Chapter 8
Spatial Models

8.1 Introduction

Economists and statisticians are rediscovering geography. Until relatively recently, most economic models essentially ignored spatial variations in data and in relationships; these were not at the heart of the issues that were considered to be interesting.

In practice, geography is important. Programs aimed at tackling poverty may find it useful to target their interventions by region or district; this requires *poverty mapping*, which we discuss more fully in Chap. 13. The patterns of demand, or household behavior, may vary from area to area, in which case policy measures may need to be calibrated to take this into account. The simple measurement of poverty will itself depend heavily on the accurate estimation of price differences across time and space.

More generally, our understanding of economic processes often requires an appreciation of the spatial dimension. For instance, why do unemployment rates, or wages, vary systematically from one district to another? What role do spatial differences play in explanations of agricultural productivity? What are the determinants of deforestation, and can we model the likelihood that a given area of forest will be cut down in the foreseeable future?

In this chapter, we explain why spatial information needs to be incorporated into many models, and how this may be done. Only a handful of studies in less-developed countries have used these techniques with household survey data, but the number is growing rapidly, and we refer to this literature at the appropriate points in the chapter.

The chapter is constructed around an example based on a study of the determinants of unemployment rates in the communities of the Midi-Pyrenées region in France, which serves to illustrate the essential methods and conclusions of spatial econometrics.

Every spatial analysis has to begin with a map, so a good place to start is with Fig. 8.1, which graphs the residuals from an ordinary least squares regression of the unemployment rate on a number of explanatory variables, for hinterlands in the

D. Haughton and J. Haughton, *Living Standards Analytics*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-0385-2_8,

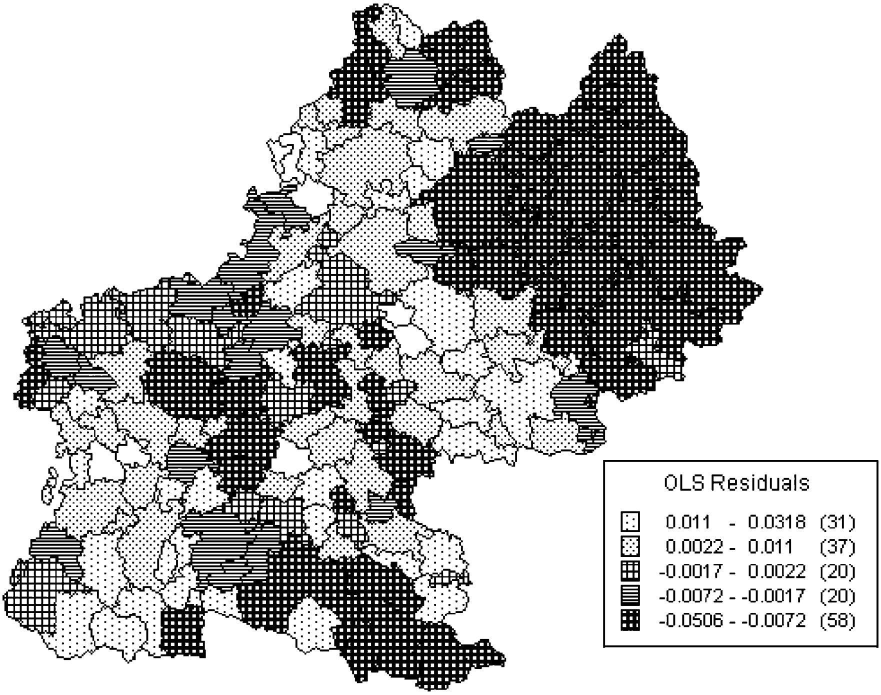

Fig. 8.1 Residuals from OLS regression of the determinants of the unemployment rate for hinterlands in the Midi-Pyrénées region of France, 1990. (*Source*: Aragon et al. 2003)

Midi-Pyrénées region of France in 1990. The interesting point is that the residuals show some spatial clumping – with negative values in the northeast, and positive values in the southwest. In other words, the unemployment rate in a given area appears to be correlated with the unemployment rate in neighboring areas, even after controlling for observable variables. Unfortunately, the use of OLS in this context will generate estimates that are inefficient and, depending on the form of the spatial links, may be biased and inconsistent. Much of this chapter is devoted to dealing with this problem.

8.2 The Starting Point: Including Spatial Variables

8.2.1 Exploratory Spatial Data Analysis

Even before introducing statistical techniques, it is helpful to examine a few basic maps in some detail, to get a “feel” for the spatial structure of the data. One useful tool is the GeoXp package developed at Université Toulouse I (Thomas-Agnan, Aragon, Ruiz-Gazen, Laurent, and Robidou, 2006) and implemented in R.

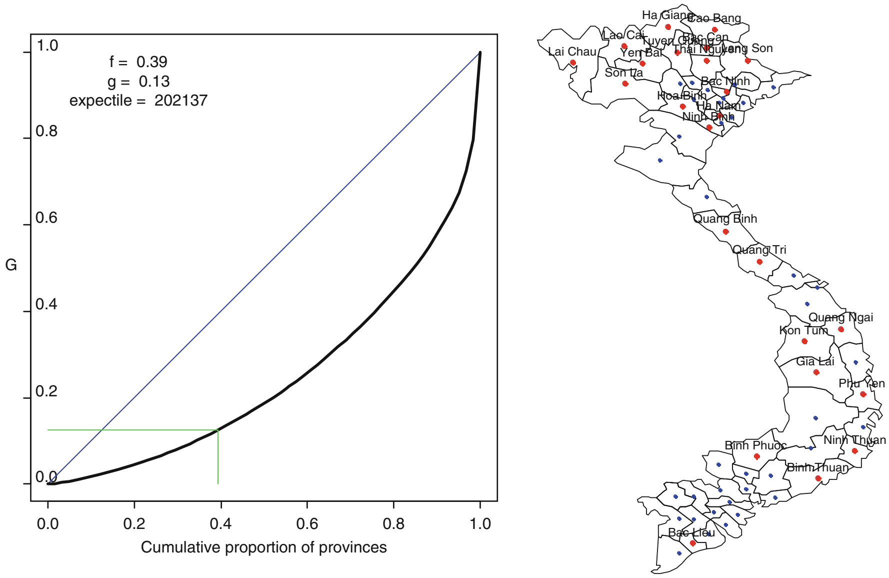

Fig. 8.2 Lorenz curve and map of GDP/capita for Vietnamese provinces, 1998

To illustrate how GeoXp works, consider Fig. 8.2, which shows on the right-hand side a map of the provinces of Vietnam, and on the left-hand side a Lorenz curve that graphs the cumulative proportion of the provinces (from poorest to richest) on the horizontal axis and the cumulative proportion of GDP/capita on the vertical axis. This is thus a measure of inequality in per capita GDP between the provinces. As the user moves a cursor along the Lorenz curve, more and more provinces are named and highlighted on the map. The snapshot in Fig. 8.2 shows the poorest two-fifths of the provinces in Vietnam, as of 1998.

Another way to use GeoXp is to create a map and an associated histogram, as done in Fig. 8.3. In this case, the variable in question is the human development index (HDI), which varies between 0 and 1 and is a weighted average of normalized measure of life expectancy at birth (1/3 weight), adult literacy (2/9 weight), gross school enrollment rates (1/9 weight), and GDP per capita (1/3 weight). By moving a cursor along the horizontal axis, it is possible to select the poorest seven provinces, which are then highlighted on the accompanying map. As one moves the cursor, the map adjusts immediately, allowing one to form a clear impression of the spatial effects.

Other commands in GeoXp help one to visualize which areas – provinces in the Vietnamese context – are most closely correlated with their neighbors.

8.2.2 Including Spatial Variables

One of the most straightforward ways to handle geographic effects is to include spatial variables directly as explanatory variables in a model. Consider, for example,

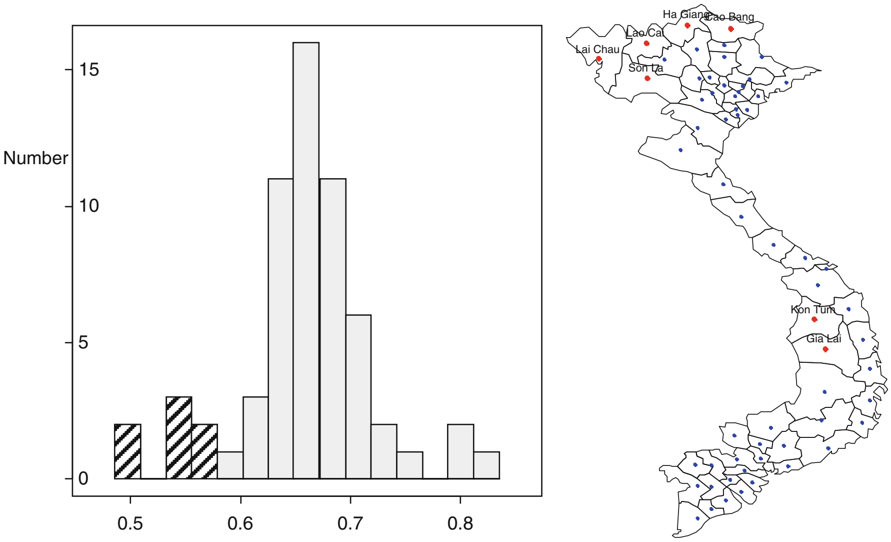

Fig. 8.3 Histomap of human development index, Vietnamese provinces, 1998

the recent study by Kaimowitz et al. (2002) of the determinants of deforestation in the province of Santa Cruz, a well-watered and relatively heavily forested area located on the eastern side of Bolivia. Based on satellite data, the researchers divided the province (other than protected areas and indigenous territories) into 24,208 relatively homogeneous polygons, representing areas that were forested in 1989. The dependent variable ("forest") is set equal to 1 if an area was still forested in 1994, and to zero otherwise. The goal is thus to identify the variables that were associated with deforestation during the 5 years up to 1994.

Forest is more likely to be felled or burned if it is accessible – closer to towns and roads – and if it is unprotected. Kaimowitz et al. also hypothesize that the most vulnerable areas are forest that are not those in areas of high rainfall (too wet) or low rainfall (not enough wood), but areas in between. They fit a logistic model, where each observation is weighted by the area of the polygon. The results of one of their more parsimonious models, including a variety of geographic variables, are shown in Table 8.1. The estimates show that areas designated for settlement, and close to the regional capital (Santa Cruz) and to trails, were more likely to lose forest cover. Curiously, areas located closer to roads were more likely to remain forested. Moreover, whether an area was designated as protected or indigenous was unrelated to deforestation (results not shown here).

Although regression estimates that include spatial variables are common enough, they typically suffer from the problem that the observations may not be independent of one another. A hilly district is likely to be close to another hilly district. Or again, shocks such as floods that hit one area are likely to affect nearby areas. These are but two examples of spatial correlation, but when they occur we need to use some additional statistical tricks.

Table 8.1 Logistic regression of the correlates of deforestation in Santa Cruz province, Bolivia, 1989–1994

	Coefficient	*t*-value
Dependent variable: =1 if area forested in 1989 and 1994, =0 if forested in 1989 but not in 1994		
Independent variables		
Intercept	0.833	3.72
Area is a forest concession (yes = 1)	1.771	5.99
Area is zoned for colonization (yes = 1)	−0.600	4.23
Soil quality (USDA soil group, from 1 through 8)	0.150	4.53
Rainfall (in millimeters p.a.)	−0.001	4.61
Distance to nearest classified road (km)	−0.051	9.61
Distance to nearest trail outside a forest concession (km)	0.008	6.04
Distance to nearest trail inside a forest concession (km)	0.003	2.21
Distance to Santa Cruz (km)	0.008	12.33

Source: Kaimowitz et al. (2002)
Note: Pseudo $R^2 = 0.23$. Number of observations: 24,208

8.3 Spatial Models

Following LeSage (1998, p. 2), we may distinguish two problems that arise when working with data that have a locational component: spatial dependence, and spatial heterogeneity. When either is present, the method of ordinary least squares is inappropriate (as we explain below), and alternative estimation methods are required.

8.3.1 Spatial Dependence

Spatial dependence occurs when the value of an observation in location i depends on observations in locations $j \neq i$. For instance, if unemployment is high in location i, then it is also likely to be high in the neighboring area j. Formally, we have

$$y_i = f(y_j), \quad i = 1, \ldots, n, \quad j \neq i, \tag{8.1}$$

where y_i is the variable of interest, such as the unemployment rate.

Such dependence may occur because it is inherently important to the problem at hand, and reflects the fundamental theorem of regional science, which states that "distance matters." Thus the price of a house may be related to the neighboring house, perhaps because they were built at the same era, or because they are equally close to the beach, or because one well-kept house begets another. To model this spatial dependence, we need to determine an appropriate form for the function $f(\cdot)$.

Spatial dependence can also arise as a consequence of measurement error, particularly if the spatial units (e.g., zip-code area, census tract, state) are not congruent

Table 8.2 Illustrative example of spatial dependence due to measurement error

District	A	B	C
Labor force	2,000	2,000	2,000
Employment in district	1,350	1,800	1,350
True unemployment in district	500	500	500
True unemployment rate in district (%)	25	25	25
Observed unemployment rate in district (%)	32.5	10	32.5

with the underlying process. We may illustrate this with a simplified version of a (semirealistic) example used by LeSage (1998, p. 4), using the numbers set out in Table 8.2. Let there be three districts, each with a labor force of 2,000 adults. Assume that the truth (which we cannot observe) is that a quarter of the residents in each district are unemployed. Of the total of 4,500 jobs, assume that a 1,800 are in the central district, while each of the other districts has 1,350 jobs. It follows that 150 people who live in districts A and C travel to the central district B in order to work. If the unemployment rate is defined as 1 – (number employed/number in the labor force) *in each district*, then the observed rate will be 10% in district B and 32.5% in each of districts A and C. The spatial pattern of unemployment observed in this case is an artifact of the geographic units employed for measurement.

The problem that arises in this case is that if we try to estimate an equation of the form

$$y_i = \mathbf{X}_i\beta + \varepsilon_i, \tag{8.2}$$

then the errors will be spatially correlated, so that $E(\varepsilon_i\varepsilon_j) \neq 0$. In this case, ordinary least squares estimation of (8.2) will not be best linear unbiased ("BLUE"), and the values of R^2 and the t-statistics may be overestimated. These issues must be addressed if one is serious about the validity of the statistical inference behind the OLS model.

8.3.2 *Spatial Heterogeneity*

Spatial heterogeneity refers to variation over space in the relationships themselves (LeSage 1998, p. 6). So, for instance, we might believe that the relationship between fertilizer inputs and the output of maize differs from place to place. Formally, in the case of a linear relationship, we have

$$y_i = \mathbf{X}_i\beta_i + \varepsilon_i, \tag{8.3}$$

where the i refers to observations at 1,. . .,n locations, $\boldsymbol{X}_i$ is a $1 \times k$ vector of explanatory variables, y_i is the dependent variable at location i, and ε_i is a stochastic disturbance with zero mean and constant variance.

If we had enough multiple observations *at each geographic point*, we could estimate (8.2). However, this is rarely possible in practice, given data limitations. Yet if we simply assume that $\beta_i = \beta, \ \forall i$, then we are not resolving the problem that the relationships may vary from place to place.

It is sometimes reasonable to assume that the location that matters is relatively broad – for instance, urban vs. rural areas. This is acceptable if we can assume that the errors are independent of one another *within each given geographic unit.*

This assumption underpins the approach taken by Ravallion and Wodon (1997) in a recent study based on household survey data from Bangladesh. The question they ask is this: "Should poverty programs target households with personal attributes that foster poverty, no matter where they live?" Their answer is "possibly not." In an economy with no apparent constraints on mobility, there may still be sizeable geographic effects on living standards, even after controlling for a wide range of observable variables.

Ravallion and Wodon divide their data into two geographically based groups, urban and rural, and estimate separate models of the determinants of consumption for each. A selection of their regression results, based on data from the 1991/92 Household Expenditure Survey of the Bangladesh Bureau of Statistics, are shown in Table 8.3. They reject the null hypothesis that all coefficients are the same in rural and urban areas – the F-value for a test restricting all 57 coefficients to be equal is 5.63, which rejects equality at the 1% level – and so conclude that "different models are determining consumption in urban and rural areas." This suggests a possible role for geographic targeting.

Their estimates show most of the expected effects: households whose members have more education, or more land, or who work in business or industry, are better off, but these effects are more pronounced in urban areas. Their study includes an extended discussion of the potential for selection bias that might arise because a household's place of residence may not be exogenous; they do not find evidence of such bias, but grant that the available test procedures have their limitations (Ravallion and Wodon 1997, Sect. 3).

8.4 Classifying Spatial Models

It is often the case that geographic variation is more fine-grained than the use of standard regression with spatial variables (as in Kaimowitz et al. 2002) would allow, or than a broad division into a few distinct categories would permit (as in Ravallion and Wodon 1997). In this case, it is necessary to specify the spatial dimensions of our models in more detail.

The most general possible model has the form

$$\mathbf{y} = \xi \mathbf{W}_1 \mathbf{y} + \mathbf{X}\beta + \mathbf{W}_2 \mathbf{X}^* \rho + (\mathbf{I} - \lambda \mathbf{W}_3)^{-1} \varepsilon, \tag{8.4}$$

Table 8.3 Estimates of influences on log real consumption, Bangladesh, 1991–1992

	Urban		Rural	
	Coefficient	SE	Coefficient	SE
Intercept	0.33	0.12	0.19	0.06
Number of children	−0.16	0.02	−0.17	0.01
Number of children squared	0.02	0.00	0.02	0.00
Number of adults	−0.10	0.02	−0.11	0.01
Number of adults squared	0.01	0.00	0.01	0.00
Education of head				
Below class 5	0.15	0.03	0.07	0.01
Class 5	0.16	0.03	0.10	0.02
Classes 6–9	0.28	0.03	0.15	0.02
Higher level	0.42	0.04	0.22	0.03
Land ownership				
0.05 to 0.49 acres	0.08	0.02	0.08	0.02
0.50 to 1.49 acres	0.07	0.03	0.17	0.02
1.50 to 2.49 acres	0.13	0.05	0.17	0.03
2.50 acres or more	0.37	0.08	0.12	0.04
Main occupation				
Factory worker, artisan	0.30	0.06	0.15	0.03
Petty trader, small businessman	0.36	0.05	0.25	0.02
Memo items				
Number of observations	1,908		3,817	
R^2	0.55		0.50	

Source: Ravallion and Wodon (1997), Table 1

Notes: The full regressions include 56 variables. The dependent variable is (the log of) consumption, normalized by a poverty line that reflects the estimated cost of living in the area where the household lives. The standard errors reflect the Huber–White correction for heteroskedasticity. All the coefficients shown are significant at the 5% level of confidence. The excluded categories for the dummy variables are illiterate (for education of head), no land (for land ownership), and landless agricultural worker (for main occupation)

where $\mathbf{y}$ is an $n \times 1$ vector of observations on the dependent variable (e.g., the unemployment rate in an area) and $\mathbf{X}$ is an $n \times k$ matrix of observations on the k independent variables; $\mathbf{X}^*$ is $\mathbf{X}$ without the column of ones corresponding to the constant term; and $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a diagonal matrix. The $\mathbf{W}_i$ are $n \times n$ spatial weights matrices that measure the strength of contiguity; we discuss their construction in more detail in the next section.

The first term on the right-hand side of (8.4) is the spatially lagged dependent variable; the third term on the right-hand side reflects the spatially lagged independent variables; and the error terms here are also spatially lagged. The model in (8.4) could also be written as

$$\mathbf{y} = \xi \mathbf{W}_1 \mathbf{y} + \mathbf{X}\beta + \mathbf{W}_2 \mathbf{X}^* \rho + u, \tag{8.5}$$

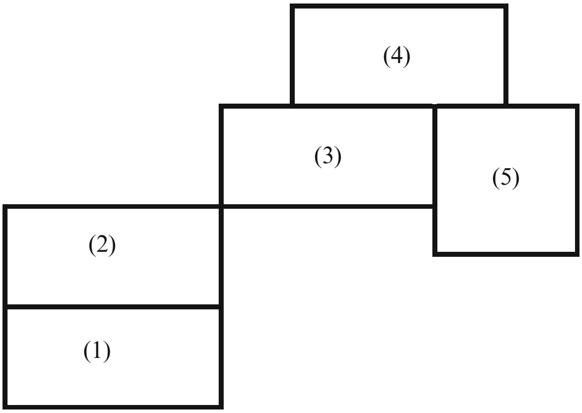

Fig. 8.4 Five regions, to illustrate concepts of contiguity

where $u = \lambda \mathbf{W}_3 u + \varepsilon,$ and ε is a zero-mean random error with diagonal covariance matrix. This simply serves to define the nature of the spatial correlation in the error terms, analogous to first-order autocorrelation in a time-series model.

8.4.1 Measuring Spatial Contiguity

Before proceeding further, we need to ask how one might measure the degree of connectedness between one area and another, with a weights matrix **W**. In this, we follow the presentation in LeSage (1998, Sect. 1.4.1). Suppose that we are studying unemployment in five regions, marked (1) through (5) in Fig. 8.4. Our interest is in creating a 5 × 5 matrix that measures the likely strength of the spillovers from one region to the next.

The key idea in a *contiguity matrix* is to set element $W_{ij} = 1$ if areas i and j are contiguous, and to 0 otherwise. A popular choice is *rook contiguity*, where $W_{ij} = 1$ if the regions share a common side. In the case of Fig. 8.4 this would generate the following matrix:

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}. \tag{8.6}$$

Note that although regions (2) and (3) touch, they do not have a common side. By convention, the diagonal elements of W are zero.

An alternative would be *queen contiguity*, where $W_{ij} = 1$ if regions i and j share a common side or vertex. There are of course other possibilities; for instance, the entries in the W matrix could be in proportion to the length of shared borders; or the time that it takes to travel between the main towns in the areas.

In practice, it is common to normalize the W matrix so that the rows sum to unity. This creates a standardized, first-order contiguity matrix, and would transform (8.6) into the following:

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}. \tag{8.7}$$

If z is a vector of observations on some variable associated with these regions, then the matrix product $z^* = Cz$ creates a variable equal to the mean of observations from the contiguous regions, so

$$\begin{pmatrix} z_1^* \\ z_2^* \\ z_3^* \\ z_4^* \\ z_5^* \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{pmatrix} = \begin{pmatrix} z_2 \\ z_1 \\ 0.5(z_4 + z_5) \\ 0.5(z_3 + z_5) \\ 0.5(z_3 + z_4) \end{pmatrix} \tag{8.8}$$

As LeSage points out, this is one approach to specifying a relationship of the form $y_i = f(y_j), j \neq i$, as set out in (8.1).

8.4.2 Types of Spatial Model

The parameters of the fully general model in (8.4) are not fully identified, but there are a number of special cases that merit further comment.

A good starting point is the relatively simple *first-order spatial autoregressive model*. In this case we assume $\beta = \rho = \lambda = 0$, and we simply have

$$y = \xi \mathbf{W} y + \varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2 I_n). \tag{8.9}$$

In this case the only variable that drives y is the value of the same variable in contiguous areas, appropriately weighted. The least-squares estimate of ξ is biased and inconsistent, but maximum likelihood estimators are available that solve this problem (see LeSage 1998, Chap. 2).

The first-order spatial autoregressive model is rather basic, and is typically augmented with other independent variables, giving rise to the *mixed autoregressive-regressive model*, also known as a *spatial lag model*, or simply the *spatial autoregressive model*, of the form

$$y = \xi \mathbf{W} y + \mathbf{X}\beta + \varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2 I_n). \tag{8.10}$$

The analog in time-series analysis would be a regression model with a lagged dependent variable; here, the dependent variable is lagged spatially rather than temporally, but otherwise the concept is the same. If OLS is applied without the spatially lagged dependent variable, the estimates of the β coefficients will be biased and inconsistent. One can test whether a spatial lag is warranted using a Lagrange multiplier test to determine whether $\xi = 0$; the LM statistic is distributed χ^2 with one degree of freedom.

With a spatially lagged dependent variable on the right-hand side, the interpretation of β changes: it only shows the immediate, but not the total, effect of a change in an explanatory variable on y. Suppose that the value of the kth explanatory variable changes; this affects y in that district – the effect that is captured by β_k – but the change in y in turn influences the value of y in neighboring districts (through the $\xi\mathbf{W}y$ term), which feeds back to affect y in the home district. The eventual effect is that y will change by $C^{ii}\beta_k$, where C^{ii} is the *(i,i)* element of $(\mathbf{I} - \xi\mathbf{W})^{-1}$.

One of the most popular models that take spatial effects into account is the *spatial errors model*. In this case we have

$$y = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2 I_n). \tag{8.11}$$

The structure of this model is analogous to a time-series model with first-order autocorrelation in the errors. A shock in one area propagates to neighboring areas; Anselin and Florax (1995) refer to this as "nuisance spatial dependence."

Traditionally, the test of whether the error structure is of the form shown in (8.11) was based on Moran's I. If $\mathbf{W}$ is a standardized contiguity matrix, then $I \equiv \varepsilon'\mathbf{W}\varepsilon/\varepsilon'\varepsilon$ should be distributed as asymptotically normal. This is of course a test for the presence of spatial effects as captured by the particular weights matrix $\mathbf{W}$; a different choice of weights matrix might generate a different value of Moran's I.

Rather than using Moran's I, it has become increasingly common to use a Lagrange multiplier test first proposed by Burridge (see Anselin 1992, p. 179; also Anselin 1998) the LM statistic is distributed χ^2 with one degree of freedom, and a large value of the statistic would suggest that a spatial error model would be appropriate.

If the spatial errors model is the correct one, but one uses least squares to estimate (8.11) on the assumption that $\lambda = 0$ (i.e., without spatially correlated errors), then the estimates of the β coefficients will be inefficient.

The spatial errors model can also be written as

$$y = \lambda\mathbf{W}y + \mathbf{X}\beta - \lambda\mathbf{W}\mathbf{X}\beta + \varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2 I_n). \tag{8.12}$$

which in turn is a special case, sometimes referred to as the common factor hypothesis model, of the more general model

$$y = \lambda\mathbf{W}y + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\rho + \varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2 I_n). \tag{8.13}$$

One can test whether the common factor hypothesis is an acceptable simplification by applying a Lagrange multiplier test to check whether $\rho = -\lambda\beta$.

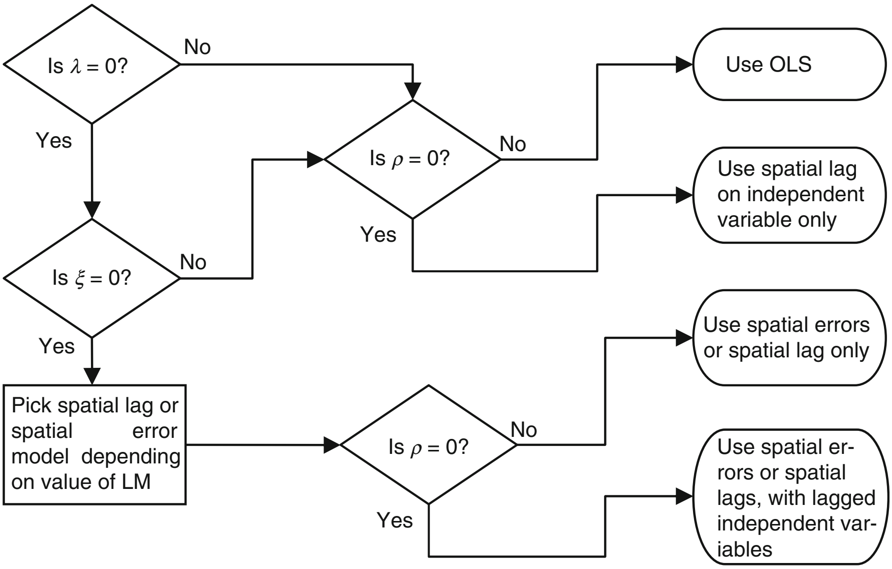
Is λ = 0?
No
Yes
Use OLS
Is ρ = 0?
No
Yes
Use spatial lag on independent variable only
Is ξ = 0?
No
Yes
Use spatial errors or spatial lag only
Pick spatial lag or spatial error model depending on value of LM
Is ρ = 0?
No
Yes
Use spatial errors or spatial lags, with lagged independent variables

Fig. 8.5 Decision diagram for choosing the most appropriate spatial model

A more *general spatial model* would allow spatial lags both in the dependent variable and in the error, and would take the form

$$y = \xi \mathbf{W}_1 y + \mathbf{X}\beta + (\mathbf{I} - \lambda \mathbf{W}_3)^{-1}\varepsilon, \quad \text{with } \varepsilon \sim N(0, \sigma^2 I_n). \tag{8.14}$$

It is possible to have $\mathbf{W}_1 = \mathbf{W}_3$, but this sometimes leads to identification problems.

It is natural to ask when the general spatial model would be preferable to a simpler model. At the intuitive level, it would make sense to use, or at least consider, such a model if there is evidence of spatial dependence in the error structure of the spatial autoregressive model; or if the contiguity differences can reasonably be assumed to differ – for instance, if one is a first-order, and the other a second-order, contiguity matrix; or if one is a contiguity matrix and the other is a matrix that shows, for instance, the distance between the central cities in the regions.

Given the abundance of choices of model, it would be valuable to have an algorithm for choosing the "best" model. Florax and Folmer (1992) suggest that one should first test whether it is appropriate to include autoregressive disturbances (i.e., is $\lambda = 0$?) or a spatially lagged dependent variable (i.e., is $\xi = 0$?). If neither should be included, one may then test whether spatially lagged *independent* variables should be included (i.e., is $\rho = 0$?).

On the other hand, if the initial test strongly suggests autoregressive errors, then this is the model to use. And if the initial test favors a lagged dependent variable, then one should test whether spatially lagged independent variables should be included as well. Anselin (1992, p. 180) proposes that as a practical matter, one could choose between a spatial error, and a spatial lag model, on the basis of which LM statistic is larger. The relevant decision diagram is set out in Fig. 8.5.

8.4.3 Illustrating the Choice of Spatial Model

We illustrate the process by which one might choose an appropriate spatial model using the study by Aragon et al. (2003) of the determinants of local unemployment rates in the Midi-Pyrénées region of France. The key results are set out in Table 8.4, where column 1 shows the estimates based on ordinary least squares estimation.

The dependent variable is the unemployment rate, which is measured in each area; these hinterlands, called "bassins de vie quotidienne" in French, are typically centered on a town. The results show that unemployment is higher if the labor force in an area has a high proportion of young workers or old people, if incomes are higher, and if an area is more densely populated. Aragon et al. argue that unemployment rates are higher in more-urbanized and richer areas as a result of the better amenities and jobs offered in these areas, which attract job-seekers who are willing to tolerate longer unemployment in the hope of obtaining higher-paying jobs in interesting locations.

The residuals from the equation whose estimates are shown in column 1 of Table 8.4 are mapped in Fig. 8.1, and are spatially clustered. This visual evidence of spatial interaction is confirmed by the Lagrange multiplier (LM) statistics reported at the bottom of column 1. Specifically, a test of $\lambda = 0$ is rejected (LM $=$ 70.4, p-value $=$ 0.00),which would argue in favor of a spatial errors model. The results of estimating such a model are displayed in column 2 of Table 8.4. A formal test shows that there is no remaining spatial lag dependence (LM $=$ 0.14, p-value $=$ 0.71). This also shows in the map of residuals from this equation, which is displayed in Fig. 8.6, and where the errors now appears to be distributed randomly across the region, in contrast to the clumping of the OLS residuals that is evident in Fig. 8.1. It is worth noting that the index of "enclavement" – essentially a measure of remoteness – is statistically significant in the OLS model, but not when one allows for spatially autocorrelated errors. This suggests that spatial autocorrelation is picking up the effects of remoteness satisfactorily.

The high value of λ, at 0.79, indicates that shocks in one area propagate outward to the neighboring areas to a very substantial degree. The pain of a local shock is thus spread widely, illustrating the strong degree of connectedness across hinterlands in the Midi-Pyrénées region.

It is worth exploring further whether the common factors model with autocorrelated errors is indeed the best choice. First one might ask whether a more general model, as in (8.13), would be preferable; however, a formal test (LR $=$ 9.4, p-value $=$ 0.23) does not allow us to reject the common factors specification, as in (8.12).

It is also clear from column 1 in Table 8.4 that a test of $\xi = 0$ is also rejected (LM $=$ 55.2, p-value $=$ 0.00), which suggests that a spatial lag model might be appropriate, even though the lower LM statistic in this case establishes a presumption in favor of a spatial error model. The results of estimating a spatial lag model are shown in column 3 of Table 8.4. Unfortunately, this model does not remove spatial autocorrelation in the errors. Moreover, it performs less well than the spatial error model as measured by the values of the AIC (Akaike information criterion)

Table 8.4 Regression results of models of the determinants of unemployment in the Midi-Pyrénées region of France, 1990

	Column 1	Column 2 Auto-correlated errors	Column 3 Lagged dependent variable	Column 4 Lagged independent vars	
				Unlagged terms	Lagged terms
	OLS	MLE	MLE	MLE	
% Lab force 20–39	0.171 (0.00)	0.224 (0.00)	0.194 (0.00)	0.206 (0.00)	..
% Pop 60–64	0.430 (0.00)	0.394 (0.00)	0.580 (0.00)	0.424 (0.00)	..
% Employment secondary sector	0.032 (0.01)	0.029 (0.00)	0.026 (0.01)	0.27 (0.00)	−0.048 (0.03)
SE monthly unemployment rates	1.228 (0.00)	0.921 (0.00)	1.345 (0.00)	0.925 (0.00)	..
Ln(population/ha)	0.010 (0.00)	0.008 (0.00)	0.010 (0.00)	0.010 (0.00)	..
Index of “enclavement”	0.074 (0.00)	0.015 (0.38)	0.056 (0.00)	0.041 (0.02)	0.078 (0.01)
Taxable income/wkr (‘000)	0.568 (0.00)	0.579 (0.00)	0.487 (0.00)	0.563 (0.00)	−0.645 (0.00)
Constant	−0.087 (0.00)	−0.110 (0.00)	−0.137 (0.00)	−0.102 (0.00)	..
Memo items					
Number of observations	174	174	174	174	
Log likelihood	454.4	489.8	478.3	493.2	
LM ($\lambda = 0?$) [test of spatial errors]	70.40 (0.00)		13.09 (0.00)	0.11 (0.73)	
λ		0.790 (0.00)			
ξ			0.456 (0.00)	0.679 (0.00)	
LM ($\xi = 0?$) [test of spatial lags]	55.16 (0.00)	0.14 (0.71)		62.41 (0.00)	
LR ($\rho = -\lambda\beta?$) [test of common factors]		9.4 (0.23)			
AIC	−890.7	−959.6	−936.5	−960.4	
BIC	−862.5	−928.3	−905.2	−919.7	

Source: Aragon et al. (2003). *Notes*: Bracketed values are *p*-values from *t*(*z*) tests. Data source is the FIDEL database, INSEE 1996. Dependent variable is the unemployment rate. *OLS* ordinary least squares, *MLE* maximum likelihood estimation, *LM* Lagrange multiplier, *LR* likelihood ratio, *AIC* Akaike information criterion, *BIC* Bayes information criterion. “..” denotes not significant at the 10% level. The index of “enclavement” is 0 if all services are found locally; a higher number means that the hinterland is more economically remote

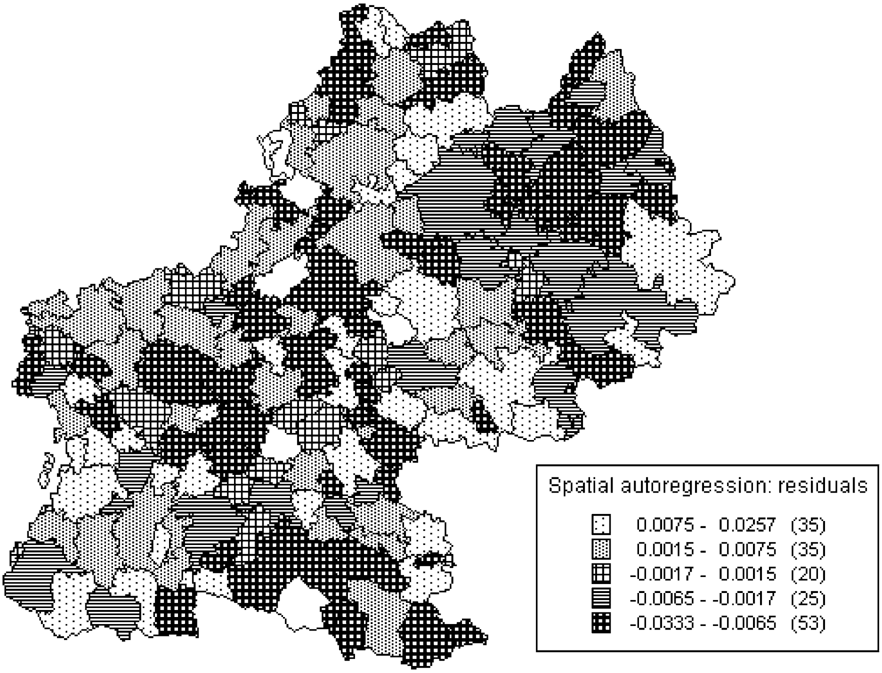

Fig. 8.6 Residuals from regression, with autocorrelated errors, of the determinants of the unemployment rate for hinterlands in the Midi-Pyrénées region of France, 1990. (*Source*: Aragon et al. 2003)

and BIC (Bayes information criterion), where lower values generally denote a more satisfactory model.

The last two columns of Table 8.4 show the results of estimating an augmented version of the spatial lag model, where not only is a lagged version of the unemployment rate included on the right-hand side, but also spatially lagged values of the independent variables. This model has virtues: there is no remaining spatial autocorrelation in the errors, and it has the highest log likelihood and lowest AIC of the models considered here. On the other hand, it is a less parsimonious model than the spatial error model, and this is reflected in lower (and hence better) value of the BIC for the latter model. When trying to choose between competing models, the BIC is typically a better guide than the AIC, because the AIC tests to favor larger and potentially overfitted models (see Haughton et al. 1990).

8.5 Other Spatial Models

There are a number of other possible approaches to modeling spatial effects; in this section we briefly summarize two of these – the spatial expansion model and geographically weighed regression. LeSage (1998) provides further details.

8.5.1 Spatial Expansion Models

When we believe that there is spatial heterogeneity, so that the relationships between variables differ from place to place, we argued in Sect. 8.3 that one might want to estimate a model of the form

$$y_i = \mathbf{X}_i\beta_i + \varepsilon_i. \tag{8.3}$$

The problem is that we rarely have enough observations at each location to allow us to estimate the β_i. However, if we could impose some structure on the β_i, then estimation might be feasible. This is the approach taken by Casetti (1972) in the spatial expansion model, where the maintained assumption is that the parameters of the model vary systematically with latitude and longitude. Each observation has coordinates of latitude and longitude, represented by $(Z_{d,i}, Z_{g,i}),\ i = 1, \ldots, n.$

To see how this model works, consider the case in which there are three independent variables (including a constant), so we have

$$y_i = \beta_{0,i} + \beta_{1,i}X_{1,i} + \beta_{2,i}X_{2,i} + \varepsilon_i, \tag{8.15}$$

where the i observations refer to the n households. The coefficients are themselves influenced by the measures of latitude and longitude, so we have

$$\begin{aligned}\beta_{0,i} &= \gamma_{0,d}Z_{d,i} + \gamma_{0,g}Z_{g,i},\\ \beta_{1,i} &= \gamma_{1,d}Z_{d,i} + \gamma_{1,g}Z_{g,i},\\ \beta_{2,i} &= \gamma_{2,d}Z_{d,i} + \gamma_{2,g}Z_{g,i}.\end{aligned} \tag{8.16}$$

After substituting (8.16) into (8.15), the model can be estimated using ordinary least squares, which yields estimates of the $2k$ parameters $\gamma_{k,d}$ and $\gamma_{k,g}$; since $k = 3$ in our example, there are six such parameters. From these one can use (8.16) to generate estimates of the $\beta_{k,I}$ coefficients for each observation. LeSage (1998) discusses the general case for k variables.

The model ensures that areas that are close neighbors will have similar coefficients, and so it does not allow for sharp discontinuities, for instance when a rich neighborhood abuts a poor area.

It is reasonable to imagine that there might be some error – captured by a stochastic term u_i – in the expansion relationship in (8.16), giving terms such as

$$\beta_{0,i} = \gamma_{0,d}Z_{d,i} + \gamma_{0,g}Z_{g,i} + u_{0,i}. \tag{8.17}$$

Substituting these into (8.15) gives a model with a composite error term that will in general be heteroscedastic. LeSage (1998, Chap. 4) discusses how to correct for this in estimating the parameters of this model.

8.5.2 Geographically Weighted Regression

If geography matters, then the relationships between the **X** and y variables may vary over space. These may be approximated by a series of locally linear regressions, much as is done in standard nonparametric regression. The key idea behind these *geographically weighted regressions* is to use distance-weighted subsamples of the data in order to estimate locally linear regression estimates at every point in space.

Each observation i has a vector-valued distance-based weight W_i. For instance, one popular method for constructing the weight function (Brunsdon et al. 1996) is to set

$$W_i^2 = \exp(-d_i/\theta), \tag{8.18}$$

where each element of the vector d_i is the geographic distance between location i and each other location in the dataset, and θ is a decay parameter. The distance vector is typically measured as

$$d_i = \sqrt{(Z_{d,i} - Z_{d,j})^2 + (Z_{g,i} - Z_{g,j})^2}, \tag{8.19}$$

where the $Z_{d,j}$ and $Z_{g,j}$ measure the latitude and longitude of the observations $j = 1,\ldots,n$. The decay parameter θ determines how quickly the influence of location tails off with greater distance. Other methods of constructing the weighting function are of course possible.

The geographically weighted regression model may be written as

$$\mathbf{W}_i^{1/2} y = \mathbf{W}_i^{1/2} X\beta_i + \varepsilon_i. \tag{8.20}$$

Here, $\mathbf{W}_i$ is an $n \times n$ diagonal matrix with the distance-based weights for observations i on its diagonal, y is an $n \times 1$ vector of observations on the dependent variable collected at n points in space, X is an $n \times k$ matrix of data on the k explanatory variables, and ε is an $n \times 1$ vector of normally distributed disturbances with constant variances (LeSage 1998, p. 155). The interesting point here is that there is a vector of parameter estimates for each of the i observations.

The estimates of β_i differ for each location, but they are based on the same sample of data, and so are not independent. This lack of independence means that one cannot draw inferences from the regression parameters. LeSage also points out that the results are highly sensitive to outliers, and to sparse data; he argues that these problems can be largely overcome by taking a Bayesian approach to estimation.

8.5.3 Spatial Effects as Random Effects

In her study of the demand for rice in Indonesia, Anne Case (1991) has survey information on 2,089 households in 141 districts in Indonesia, where each district is included only if it abuts at least one other district in the sample. She specifies a model with both spatial errors and spatial lags, of the form

$$\mathbf{y} = \xi \mathbf{W}\mathbf{y} + \mathbf{X}\beta + u, \tag{8.21a}$$

$$u = \lambda \mathbf{W} u + \varphi + \varepsilon. \tag{8.21b}$$

There are T districts with N households per district, so ε is a $TN \times 1$ vector of random errors with zero mean and constant variance. There is also a district-specific random error term φ, where we have $E(\varphi_k) = 0$ for household k in district i, and $E(\varphi_k \varphi_j) = {\sigma_\varphi}^2$ if $j \in i$ but is 0 otherwise. Note that we could rewrite (8.21a) and (8.21b) as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon + [\xi \mathbf{W}\mathbf{y} + \lambda \mathbf{W} u + \varphi]. \tag{8.22}$$

The bracketed expression in (8.22) is the same for every observation in a given district and is in effect a vector of "constrained random effects." This random effect is composed of ξ times the average value of **y** in abutting districts, plus λ times the average error in abutting districts, plus a nonspatial district-specific error (Case 1991, p. 958).

We may also think of (8.22) as a special case of a model that incorporates fixed effects for every district, and that takes the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{D}\theta + \varepsilon, \tag{8.23}$$

where **D** is a $TN \times (T - 1)$ matrix of district dummy variables. One may test formally whether the model in (8.22) represents an appropriately constrained version of the fixed effects model in (8.23) using a Hausman test, although there are also times when it is more helpful to work with the spatial effects structure – for instance, if the question of interest is whether households mimic their neighbors.

It is also possible to use multilevel models, with random effects that capture geographic disparities that are unexplained by the available predictors; we return to this topic in Chap. 12 (see too Haughton and Phong 2010).

8.6 Conclusion

Data from living standards surveys always have a spatial dimension, and a growing number of studies explicitly take this into account. For instance, in his study of the impact of the Impres Desa Tertingal program in Indonesia – a program of block grants to poor areas – Daimon (2001) finds that a spatial lag model is called for. And Druska and Horrace (2004) use spatial techniques to improve their measures of the efficiency of rice farmers in Indonesia, estimating a stochastic frontier using panel data. In this case productivity shocks, which could be related to weather or unobserved influences, spill over from one area to the next.

As the technology for estimating spatial models becomes more accessible, and geographic data become more easily available, it will become increasingly standard to consider spatial models. The use of spatial models does not always change the standard coefficient estimates by much, but it changes the way in which we think about the spread of economic effects. This is how economists and statisticians are rediscovering geography.

8.6.1 Estimating Spatial Models

To estimate spatial models, a relatively straightforward solution is to draw on the Matlab routines that have been developed and well-documented by Jim LeSage (LeSage 1998). Luc Anselin's *SpaceStat* (Anselin 1992) is still used by some. For users of Stata, Maurizio Pisati has written a routine for spatial regression called `spatreg`, and Mark Pearce has developed a routine called `gwr` that fits geographically weighted regression. There is now a large library of relevant routines in R, most notably the `spdep` package maintained by Roger Bivand, and the `spgwr` package that does geographically weighted regression.

References

Anselin, Luc. 1992. *SpaceStat tutorial*. Urbana-Champaign: University of Illinois.

Anselin, Luc. 1998. *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic.

Anselin, Luc, and R. Florax. 1995. Introduction. In *Net directions in spatial econometrics*, ed. L. Anselin and R. Florax. Berlin: Springer.

Aragon, Yves, Dominique Haughton, Jonathan Haughton, Eve Leconte, Eric Malin, Anne Ruiz-Gazen, and Christine Thomas-Agnan. 2003. Explaining the pattern of regional unemployment: The case of the Midi-Pyrénées region. *Papers in Regional Science* 82: 155–174.

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

Case, Anne. 1991. Spatial patterns in household demand. *Econometrica* 59(4): 953–965.

Casetti, E. 1972. Generating models by the expansion method: Applications to geographic research. *Geographical Analysis* 4: 81–91.

Daimon, Takeshi. 2001. The spatial dimension of welfare and poverty: Lessons from a regional targeting programme in Indonesia. *Asian Economic Journal* 15(4): 345–367.

Druska, Viliam, and William Horrace. 2004. Generalized moments estimation for spatial panel data: Indonesia rice farming. *American Journal of Agricultural Economics* 86(1): 185–198.

Florax, R., and H. Folmer. 1992. Specification and estimation of spatial linear regression models: Monte Carlo evaluation and pre-test estimators. *Regional Science and Urban Economics* 22: 405–432.

Haughton, D., J. Haughton, and A. Izenman. 1990. Information criteria and harmonic models in time-series analysis. *Journal of Statistical Computation and Simulation* 35: 187–207.

Haughton, Dominique, and Nguyen, Phong. 2010. Multilevel models and inequality in Vietnam. *Journal of Data Science* 8: 289–306.

INSEE. 1996. *FIDEL Base de Données – 1996: Guide de L'Utilisateur*. Paris: Institut National de la Statistique et des Études Économiques.

Kaimowitz, David, Patricia Mendez, Atie Puntodewo, and Jerry Vanclay. 2002. Spatial regression analysis of deforestation in Santa Cruz, Bolivia. In *Land use and deforestation in the Amazon*, ed. C.H. Wood and R. Porro. Gainesville: University Press of Florida.

LeSage, James. 1998. *Spatial econometrics*. Toledo: Department of Economics, University of Toledo.

Ravallion, Martin, and Quentin Wodon. 1997. Poor areas, or only poor people? Policy Research Working Paper 1798. Washington, DC: World Bank.