be seen at http://geog-www.sbs.ohio-state.edu/faculty/mkwan/gis-t/). Moreover, their comparison is useful to glean potential relationships among them. An outcome of this exploratory analysis is a strong suggestion of spatial association between the locations of non-employment activities and workplaces. Clearly, findings such as this can pave the way to more targeted attempts to explain the spatio-temporal relationships between different types of activities and the resulting travel patterns (e.g., most non-work activities are undertaken near workplaces).

As the above example suggests, the spatial analysis of activity data, by way of describing and summarizing the data, can assist the analyst in defining, refining, and/or re-defining hypotheses. Visualization and exploration do not, however, explain travel behavior – an objective that requires a more formal modeling approach. One such a model, proposed by Bhat and Zhao (2002), incorporates the effect of spatial heterogeneity into a study of household shopping stops. The challenge faced by conventional (i.e., aspatial) approaches to modeling stop generation relates to potential instability in the relationships between the outcome (i.e., stops) and the explanatory variables. In other words, conventional approaches do not accommodate local variations in these relationships. Variation, on the other hand, is likely to occur in the analysis for reasons that include the possibility of intrinsic behavioral differences at different geographical locations, and/or incomplete information regarding relevant spatial attributes affecting decision-making behavior. In order to accommodate these concerns within their modeling framework, Bhat and Zhao (2002) adopt a multilevel approach and adapt it to produce a logit model suitable for the special needs of transportation models. In the study, multiple-stop shopping trips are analyzed as a bi-level problem by nesting individuals into traffic analysis zones (TAZs). The results of the analysis show that the spatial model gives a superior statistical fit. Also, the researchers report significant heterogeneity in the response to some factors. For instance, propensity to stop for shopping relates to the level of accessibility, but this effect may be important only when accessibility levels are low. Moreover, there is a higher variability in propensity to stop across rural and sub-urban zones when compared to urban zones. These potentially useful insights, it should be noted, would have remained uncovered by a non-spatial model that assumed away heterogeneity.

The study by Bhat and Zhao (2002), like most urban transportation studies (e.g., Bolduc *et al.*, 1995), employs a spatial framework consisting of TAZs. While the design of 'optimal' TAZs for use in aggregate models of urban travel demand (i.e., the Urban Transportation Modeling System) has received some attention in recent years (Bennion and O'Neill, 1994; You *et al.*, 1997a,b; Ding, 1998), inspired in part due to concerns about the MAUP, very few studies in urban transportation have investigated MAUP effects formally. One such study, however, is that by Horner and Murray (2002) who explore the degree to which a measure known as excess commuting is influenced by the MAUP. As argued by Scott *et al.* (1997), excess commuting is useful in an environment where sustainable transport is a desired goal of policy makers because the measure provides an objective assessment of commuting efficiency (i.e., the degree to which people live near their jobs) in an urban area. Excess commuting (E) is calculated for workers as follows:

$$E = \left(\frac{T_a - T_r}{T_a}\right) \times 100, \tag{14}$$

(14) where T_a is the average actual commute and T_r is the average required commute (i.e., a theoretical minimum that is obtained by solving for the transportation problem). Both are measured in terms of travel cost, which can be expressed in units of time or distance. Horner and Murray (2002) suggest that the MAUP manifests itself in excess commuting in two interrelated ways. First, as areal units increase in size, the likelihood that travel will be assigned outside zones decreases. This, by definition, does not reflect excess commuting. In fact, as spatial aggregation proceeds, T_r will be biased upwards such that it approaches T_a. Second, changes in zoning schemes will affect calculations of inter- and intra-zonal travel costs. Obviously, this suggests that estimates of excess commuting could be highly variable both across scales (i.e., aggregation effect) and at the same scale (i.e., zoning effect).

Horner and Murray (2002) use Boise (Idaho) for their analysis. Its 286 TAZs are systematically aggregated using the Theissen region approach to form new zoning systems. Specifically, 100 unique systems are generated for each of a series of aggregation levels ranging from 25 to 275 zones in increments of 25 zones. The excess commute for the original zoning system (i.e., 286 TAZs) is estimated at 48%. The results of the study demonstrate clearly that as aggregation proceeds, estimates of excess commuting decrease at an increasing rate. For instance, at 25 zones, the average estimate is 26%, which is far less than that for 286 zones. Furthermore, the results show that at any given aggregation level, a range of estimates exists. Moreover, the range increases as aggregation proceeds. Again, at 25 zones, the estimates of excess commuting range from 12% to 37%.

The results of Horner and Murray's (2002) study suggest that the MAUP warrants greater consideration in the analysis of transportation and travel behavior than it has received to date. This is necessary to ensure that analytical results capture the processes being scrutinized and not arbitrary effects of the zoning systems used. Researchers should also remember that the most commonly used zoning system in urban transportation, traffic analysis zones, was created with a specific purpose in mind – aggregate travel demand modeling. For this reason alone, it may not be appropriate for all types of analyses, especially those pertaining to disaggregate travel behavior. For instance, one cannot help but wonder whether the results from Bhat and Zhao's (2002) study capture the process under scrutiny or are simply artifacts of the TAZs. Obviously, this would require further investigation by the researchers.

Other examples in the transportation literature deal explicitly with spatial dependencies or heterogeneity. Wang

(2001), for instance, presents an analysis of intraurban commuting variations that uses spatially autoregressive models. Commuting studies can be and have been conducted using conventional techniques – a spatial model in this case adds value to the analysis by providing more reliable estimates. In other situations, however, there is no better or even worse alternative to the use of spatial statistics. An investigation by Steenberghen *et al.* (2004), which identified accident-prone areas in a Belgian road network, is an example of a study where spatial association in transportation data is the direct objective of the analysis. In this case, spatial analytical methods do not improve results that could be obtained by other means – they actually make the research question possible.

Spatial land-use analysis

The theoretical foundation for many problems in urban analysis originates from basic economic theory that demonstrates the fundamental relationships between the pattern of land uses, land values and commuting in a monocentric city (i.e., Alonso, 1964). This theory relates to a host of important urban topics, from housing to development to transportation infrastructure, and has therefore received considerable attention over the years. Related directly to these topics is the issue of urban organization – that is, the question of how to describe and explain the distribution of population, land values, employment and other structural variables in a city.

Numerous studies, over the years, have related population density to distance from a central business district (CBD). Since theory indicates that population density in a monocentric city should decrease with distance from the centre of the city, a function commonly used to describe this effect is the negative exponential, which is closely related to the relatively simple analytical framework required by economic theory. It has been noted, however, that the exponential function is not sufficiently sophisticated to explain the spatial variability observed in many empirical situations. Growth patterns that produce irregularities at different distances from the centre of the city (e.g., green rings, employment subcenters) and/or directional variation due to topographic factors, economically attractive locations or differential accessibility levels, represent situations that are at odds with the homogeneous space formulation implied by the simple exponential function. Some of these situations may have contributed to reduce the explanatory power of the function (e.g., Crampton, 1991).

A number of researchers have applied spatial methods to grapple with the above issues. Bender and Hwang (1985), for example, developed a switching regression that takes into account variations with distance from the CBD. They applied it to the study of land price profiles and employment subcenters. More recently, Alperovich and Deutsch (2002) proposed a more flexible switching regression specification that controls for directional variation in distance-decay. They applied their model to data from the municipality of Tel-Aviv/Yaffo (Israel) to find that the area under investigation can be divided into two separates regimes, roughly described as north and south of the CBD. Density functions for these two regimes have their own parameters, and produce results that indicate a satisfactory fit of the model to population density gradients to the north of the CBD ($\overline{R}^2 = 0.543$), but not to the south ($\overline{R}^2 = 0.000$ and an inverted but non-significant gradient). The coefficient of determination for the conventional model is $\overline{R}^2 = 0.059$. Interestingly, the thesis of separation of regimes is further supported in this case study by qualitative survey data. These data also contribute to explain the exceptional character of one tract to the north of the CBD that is classified by the model along with less attractive tracts to the south of the city.

Switching regressions, the example suggests, can successfully identify heterogeneous regimes of population density or other attributes, such as land prices, in a city. A challenging aspect of the approach, on the other hand, is how to define the location and number of switches. A number of different criteria can be applied to this end, including distance to the CBD, distance to subcenters or political boundaries. Unfortunately, these criteria are not exempt from some degree of arbitrariness. Using political boundaries implies that distance measured from a given point (e.g., the centroid of a zone or a representation of the CBD) will depend on the zoning system used (Okabe and Tagashira, 1996). Misidentification of the location of the CBD, moreover, has been shown to lead to biased population density gradients (Alperovich and Deutsch, 1992). A more fundamental question, however, is what and where is the CBD or a subcenter. Many methods currently in use define a subcenter as a site with significantly larger employment density than neighboring locations that has an impact on overall employment density. As discussed by McMillen (2001), these methods depend on arbitrary definitions of size (how large is large?), and often require substantial knowledge of the study area to become operational. Sometimes the definitions must be tweaked in order to give reasonable results (see, for example, McMillen and McDonald, 1998).

It is interesting to note, against this backdrop, that the use of recent developments in spatial statistics can help to overcome these limitations – as recent studies that apply local forms of spatial analysis in an urban context show. Baumont *et al.* (2004), for example, use LISA statistics to study population and employment variations in a French city, whereas a study by Páez *et al.* (2001) used a statistic of the $G_i(d)$ family to explore land price variation in a Japanese city. The results from these studies suggest that a number of population and employment subcenters exist in the French city, while the Japanese city appears to be strongly monocentric in terms of its land-price structure. Exploratory spatial data analysis in both papers, moreover, guides model development in frameworks that use spatially autoregressive models (Baumont *et al.*, 2004) and spatially switching-spatially autoregressive models (Páez *et al.*, 2001). The findings reported confirm the importance of identifying the CBD and possibly other subcenters. Ordinary models that ignore the spatial characteristics of the process, it is shown, tend to under-perform the spatial models, and also may mask other interesting results. The spatial model used in the Japanese analysis, for example, suggests that land prices are significantly related to distance to two emerging subcenters, but distance to the CBD is only significant in the central part of the city. The emerging subcenters, intriguingly, do not have a significant presence in the overall land-price landscape in the city.

A different approach to identify subcenters that uses geographically weighted regression (GWR) is developed by McMillen (2001). A feature of GWR that is important within this context is its role as a smoother, meaning that it can be used to approximate to a very high level of accuracy the observed surface (e.g., population density, land values) By adopting a predetermined window size to give a desired level of smoothing, McMillen (2001) shows how GWR can be used in a two-stage procedure that involves the use of semi-parametric methods in the second stage to identify urban subcenters in polycentric cities. Use of this method allowed him to identify subcenters in a large number of American cities, including Chicago, Dallas and Los Angeles, with little or no knowledge of the local situation as a pre-requisite to the procedure.

The above examples illustrate that local methods of spatial analysis can lead to more efficient analysis of urban data in the context of studying the spatial structure of metropolitan areas. In addition to these, other examples exist that apply spatial statistical methods to urban land-use analysis. Spatially autoregressive models, for example, have been used to analyze housing prices and neighborhood effects (Tse, 2002) and the effect of transportation infrastructure on housing prices (Haider and Miller, 2000). The concept of spatial association is also intimately related to the analysis of land-use change and development, a topic where an underlying question is the extent to which the process is conditioned by space. Páez and Suzuki (2001) applied a dynamic spatial logit model to investigate the effect of neighborhood effects on land-use change. Their findings, using two case studies, suggest that residential or commercial construction tends to encourage the same type of development in neighboring zones. On the other hand, distance to transportation facilities does not seem to influence land-use change. Cuthbert and Anderson (2002a,b) also addressed the question of urban form and land development, but using individual-level (i.e., disaggregate) parcel data from Halifax-Dartmouth (Canada). Kernel estimations and nearest neighbor analysis (i.e., tools to study spatial association in point patterns) afford them insights into an evolving urban form possibly undergoing a multinucleation process.

Other studies have also turned their attention to disaggregate analysis, but from the perspective of locational decisions by individual households and firms. Along this line of inquiry, researchers have become interested in the possible effect of unexplained spatial variability on individual decisions (Bhat and Guo, 2004), the interaction between spatial alternatives and/or decision makers (Mohammadian and Kanaroglou, 2003) and a combination of both effects (Miyamoto *et al.*, 2004). Also, the MAUP has received attention in the context of modeling the residential location choices of households (Guo and Bhat, 2004). Results from these studies suggest that incorporating spatial analytical issues in disaggregate analysis is important not only on technical grounds, but also for the realistic assessment of locational and other transportation-related decisions as a function of transportation systems, land use and socio-economic variables.

Summary and conclusions

This paper has reviewed a number of developments in spatial statistics and, by means of recent examples from the scientific literature, has shown how these developments can improve urban modeling practice and potentially also our understanding of urban processes. A limitation to the application of spatial methods was, for a long time, the availability, or lack thereof, of adequate software and training materials, textbooks and data. This limitation has largely disappeared. Currently, the urban analyst can count on the advantages of using GIS to manage, process, and visualize data, combined with the ability of spatially analyzing data by means of specialized software. Existing packages include commercial software (e.g., SPACESTAT by BioMedware Inc. and HLM by Raudenbush *et al.* [2001]) and those available from researchers (e.g., the SAGE project [Wise *et al.*, 2001] and James P. LeSage's spatial econometrics MATLAB library [www.spatial-econometrics.com]). In addition to software, a wealth of spatial data has become available (e.g., the Bureau of Transportation Statistics of the United States Department of Transportation provides georeferenced data on its website [http://www.transtats.bts.gov/]), and a large number of books now exist that present spatial statistical methods at different levels of technical expertise, from introductory to advanced (e.g., Odland, 1987; Anselin, 1988; Griffith, 1988; Haining, 1990; Cressie, 1993; Bailey and Gatrell, 1995; Griffith and Layne, 1999).

The examples in this paper represent an incipient trend in urban analysis towards the use of increasingly sophisticated spatial statistical tools. The influence, moreover, is reciprocal, as there are questions of interest in urban analysis that help motivate research in spatial analysis. Overall, there seems to be considerable potential for cross-fertilization between the two fields. In the future, it seems likely that development and application of spatial statistical methods for urban analysis will tend to draw more heavily on various sources of theory. To date, and despite calls for theory-driven model development (e.g., Anselin, 1988), at times there appears to be an excessive focus on the application of the methods for the application's sake. Recent research in various fields, on the other hand, now provides rich sources of theory to inspire the derivation of spatially explicit models. Sources of theory include the 'new economic geography' (NEG) of regional science (Fujita *et al.*, 1999), a field concerned with the spatial distribution of economic activity; the economic theory of externalities, public goods and club goods (Cornes and Sandler, 1996); and the sociological concept of social networks, which provides the framework for the analysis of social interaction in individual and group behavior (Wasserman *et al.*, 1994). Efforts to bridge the gap between highly abstract NEG theory and other approaches to economic geography, for example, have been greatly facilitated by the use of spatial statistical models (Fingleton, 2000). Griffith's (1999) discussion of the sources of theory in regional science, moreover, intersects some problems of interest for urban and spatial analysis, such as the location of firms and the estimation of urban population density functions.

On a related note, it appears likely that the greater emphasis on behaviorally-based explanations for urban processes (Kanaroglou and Scott, 2002) will require more extensive use of discrete choice models with spatial effects, including interactions, inter-related decision making and population heterogeneities. The basic technical framework for this type of research has been laid out in the form of the mixed logit model with its highly flexible covariance structure. This model has already found application in the spatial analysis of urban processes (see Mohammadian and Kanaroglou, 2003; Bhat and Guo, 2004; Miyamoto *et al.*, 2004). Future research should lead to a better integration of this technical basis with concepts derived from the theories of economic externalities and social networks. Such research holds the potential to produce analytical tools capable of addressing richer and more nuanced urban research questions.

Finally, it must be noted that although closely related to the topic of urban analysis, space limitations have prevented this review from covering other important topics such as the analysis of systems of cities and urban environmental systems. Interesting developments in these fields include work by Ramos and Silva (2003) that used local statistics of spatial association to delineate metropolitan areas; the analysis of spatial spillovers in the provision of transportation infrastructure (Rietveld and Wintershoven, 1998); research by Pacheco and Tyrrell (2002) that explored the existence of clusters of households in a system of cities; the analysis, using spatially autoregressive models, of network accessibility and the distribution of economic activity in Eastern Asia (Páez, 2004); and the use of spatial interpolation techniques to derive urban pollution maps (Kanaroglou *et al.*, 2002). We anticipate that future reviewers of spatial and urban analysis will be hard pressed to cover in a concise fashion the exploding amounts of research that will likely continue to appear in the literature.

Acknowledgements

An earlier version of this paper was presented in the 9th International Conference on Computers in Urban Planning and Urban Management in Sendai, Japan (May, 2003). The comments of the participants in the Spatial Analysis session are gratefully acknowledged. Two anonymous reviewers made useful suggestions that helped to improve the quality of the paper. This research was partially supported by grants 261872-03 (NSERC), 0136818 (NSF) and DEC#SCO (GEOIDE).

References

Abed J. and Kaysi I., 2003: Identifying urban boundaries: application of remote sensing and geographic information system technologies. *Canadian Journal of Civil Engineering* **30**: 992—999.

Alonso W., 1964: *Location and Land Use*. Harvard University Press, Cambridge.

Alperovich G. and Deutsch J., 1992: Population-density gradients and urbanization measurement. *Urban Studies* **29**: 1323–1328.

Alperovich G. and Deutsch J., 2002: An application of a switching regimes regression to the study of urban structure. *Papers in Regional Science* **81**: 83–98.

Amrhein C.G., 1995: Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environment and Planning A* **27**: 105–119.

Amrhein C.G. and Flowerdew R., 1992: The effect of data aggregation on a Poisson regression-model of Canadian migration. *Environment and Planning A* **24**: 1381–1391.

Anselin L. and Griffith D.A., 1988: Do spatial effects really matter in regression-analysis. *Papers of the Regional Science Association* **65**: 11–34.

Anselin L., 1988: *Spatial Econometrics: Methods and Models*. Kluwer, Dordrecht.

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

Bailey T.C. and Gatrell A.C., 1995: *Interactive Spatial Data Analysis*. Addison Wesley Longman, Essex.

Barredo J.I., Demicheli L., Lavalle C., Kasanko M. and McCormick N., 2004: Modelling future urban scenarios in developing countries: an application case study in Lagos, Nigeria. *Environment and Planning B* **31**: 65–84.

Batty M. and Sikdar P.K., 1982a: Spatial aggregation in gravity models. 1. an information-theoretic framework. *Environment and Planning A* **14**: 377–405.

Batty M. and Sikdar P.K., 1982b: Spatial aggregation in gravity models. 2. one-dimensional population-density models. *Environment and Planning A* **14**: 525–553.

Batty M. and Sikdar P.K., 1982c: Spatial aggregation in gravity models. 3. two-dimensional trip distribution and location models. *Environment and Planning A* **14**: 629–658.

Batty M. and Sikdar P.K., 1982d: Spatial aggregation in gravity models. 4. generalizations and large-scale applications. *Environment and Planning A* **14**: 795–822.

Baumont C., Ertur C. and Le Gallo J., 2004: Spatial analysis of employment and population density: The case of the aglomeration of Dijon 1999. *Geographical Analysis* **36**: 146–176.

Bender B. and Hwang H., 1985: Hedonic housing price indices and secondary employment centers. *Journal of Urban Economics* **17**: 90–107.

Bennion M.W. and O'Neill W.A., 1994: Building transportation analysis zones using geographic information systems. *Transportation Research Record* **1429**: 49–56.

Bhat C. and Zhao H.M., 2002: The spatial analysis of activity stop generation. *Transportation Research B* **36**: 557–575.

Bhat C.R. and Guo J., 2004: A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Transportation Research B* **38**: 147–168.

Black J.A., 1981: *Urban Transport Planning*. Croom Helm, London.

Blalock H., 1964: *Causal Inferences on Nonexperimental Research*. University of North Carolina Press, Chapel Hill.

Bolduc D., Dagenais M.G. and Gaudry M.J.I., 1989: Spatially autocorrelated errors in origin-destination models: A new specification applied to aggregate mode choice. *Transportation Research B* **23**: 361–372.

Bolduc D., Laferriere R. and Santarossa G., 1992: Spatial autoregressive error-components in travel flow models. *Regional Science and Urban Economics* **22**: 371–385.

Bolduc D., Laferriere R. and Santarossa G., 1995: Spatial autoregressive error components in travel flow models: an application to aggregate mode choice. In Anselin, L. and Florax, R. J.G.M. (eds), New Directions in Spatial Econometrics, pp. 96–108. Springer-Verlag, Berlin.

Briggs D.J., Collins S., Elliott P., Fischer P., Kingham S., Lebret E., Pryl K., VanReeuwijk H., Smallbone K. and VanderVeen A., 1997: Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science* **11**: 699–718.

Brunsdon C., Fotheringham A.S. and Charlton M., 1999: Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science* **39**: 497–524.

Brunsdon C., Fotheringham A.S. and Charlton M.E., 1996: Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* **28**: 281–298.

Ceccato V., Haining R. and Signoretta P., 2002: Exploring offence statistics in Stockholm City using spatial analysis tools. *Annals of the Association of American Geographers* **92**: 29–51.

Cleveland W.S., 1979: Robust locally weighted regression and smoothing of scatterplots. *Journal of the American Statistical Association* **74**: 823–836.

Cliff A.D. and Ord J.K., 1981: *Spatial Processes: Models and Applications*. Pion, London.

Cornes R. and Sandler T., 1996: *The Theory of Externalities, Public Goods, and Club Goods*. 2nd Edition: Cambridge University Press, Cambridge.

Craglia M., Haining R. and Signoretta P., 2001: Modelling high-intensity crime areas in English cities. *Urban Studies* **38**: 1921–1941.

Craglia M., Haining R. and Wiles P., 2000: A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies* **37**: 711–729.

Crampton G.R., 1991: Residential density patterns in London: any role left for the exponential density gradient. *Environment and Planning A* **23**: 1007–1024.

Cressie N.A.C., 1993: *Statistics for Spatial Data*. John Wiley & Sons, New York.

Cuthbert A.L. and Anderson W.P., 2002a: An examination of urban form in Halifax-Dartmouth: alternative approaches in data. *Canadian Journal of Urban Research* **11**: 213–237.

Cuthbert A.L. and Anderson W.P., 2002b: Using spatial statistics to examine the pattern of urban land development in Halifax-Dartmouth. *Professional Geographer* **54**: 521–532.

De la Barra T., 1989: *Integrated Land Use and Transport Modelling*. Cambridge University Press, Cambridge.

Ding C., 1998: The GIS-based human-interactive TAZ design algorithm: examining the impacts of data aggregation on transportation-planning analysis. *Environment and Planning B* **25**: 601–616.

Du G., 2001: Using GIS for analysis of urban systems. *GeoJournal* **52**: 213–221.

Duncan C. and Jones K., 2000: Using multilevel models to model heterogeneity: potential and pitfalls. *Geographical Analysis* **32**: 279–305.

Fingleton B., 2000: Spatial econometrics, economic geography, dynamics and equilibrium: a 'third way'? *Environment and Planning A* **32**: 1481–1498.

Flint C., 2002: The theoretical and methodological utility of space and spatial statistics for historical studies: the Nazi Party in geographic context. *Historical Methods* **35**: 32–42.

Fotheringham A.S., 1989: Scale-independent spatial analysis. In Goodchild, M. and Gopal, S. (eds), Accuracy of Spatial Databases, pp. 221–228. Taylor and Francis, London.

Fotheringham A.S., 2000: GIS-based spatial modelling: a step forwards or a step backwards. In Fotheringham, A.S. and Wegener, M. (eds), Spatial Models and GIS: New Potential and New Models, pp. 21–30. Taylor and Francis, London.

Fotheringham A.S. and Brunsdon C., 1999: Local forms of spatial analysis. *Geographical Analysis* **31**: 340–358.

Fotheringham A.S., Brunsdon C. and Charlton M., 2000: *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage, London.

Fotheringham A.S., Densham P.J. and Curtis A., 1995: The zone definition problem in location-allocation modeling. *Geographical Analysis* **27**: 60–77.

Fotheringham A.S. and Wong D.W.S., 1991: The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* **23**: 1025–1044.

Fujita M., Krugman P. and Venables A.J., 1999: *The Spatial Economy: Cities, Regions and International Trade*. MIT Press, Cambridge.

Gehlke C.E. and Biehl K., 1934: Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association* **29**: 169–170.

Getis A. and Ord J.K., 1993: The analysis of spatial association by use of distance statistics. *Geographical Analysis* 25: 276–276.

Goodchild M.F., 1979: Aggregation problem in location-allocation. *Geographical Analysis* **11**: 240–255.

Griffith D.A., 1988: *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*. Kluwer, Dordrecht.

Griffith D.A., 1999: Statistical and mathematical sources of regional science theory: map pattern analysis as an example. *Papers in Regional Science* **78**: 21–45.

Griffith D.A. and Layne L.J., 1999: *A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Data Sets*. Oxford University Press, New York.

Guo J.Y. and Bhat C.R., 2004: Modifiable areal units: a problem or a matter of perception in the context of residential location choice modeling? *Transportation Research Record*, forthcoming.

Haider M. and Miller E.J., 2000: Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. *Transportation Research Record* **1722**: 1–8.

Haining R., 1990: *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge.

Hall P., 1986: The state of cities and the state of planning. In Hutchinson, B.G. and Batty, M. (eds), Advances in Urban Systems Modelling, pp. 3–13. Elsevier Science, Amsterdam.

Harris B., 1985: Urban simulation-models in regional science. *Journal of Regional Science* **25**: 545–567.

Harris B. and Batty M., 1993: Locational models, geographic information and planning support systems. *Journal of Planning Education and Research* **12**: 184–198.

Hodgson M.J., Shmulevitz F. and Körkel M., 1997: Aggregation error effects on the discrete-space *p*-median model: the case of Edmonton, Canada. *Canadian Geographer* **41**: 415.

Holt D., Steel D.G., Tranmer M. and Wrigley N., 1996: Aggregation and ecological effects in geographically based data. *Geographical Analysis* **28**: 244–261.

Horner M.W. and Murray A.T., 2002: Excess commuting and the modifiable areal unit problem. *Urban Studies* **39**: 131–139.

Innes J.E. and Simpson D.M., 1993: Implementing GIS for planning: lessons from the history of technological innovation. *Journal of the American Planning Association* **59**: 230–236.

Jerrett M., Burnett R.T., Goldberg M.S., Sears M., Krewski D., Catalan R., Kanaroglou P., Giovis C. and Finkelstein N., 2003: Spatial analysis for environmental health research: concepts, methods, and examples. *Journal of Toxicology and Environmental Health A* **66**: 1783–1810.

Jones K., 1991: Specifying and estimating multilevel models for geographical research. *Transactions of the Institute of British Geographers* **16**: 148–159.

Jones K. and Bullen N., 1994: Contextual models of urban house prices: a comparison of fixed-coefficient and random-coefficient models developed by expansion. *Economic Geography* **70**: 252–272.

Kanaroglou P.S. and Scott D.M., 2002: Integrated urban transportation and land-use models. In Dijst, M., Schenkel, W. and Thomas, I. (eds), Governing Cities on the Move: Functional and Management Perspectives on Transformations of European Urban Infrastructures, pp. 42–72. Ashgate, Aldershot.

Kanaroglou P.S., Soulakellis N.A. and Sifakis N.I., 2002: Improvement of satellite derived pollutions maps with the use of a geostatistical interpolation method. *Journal of Geographical Systems* **4**: 193–208.

King G., 1997: *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton.

Krieger M.H., 1991: Segmentation and filtering into neighborhoods as processes of percolation and diffusion: stochastic-processes (randomness) as the null hypothesis. *Environment and Planning A* **23**: 1609–1626.

Kwan M.P., 2000: Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research C* **8**: 185–203.

Landis J. and Zhang M., 1998a: The second generation of the California urban futures model. part 1: model logic and theory. *Environment and Planning B* **25**: 657–666.

Landis J. and Zhang M., 1998b: The second generation of the California urban futures model. part 2: specification and calibration results of the land-use change submodel. *Environment and Planning B* **25**: 795–824.

Landis J. and Zhang M., 2000: Using GIS to improve urban activity and forecasting models: three examples. In Fotheringham, A.S. and Wegener, M. (eds), Spatial Models and GIS: New Potential and New Models, pp. 63–81. Taylor and Francis, London.

Levine J. and Landis J., 1989: Geographic information systems for local planning. *Journal of the American Planning Association* **55**: 209–220.

Liu S.X. and Zhu X.A., 2004: Accessibility analyst: an integrated GIS tool for accessibility analysis in urban transportation planning. *Environment and Planning B* **31**: 105–124.

McMillen D.P., 1996: One hundred fifty years of land values in Chicago: a nonparametric approach. *Journal of Urban Economics* **40**: 100–124.

McMillen D.P., 2001: Nonparametric employment subcenter identification. *Journal of Urban Economics* **50**: 448–473.

McMillen D.P. and McDonald J.F., 1998: Population density in suburban Chicago: a bid-rent approach. *Urban Studies* **35**: 1119–1130.

Meyer M.D. and Miller E.J., 2001: *Urban Transportation Planning*. 2nd Edition: McGraw-Hill, New York.

Miller H.J., 1999: Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T). *Geographical Analysis* **31**: 373–399.

Miller H.J. and Shaw S.L., 2001: *Geographic Information Systems for Transportation: Principles and Applications*. Oxford University Press, Oxford.

Miyamoto K., Vichiensan V., Shimomura N. and Páez A., 2004: Discrete choice model with structuralized spatial effects for location analysis. *Transportation Research Record*, forthcoming.

Moeckel R., Spiekermann K., Schürmann C. and Wegener M., 2003: Microsimulation of urban land use. Paper presented at the 9th Conference on Computers in Urban Planning and Urban Management, Sendai, Japan, 27–29 May.

Mohammadian A. and Kanaroglou P.S., 2003: Applications of spatial multinomial logit model to transportation planning. Paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, Switzerland, 10–14 August.

Murdoch J.C., Sandler T. and Sargent K., 1997: A tale of two collectives: sulphur versus nitrogen oxides emission reduction in Europe. *Economica* **64**: 281–301.

Murray A. and Gottsegen J., 1997: The influence of data aggregation on the stability of p-median location model solutions. *Geographical Analysis* **29**: 200–213.

Odland J., 1987: *Spatial Autocorrelation*. Sage Publications, Newbury Park.

Okabe A. and Tagashira N., 1996: Spatial aggregation bias in a regression model containing a distance variable. *Geographical Systems* **3**: 77–99.

Okunuki K., 2001: Urban analysis with GIS. *GeoJournal* **52**: 181–188.

Openshaw S., 1977a: Geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers* **2**: 459–472.

Openshaw S., 1977b: Optimal zoning systems for spatial interaction models. *Environment and Planning A* **9**: 169–184.

Openshaw S., 1978a: An optimal zoning approach to the study of spatially aggregated data. In Masser, I. and Brown, P.J.B. (eds), Spatial Representation and Spatial Interaction, pp. 93–113. Martinus Nijhoff, Leiden.

Openshaw S., 1978b: Empirical-study of some zone-design criteria. *Environment and Planning A* **10**: 781–794.

Openshaw S. and Baxter R.S., 1977: Algorithm 3 – procedure to generate pseudo-random aggregations of n-zones into m-zones, where m is less than n. *Environment and Planning A* **9**: 1423–1428.

Openshaw S. and Rao L., 1995: Algorithms for reengineering 1991 census geography. *Environment and Planning A* **27**: 425–446.

Openshaw S. and Taylor P.J., 1979: A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In Wrigley, N. (ed), *Statistical Applications in the Spatial Sciences*, pp. 127–144. Pion, London.

Ord J.K. and Getis A., 1995: Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* **27**: 286–306.

Ord J.K. and Getis A., 2001: Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science* **41**: 411–432.

Ortúzar J.D. and Willumsen L.G., 2001: *Modelling Transport*. 3rd Edition: Wiley, New York.

Pacheco A.I. and Tyrrell T.J., 2002: Testing spatial patterns and growth spillover effects in clusters of cities. *Journal of Geographical Systems* **4**: 275–285.

Pacione M., 1990: *Urban Problems: An Applied Urban Analysis*. Routledge, London.

Páez A., 2004: Network accessibility and the distribution of economic activity in Eastern Asia. *Urban Studies* (forthcoming).

Páez A. and Suzuki J., 2001: Transportation impacts on land use change: an assessment considering neighborhood effects. *Journal of the Eastern Asia Society for Transportation Studies* **4**: 47–59.

Páez A., Uchida T. and Miyamoto K., 2001: Spatial association and heterogeneity issues in land price models. *Urban Studies* **38**: 1493–1508.

Peeters D. and Ruhigira J.D., 2004: Using the transportation problem to measure the proximity between two sets of weighted points in geographical space. *Papers in Regional Science* **72**: 337–347.

Putman S.H. and Chung S.H., 1989: Effects of spatial system-design on spatial interaction models.1. the spatial system definition problem. *Environment and Planning A* **21**: 27–46.

Quandt R., 1958: The estimation of parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* **53**: 873–880.

Ramos R.A.R. and Silva A.N.R., 2003: A data-driven approach for the definition of metropolitan regions. Paper presented at the 9th Conference on Computers in Urban Planning and Urban Management, Sendai, Japan, 27–29 May.

Raudenbush S.W., Bryk A.S., Cheong Y.F. and Congdon R., 2001: *HLM 5: Hierarchical Linear and Nonlinear Modeling*. 2nd Edition: Scientific Software International, Illinois.

Rietveld P. and Wintershoven P., 1998: Border effects and spatial autocorrelation in the supply of network infrastructure. *Papers in Regional Science* **77**: 265–276.

Robinson T.P., 2000: Spatial statistics and geographical information systems in epidemiology and public health. *Advances in Parasitology* **47**: 81–128.

Scott D.M., Kanaroglou P.S. and Anderson W.P., 1997: Impacts of commuting efficiency on congestion and emissions: case of the Hamilton CMA, Canada. *Transportation Research D* **2**: 245–257.

Spiekermann K. and Wegener M., 2000: Freedom from the tyranny of zones: towards new GIS-based spatial models. In Fotheringham, A.S. and Wegener, M. (eds), pp. 45–61. Taylor and Francis, London.

Stanilov K., 2003: Accessibility and land use: the case of suburban Seattle, 1960–1990. *Regional Studies* **37**: 783–794.

Steel D.G. and Holt D., 1996: Rules for random aggregation. *Environment and Planning A* **28**: 957–978.

Steenberghen T., Dufays T., Thomas I. and Flahaut B., 2004: Intra-urban location and clustering of road accidents using GIS: a Belgian example. *International Journal of Geographical Information Science* **18**: 169–181.

Tagashira N. and Okabe A., 2002: The modifiable areal unit problem in a regression model whose independent variable is a distance from a predetermined point. *Geographical Analysis* **34**: 1–20.

Thill J.C., 2000: Geographic information systems for transportation in perspective. *Transportation Research C* **8**: 3–12.

Tobler W.R., 1989: Frame independent spatial analysis. In Goodchild, M. and Gopal, S. (eds), Accuracy of Spatial Databases, pp. 115–122. Taylor and Francis, London.

Tranmer M. and Steel D.G., 1998: Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A* **30**: 817–831.

Tse R.Y.C., 2002: Estimating neighbourhood effects in house prices: towards a new hedonic model approach. *Urban Studies* **39**: 1165–1180.

Vichiensan V., Miyamoto K. and Shizukuishi K., 2001: GIS test bed building for land-use and transportation models. Paper presented at the 7th International Conference on Computers in Urban Planning and Urban Management, Hawaii, 14–16 July.

Wang D.G. and Cheng T., 2001: A spatio-temporal data model for activity-based transport demand modelling. *International Journal of Geographical Information Science* **15**: 561–585.

Wang F.H., 2001: Explaining intraurban variations of commuting by job proximity and workers' characteristics. *Environment and Planning B* **28**: 169–182.

Wasserman S., Faust K. and Iacobucci D., 1994: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Wegener M., Gnade F. and Vannahme M., 1986: The time scale of urban change. In Hutchinson, B.G. and Batty, M. (eds), Advances in Urban Systems Modelling, pp. 175–197. Elsevier Science, Amsterdam.

Whittle P., 1954: On stationary processes in a plane. *Biometrika* **41**: 434–439.

Wise S., Haining R. and Ma J.S., 2001: Providing spatial statistical data analysis functionality for the GIS user: the SAGE project. *International Journal of Geographical Information Science* **15**: 239–254.

Wong D.W.S., 1996: Aggregation effects in geo-referenced data. In: Arlinghaus S.L., Griffith D.A., Drake W.D. and Nystuen J.D. (eds), *Practical Handbook of Spatial Statistics*, pp. 83–106. CRC Press, Boca Raton.

Wong D.W.S., Lasus H. and Falk R.F., 1999: Exploring the variability of segregation index D with scale and zonal systems: an analysis of thirty US cities. *Environment and Planning A* **31**: 507–522.

You J., Nedoviæ-Budiæ Z. and Kim T.J., 1997a: A GIS-based traffic analysis zone design: technique. *Transportation Planning and Technology* **21**: 45–68.

You J.S., Nedoviæ-Budiæ Z. and Kim T.J., 1997b: A GIS-based traffic analysis zone design: implementation and evaluation. *Transportation Planning and Technology* **21**: 69–91.