Diagnosing, Modeling, Interpreting, and Leveraging Spatial Relationships in Time -Series-Cross-Section Data

Social scientists have long recognized that observations in time-series-cross-section (TSCS) datasets usually correlate across time and space. Today, standard practice in political science is to model dynamics (*i.e.*, temporal dependence) directly, typically with lags of the dependent variable, but to address spatial dependence solely by applying panel-corrected (robust) standard-errors (PCSE's), thereby treating spatial dependence as a *nuisance* (Beck and Katz's 1996 terminology). As we explain below, however, direct modeling of spatial dependence (plus robust standard-errors perhaps) is superior, regardless of the analyst's substantive interest in these relationships. Directly modeling spatial dependence always enhances efficiency and, moreover, is often necessary to obtain unbiased and consistent estimates of the coefficients on non-spatial regressors and all associated hypothesis tests and confidence intervals. Understandably, analysts less interested in spatial relationships *per se* will want to employ simple proxies for more complicated diffusion processes. Those directly interested, on the other hand, will prefer more sophisticated modeling techniques for estimating the dyadic patterns of diffusion in their data. In this project, we use Monte Carlo experiments to evaluate the performance of several simple and sophisticated estimators under three important types of spatial correlation (the more-complex methods being spatial analogues to estimators devised for dynamic panel models: *e.g.*, Hsiao 1986; Baltagi 1995). We also develop a set of techniques and guidelines to help analysts diagnose spatial correlation and choose and interpret appropriate estimators based on their objectives.

We thus address this project both to social-science researchers directly interested in spatial relationships (*substance*) and to those primarily concerned to make optimal inferences regarding other substantive relationships given spatially dependent data (*nuisance*). Building from analogies to similar, better-explored issues arising from temporal dependence, and through analytic derivation and Monte Carlo experimentation, we will: (1) detail conditions under which failing to model spatial dependence or relegating its role to standard-error adjustment biases other coefficient estimates or *merely* induces inefficiency, exploring the magnitudes of these biases and inefficiencies under differing degrees and natures of spatial dependence; (2) distinguish conceptually spatial diffusion from spatially correlated responses to omitted spatially correlated factors and evaluate alternative approaches to making such distinctions empirically; (3) develop and explore the properties of several parametric, semi-, and non-parametric methods of testing for spatially correlated disturbances, in the presence or absence of spatial lags in the model; (4) compare the properties of simple proxies for full models of the true spatial-diffusion or common omitted-factors—*e.g.*, spatial dummies or symmetric spatial-lags comprised of averages of other cross-section units' dependent variables each time-period—to each other, to PCSE's alone, and to different techniques for estimating the full model; and (5) explore the properties of leveraging spatial dependence to aid identification of endogenous systems by using spatial lags as *quasi-instruments* (Bartels 1991) under differing magnitudes of the relevant spatial correlations and causal arrows: $\mathbf{X}$? $\mathbf{Y}$, $\mathbf{X}_i$? $\mathbf{X}_j$, and $\mathbf{Y}_i$? $\mathbf{Y}_j$. As we aim to help analysts of varying statistical sophistication and substantive interest in spatial relationships develop intuitions and techniques of use in diagnosing spatial correlation and choosing and interpreting appropriate estimators for their objectives, we will also (6) create and publish free statistical-software algorithms in widely used software packages, such as Stata, to implement all of our suggested techniques and, where possible and productive, pedagogical modules to help teach them.

(We note that Beck and Katz, who introduced the extremely useful and now almost universally employed panel-corrected standard-errors, PCSE's, have a current project that likewise addresses some gaps in political methodology regarding spatial dependence. Although they intend to explore some applications of spatial lags to model spatial dependence related to our intended explorations, our foci in that and in other regards differ from and complement theirs. They propose to consider (a) non-geographic analogues to spatial-lag models of contemporaneous correlation, and (b) modeling approaches to spatial parameter-heterogeneity, and models that address spatial dependence in (c) nominal and in (d) non-stationary data. Complementarily, as just enumerated above and as elaborated below, we propose to explore (a) the conditions that determine whether and to what degree failing to model spatial dependence directly biases other coefficient estimates or *merely* induces inefficiency, (b) approaches to distinguishing spatial diffusion from spatially correlated responses conceptually and empirically, (c) tests for spatially correlated disturbances, with and without spatial lags in the model, (d) comparison of simple spatial-indicator and spatial-lag proxies to each other, to PCSE's alone, and to fuller, more-complex models of spatial dependence, and (e) the leveraging of spatial dependence to aid identification of endogenous systems by using spatial lags as *quasi-instruments* (see Bartels 1991).)

1. Thinking about Spatial Relationships

Much political-science research uses time-series-cross-section data: datasets that contain observations on several cross-sectional units over multiple time-periods. The units are typically interdependent, and this interdependence, which may or may not relate to geographic proximity, is termed *spatial*. More precisely, spatially correlated units are cross-sectional units with variables that correlate *contemporaneously*. (One should note that what is contemporaneous depends upon the data's level of temporal aggregation. Spatial processes that take weeks or months appear instantaneous in annual data.) While political scientists widely recognize this interdependence, the difficulties that spatial correlation can create for empirical research have received little consideration beyond Beck and Katz's seminal work. Along with the other NSF-funded project just mentioned, our project seeks to redress key aspects of these gaps in the political methodology literature.

We begin conceptually, noting two distinct sources of spatial correlation: spatial diffusion across units and common (or correlated) shocks to all units. In the former process, an internal change within one unit affects other units. Common shocks, contrarily, originate from an external source and affect multiple units simultaneously. Outcomes in a group of oil-dependent economies, *e.g.*, may correlate spatially because each experiences the same external price-shocks. To develop these ideas formally, we start with a basic model

$$\mathbf{y} = \mathbf{X}\text{ß} + \mathbf{e}, \qquad (1)$$

where **y** is a *TxN* matrix of *T* outcomes (1 *per* time period) for N units (1 *per* cross-sectional/spatial unit), **X** is a *TxNK* matrix of values on *K* independent variables, **e** is a *TxN* matrix of spatially correlated errors, and **ß** is an *NKxN* matrix of coefficients. To simplify exposition, we assume henceforth that *K=1* and that the coefficients **ß** are identical in each spatial unit. Under these assumptions, **ß** simplifies to an *NxN* matrix with common diagonal element *ß* and zeros off-diagonal. (More generally, $\text{ß}=\mathbf{I}\otimes\mathbf{S}$, where **I** is an *NxN* identity matrix and **S** is a *Kx1* vector of coefficients.) Decomposing **e** into component parts, we have

$$\mathbf{y} = \mathbf{S}\,\mathbf{W}\,? + \mathbf{X}\,\text{ß} + ? + ? \qquad (2)$$

where **W** is an *NxN* spatial diffusion matrix, which contains zeros along the diagonal ($w_{ii}=0$) and some non-zero elements off the diagonal; the cell w_{ij} measures the degree of diffusion from unit *i* to unit *j* and w_{ji} measures the degree of diffusion from unit *j* to unit *i*. Note that **W**, which fully describes the pattern of spatial diffusion, has up to $N\cdot(N-1)$ unique elements. Thus, the number of spatial-diffusion terms to estimate grows much faster than *N*, which, except for exceptionally large *T* relative to *N*, underscores the magnitude of the challenge spatial diffusion presents for empirical researchers, and explains why those directly interested in such spatial relationships require some model (*i.e.*, theoretically reduced parameterization) of the diffusion process, and why those less-substantively interested require some simplified proxy for them. The matrix **?** is a *TxN* matrix with a common row element, representing the spatially common component of the shock in each time period; and **?** is a *TxN* matrix with independent elements, *i.e.*, each element represents the component of the shock unique to each spatial unit in each time period. The *TxN* matrix **S** is the sum of **X**, **?**, and **?**, so that **SW** is the *spatial lag* (see Anselin 1988), each element of which is a diffusion-matrix-weighted average of the independent variables and shocks elsewhere in the cross-section that time period. Similar to **ß**, the *NxN* matrix **?** has zeros off the diagonal and a common diagonal element *?*, giving the coefficient on the spatial lag. With these definitions, we can write the equation for each element of the outcome matrix as

$$y_{n,t} = r\,\mathbf{s}_\mathbf{t}\mathbf{w}_\mathbf{n} + b\,x_{n,t} + h_t + u_{n,t}, \qquad (3)$$

where $\mathbf{s}_\mathbf{t}$ (*1xN*) is the sum of the t^{th} rows from **X**, **?**, and **?**, and $\mathbf{w}_\mathbf{n}$ (*Nx1*) is the n^{th} column from **W**. Note the first and penultimate terms reflect what we call *spatial diffusion* and *common shocks* respectively.

Most political scientists view such spatial dependence as *nuisance* (Beck and Katz 1996), relegating it to a mere standard-error adjustment role *via* panel-corrected (robust) standard-errors. That so few political scientists model the spatial relationships in their data directly is surprising (Ward and O'Loughlin 2002 may evidence a recent change), though, because these spatial relationships not only map how shocks in a particular area have effects throughout a region (first and penultimate terms) but also determine the cumulative effect of such shocks *via* diffusion (first term). In this project, we demonstrate why best practice is always to model such spatial dependence, regardless of one's substantive interest in these relationships. This approach adds efficiency and, under many circumstances, is necessary to obtain unbiased and consistent estimates of the coefficients on non-spatial regressors (and hypothesis tests and confidence intervals along with them).

Understandably, analysts less interested in spatial relationships *per se* will want to employ simple proxies for such complex processes. Those more-directly interested, conversely, will prefer more-sophisticated modeling techniques to estimate the substantively rich patterns of diffusion in their data. Accordingly, we explore three methods of addressing spatial correlation along a continuum in terms of simplicity. One simple way to address spatial correlation is to include time-period dummies. In theory, this approach assumes that the spatial correlation arises from common external shocks alone, but, in practice, the dummies will also partially proxy for any spatial diffusion present. A slightly less simplistic method might include average values of the dependent variable in the N-1 other cross-sectional units for each time period (see, *e.g.*, Franzese 1999, 2002a, 2002b). Theoretically, this approach assumes a symmetric spatial-diffusion process (*i.e.*, each unit affects all other units and all other units affect it equally) without common shocks, but again, in practice, it will proxy partially for both asymmetrical diffusion and common shocks. Thus, the estimated coefficients on either of these simple spatial proxies will have ambiguous interpretation in practice, although this may be less problematic for analysts less-centrally interested in interpreting the spatial dependence. More thorough approaches would estimate the patterns and strength of spatial diffusion in one's data, for which, again, relatively simpler or more-complicated estimation techniques exist (*e.g.*, 2-stage or FIML estimation).

Accordingly, one part of our project will be to evaluate how well, under varying independent-variable and stochastic-term conditions, these and other such simple or complex approaches to spatial dependence can distinguish spatial diffusion from common shocks, and how well they can estimate the spatial and non-spatial regressors' coefficients and standard errors, relative to each other and to estimating PCSE's alone. Given this evaluation, which will suggest some form of direct model for spatial dependence, simpler or more-complex depending on the degree and pattern of those correlations, researchers will logically need tools to assess the spatial correlation in their data and, ultimately, in their estimated residuals. Most critically, note that, as with temporal lags, spatial lags can induce bias if and to the degree that stochastic components retain correlation controlling for the model of the systematic component. Accordingly, our project will next develop and explore the properties of several parametric, non-, and semi-parametric tests for spatial correlation in the presence or absence of spatial lags in the model (see below). Then, we explore the properties of leveraging spatial correlation to help identify endogenous systems of equations. In political economy, *e.g.*, many social science researchers (e.g., Alvarez, Garrett, and Lange 1991; Beck, Katz, Alvarez, Garrett, and Lange 1993, Franzese 1999, 2002a, 2002b) include economic conditions abroad as controls or substantively central factors in their analyses, assuming them exogenous. Insofar as this exogeneity assumption holds, these factors might prove highly useful as instruments for identifying endogenous systems. Unemployment abroad, *e.g.*, might instrument for domestic unemployment in estimating the effect of unemployment on government spending, which latter some suggest partially causes the former, implying endogeneity and the need to identify by instrumentation or some other means. However, if outcomes abroad affect domestic conditions, one criterion for a useful instrument, then quite likely domestic conditions affect outcomes abroad, violating the other condition for a perfect instrument. Thus, researchers attempting to leverage spatial correlation to help identify endogenous systems, or even simply employing external conditions as an assumed-exogenous factor explaining domestic outcomes—and these possible applications extend well beyond political economy or political science to any study in time-series-cross-section data—must recognize that such external conditions are partly endogenous and so only *quasi-instruments* (Bartels 1991). We will therefore explore the properties of leveraging spatial dependence to aid identification of endogenous systems by using spatial lags as *quasi-instruments* (Bartels 1991) under differing magnitudes of the relevant spatial correlations and causal arrows: $\mathbf{X}$? $\mathbf{Y}$, $\mathbf{X}_i$? $\mathbf{X}_j$, and $\mathbf{Y}_i$? $\mathbf{Y}_j$. Finally, to maximize the broader impact of our study within and beyond social science, we will create and publish free statistical-software algorithms and pedagogical modules to further the dissemination and use of our suggested techniques.

In these explorations, we will consider at least three types of spatial dependence: 1) spatially correlated disturbances with orthogonal regressors, 2) spatially correlated disturbances with spatially correlated regressors, and 3) orthogonal disturbances with spatially correlated and endogenous regressors. We introduce these three cases below, describing the methodological problems that arise and highlighting the questions for which researchers who work with TSCS data, both those directly interested in spatial relations and those not, need answers. For illustration purposes, the proposal provides the results of one-shot simulations for each case. Obviously, these simulations and the discussion below are preliminary, one central task of this project being, of course, to develop these one-shot simulations into a full set of Monte Carlo experiments. In the final section, we review our plan of research.

2. Spatially Correlated Disturbances with Exogenous and Orthogonal Regressors

We start with the case of spatially correlated errors and spatially orthogonal regressors, which is represented by the following set of variance-covariance matrices:

$$\begin{aligned} &\mathbf{var[e]} = s^2\mathbf{C} \\ &\textit{and} \\ &\mathbf{var[X]} = y^2\mathbf{I} \end{aligned} \qquad (4)$$

The parameters s^2 and $?^2$ are scalars. The matrix **C** is *NxN* with ones along the diagonal and non-zero off-diagonal elements. The variance-covariance matrix **var[e]** is, therefore, an *NxN* matrix with s^2 as the diagonal and non-zero off-diagonals. These off-diagonals are a function of common shocks. The matrix I is an *NxN* identity matrix. The variance-covariance matrix var[X] is an *NxN* diagonal matrix with $?^2$ the diagonal.

Efficiency in Estimating b (Table 1, Analyses II-III)

One reason that analysts might want to model the spatial relationships in their TSCS data is to improve the precision of the estimates on their non-spatial regressors. The basic idea is to *soak up* some of the residual variance and thus reduce the standard errors of $\hat{\beta}$. To provide a check on this intuition and get some sense of the magnitude of these efficiency gains, we provide the results of a one-shot simulation exercise in Table 1, Analyses I-III. The true model that generated the dataset of 20 cross-sectional units over 30 time periods for this exercise is $y_{n,t} = .2\mathbf{s_t w_n} + .8x_{n,t} + h_t + n_{n,t}$, where each $x_{n,t}$, $?_t$, and $?_{n,t}$ is a random draw from a *N(0,1)* distribution; and **W** is the following, arbitrary, asymmetrical spatial-diffusion matrix.

$$\mathbf{W} = \begin{pmatrix}
0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\
0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 \\
0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 \\
0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 \\
0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 \\
0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 & 0.6 \\
0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 & 0.7 \\
0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.8 \\
0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 \\
0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} \qquad (5)$$

As a benchmark against which to gauge the efficiency gains, we regress **Y** on a constant and **X**. The estimated coefficient on **X** will be inefficient but unbiased, because the omitted spatially correlated factors do not correlate with the included independent (*ergo* spatially uncorrelated) $x_{n,t}$ draws. Table 1 (Column 1)

Table 1: Spatially Correlated Disturbances with Orthogonal Regressors (Simulated Data)

(True Model: $y_{n,t} = .2*\mathbf{s_t w_n} + .8*x_{n,t} + h_t + u_{n,t}$)

Variable	Analysis I	Analysis II	Analysis III	Analysis IV	Analysis V
X	0.62645**	0.73229**	0.73618**	0.78320**	0.75976**
	(.09665)	(.04264)	(.04216)	(.04682)	(.04361)
$\bar{y}_{\sim n,t}$		0.99918**			
		(.02104)			
SW				0.38972**	0.08439**
(Spatial Lag)				(.01038)	(.02622)
Fixed Period Effects	No	No	Yes	No	Yes
No. of Observations	600	600	600	600	600

**p-value < .01, *p-value < .05

reports the results of this regression. The estimated standard error on $\hat{\beta}$, which is biased, is almost 0.10.

We consider two simple ways to model the spatial correlation. One approach includes the average value of the dependent variable in other cross-sectional unit, $\bar{y}_{\sim n,t}$, on the right-hand-side. In this context, $\bar{y}_{\sim n,t}$ essentially proxies for $?_t$ and $\mathbf{s_t w_n}$. The other adds period dummies. Severe multicolinearity will usually prohibit estimating a regression model with both $\bar{y}_{\sim n,t}$ and period dummies on the right-hand-side, which underscores further the importance of determining conditions under which one outperforms the other. With **X** orthogonal to both the spatial lag and common shocks, we would expect either strategy to estimate $\hat{\beta}$ without bias; which strategy will produce more-efficient $\hat{\beta}$ estimates is unclear *a priori*. This will presumably depend on the share of **Y**'s total variation that the common shocks can explain relative to the number of time periods in the data set. If that ratio is small, the efficiency costs from lost degrees of freedom will outweigh any explanatory power advantage, so $\bar{y}_{\sim n,t}$ will dominate the set of period dummies, and *vice versa*.

In the one-shot simulation, the time-period dummies (Analysis III) produce a slightly more precise $\hat{\beta}$ estimate (Table 1, Analysis II *vs.* Analysis III), but this will not always hold. Both approaches dominate the simple regression of **Y** on **X**, however, cutting the standard error by more than half. Note that we have not yet explored how accurately this nominal standard-error reduction reflects any true increased precision, as that would require Monte Carlo experimentation. Note also that one can interpret the estimated coefficient on $\bar{y}_{\sim n,t}$ in Analysis II as an estimate neither of *?* nor of some average common shock. The estimated coefficients on the period dummies likewise lack simple interpretation.

Estimating the Coefficient on the Spatial Lag (Analyses IV-V)

Seeking an unbiased and efficient estimate of the spatial-lag coefficient, *?*, as a substantive aspect of the diffusion process complicates matters. Methodologically, the problem surrounds the difficulty distinguishing spatial diffusion from common shocks, especially in small samples, and these difficulties arise regardless of whether analysts seek to make this distinction in their models.

We assume for now that the analyst knows and so need not estimate the spatial diffusion matrix, **W**, allowing us to isolate the problem of distinguishing common shocks from spatial diffusion. Table 1, Analyses IV-V reports the results. If one ignores the common shocks, these omitted fixed period effects will correlate positively with the true spatial lag, biasing the $\hat{\rho}$ estimate upward. In the one-shot simulation (see Table 1, Analysis IV), the estimated spatial-lag coefficient, $\hat{\rho}$, 0.3897, exceeds the true ?, 0.2, by 0.1897, with an estimated standard error of just 0.0104. This *asymptotic bias* will not decrease as T grows large. By contrast, the $\hat{\beta}$ estimate is within one standard error of its true value. Conversely, if one includes fixed period-effects in the model, and *T* is non-infinite, the $\hat{\rho}$ estimate incurs a *spatial Hurwicz bias* downward. This estimator parallels the LSDV estimator discussed extensively in the dynamic panel models literature from econometrics (Hsiao 1986, Baltagi 1995). One can see this in the regression results above (Table 1, Analysis V; Dummy Coefficients Omitted). The estimated $\hat{\rho}$ is 0.1156 lower than the true ?, with an estimated standard error of 0.0262. This *spatial Hurwicz bias* is a small-sample property that decreases with *T*. Again, the $\hat{\beta}$ estimate is within one standard error of its true value (because **X** remains spatially orthogonal).

For analysts who have a substantive interest in spatial relationships, this apparent no-win situation is troubling. How can one obtain good estimates of both *ß* and ?? The dynamic panel models literature has proposed several alternatives to the LSDV estimator: a two-stage and FIML estimator (Anderson and Hsiao 1981), a GMM estimator (Arellano and Bond 1991), and a corrected LSDV estimator (Kiviet 1995). We expect similar approaches to bear fruit in the spatial context and will explore that conjecture in this project.

3. Spatially Correlated Disturbances with Spatially Correlated (Exogenous) Regressors

With spatially correlated disturbances but spatially orthogonal **X**, obtaining efficient and unbiased $\hat{\beta}$ estimates raises no insurmountable challenges. When both regressors and disturbances correlate spatially, however, the difficulties increase appreciably. Formally,

$$\mathbf{var[e]} = s^2\mathbf{C} \quad \textit{and} \quad \mathbf{var[X]} = y^2\mathbf{W} \qquad (6)$$

Now, the variance-covariance matrix **var[X]** contains non-zero elements off the diagonal. The covariance in **X** across units arises from spatial diffusion. Again, we use the model $y_{n,t} = .2*\mathbf{s_t w_n} + .8*x_{n,t} + h_t + u_{n,t}$ to generate a *30x20* dataset and conduct a one-shot simulation exercise. If the analyst does not account for the spatial relationships in any way (Table 2a, Column 1), the $\hat{\beta}$ estimate will be very inefficient (and biased upward, we expect). Note that the estimated standard error on the coefficient, which is biased, exceeds 0.15.

Estimating b with Simple Spatial Proxies (Table 2, Analyses II-III)

Can one improve the efficiency of the $\hat{\beta}$ estimate (*i.e.*, reduce its standard error) simply and innocuously? Unfortunately: no, because either simple proxy, the spatial indicators or the spatial-lag averages, ultimately offers an imperfect substitute for the true diffusion and shock variables. Because **X** correlates spatially, an omitted-variable problem arises from introducing the spatial proxies, which will almost always be endogenous in this case—*i.e.*, correlated with the disturbance. If the spatial-diffusion matrix is symmetric and period effects omitted, the $\hat{\rho}$ estimate of spatial-diffusion will be biased upward because the omitted common shocks will correlate positively with the spatial lag. In the one-shot simulation, the $\hat{\rho}$ estimate exceeds r by 0.6481 (Table 2a, Analysis II). Moreover, the $\hat{\beta}$ estimate understates *ß* by approximately 25% because the spatial lag steals explanatory power from the **X** variable (Achen 2000 discusses a similar effect of temporal lags). Conversely, with the spatial lag omitted and only period dummies included (Table 2a, Analysis III), the omitted lag correlates positively with the included **X,** and the $\hat{\beta}$ estimate will be biased upward. In the simulation, the bias is +0.0721 with an estimated standard error of 0.039.

Getting Good Estimates of Both r *and* b *(Table 2, Analyses IV and V)*

The poor results in Analyses II and III do not derive from the spatial-diffusion matrix being unknown or from the inability of period dummies and dependent-variable averages to enter these models simultaneously. Were **W** known, the analyst could include the true spatial lag in the regression model as we do in Analysis IV. As seen, however, the $\hat{\rho}$ coefficient, even on the true spatial lag, remains overestimated, although the degree of bias decreases radically from Analysis II. Moreover, $\hat{\beta}$ now underestimates *ß* by about 50%. Again, this because the (incorrectly estimated) spatial lag robs some explanatory power from **X** (*a la* Achen 2000).

Table 2a: Spatially Correlated Disturbances and Regressors (Simulated Data)

	(True Model: $y_{n,t} = .2 * \mathbf{s_t w_n} + .8 * x_{n,t} + h_t + u_{n,t}$)				
Variable	Analysis I	Analysis II	Analysis III	Analysis IV	Analysis V
X	0.9436**	0.59042**	0.87214**	0.4402**	0.7643**
	(.15084)	(.03595)	(.03861)	(.05234)	(.03404)
$\overline{y}_{\sim n,t}$		0.84807**			
		(.02545)			
SW				0.3149**	0.1741**
(Spatial Lag)				(.01560)	(.01202)
Fixed Period Effects	No	No	Yes	No	Yes
No. of Observations	600			600	600

**p-value < .01, *p-value < .05

Table 2b: Modeling Unemployment with Spatial and Temporal Lags (OECD, 1966-1990)

Variable	Analysis I	Analysis II	Analysis III
$\text{Unemployment}_{t-1}$	0.9179**	0.0309	0.0562*
(Temporal Lag)	(.01794)	(.02179)	(.02863)
$\text{Unemployment}_{\sim n,t}$		0.9697**	0.9448**
(Spatial Lag)		(.02256)	(.02952)
Fixed Unit Effects	Yes	Yes	Yes
Fixed Period Effects	No	No	Yes
No. of Observations	350	350	350
R^2	0.936	.990	.990

Data Source: Garrett (1998)
**p-value < .01, *p-value < .05

With **W** asymmetric, distinguishing common shocks from spatial diffusion becomes more feasible. More precisely, ability to discern spatial diffusion from spatially correlated responses to omitted spatially correlated shocks increases, we expect, with the difference in their incidence patterns (and may depend, we conjecture, non-monotonically on the number of cross-sectional units). In our one-shot simulations, *e.g.*, the analyst could fruitfully include the true spatial lag (or an appropriately parameterized model thereof) and period dummies in the model both. This spatial analogue to the LSDV estimator for dynamic panel-models reduces the problems in estimating *ß* to those arising in that parallel context. Namely, with both spatial lags and period indicators, the spatial-lag estimates are biased downward in limited samples, but, without period indicators, the spatial-lag estimates are biased upward asymptotically. In our simulation, the *spatial Hurwicz bias* is relatively small (Table 2, Column 3); $\hat{\rho}$ underestimates *r* by 0.0259, with an estimated standard error of 0.0120. $\hat{\beta}$ also underestimates *ß* by approximately one standard deviation. Again, to get both good $\hat{\rho}$ and $\hat{\beta}$ estimates, we expect analysts will have to apply spatial analogues to the more-sophisticated methods mentioned above.

As one final illustration of this *Hurwiczian* dilemma, consider some actual unemployment data for OECD countries between 1966 and 1990. Most analysts agree that unemployment rates persist strongly over time, and none would argue that the unemployment rate at time *t-1* has no effect on the rate at time *t*. Regressing annual unemployment rates on their temporal lag strongly supports this belief (Table 2b, Analysis I). The estimated temporal-lag coefficient is 0.9179, which is likely too large (although some argue strong hysteresis plagues unemployment). Now introduce an asymmetric spatial-lag to the model; *i.e.*, include as a regressor in the model the predicted unemployment rate from an auxiliary regression with the unemployment rates in other countries that year as regressors. With this spatial lag inserted, the coefficient on the temporal lag almost vanishes, dropping to 0.0309, and becomes statistically insignificant (Table 2b, Analysis II). The omission of fixed period effects in Analysis II causes some of this, biasing the spatial-lag estimate upward. Given our asymmetric spatial-lag, the model can accommodate common, fixed period-effects. Doing so, the estimated temporal-lag and spatial-lag coefficients rise and decline, respectively, as expected; although the temporal-lag estimated remains small, it at least regains its statistical significance. Notice from this example and those above that a spatial diffusion process also implies a spatial analogue to the *long-run multiplier* in dynamic models. In a model with a simple one-period temporal lag, a coefficient of *ß* on *X* and of *?* on y_{t-1} implies a long-run effect of a permanent 1-unit increase in *X* of *ß/(1-?)*. The symmetric spatial-diffusion parameter in the model, $\bar{y}_{\sim n,t}$, similarly implies that the post-diffusion total-impact of an exogenous unit-change in *X*, all of which accrues simultaneously given the contemporaneousness of the spatial lag, is *ß/(1-?)*. Thus, in the last two columns of Table 2b, the post-diffusion total-impact of y_{t-1}, is $\hat{\beta}_{y(t-1)} / (1-\hat{\rho}_{\bar{y}(\sim n,t)})$, or 1.0198 and 1.0181 respectively: now likely too large once again (in fact, explosive). Our study seeks to offer some guidance in complicated, but common, contexts such as these with both spatial and temporal dynamics prominent.

4. Spatially Correlated and Endogenous Regressors

Finally, we consider the case of spatially correlated <u>and</u> endogenous regressors. Formally,

$$\mathbf{var[X]} = y^2\mathbf{C} \quad \textit{and} \quad \mathbf{cov[X,?]} = x^2\mathbf{I} \qquad (7)$$

The variance-covariance matrix **var[X]** is an *NxN* matrix with $?^2$ as the diagonal and non-zero elements off-diagonal. The covariance matrix for **var[X,?]** is an *NxN* matrix with $?^2$ as the diagonal and zeros off-diagonal. Here, though, we use the model $y_{n,t} = 0.8 \cdot x_{n,t} + \upsilon_{n,t}$; $x_{n,t} = \eta_t + 0.9 \cdot \upsilon_{n,t}$ to generate the *30x20* dataset in which to conduct our one-shot simulation exercise. The model assumes that the disturbances to $y_{n,t}$ and $x_{n,t}$ correlate, reflecting the endoegneity of the regressors, and the spatial correlation in **X** to be caused by the common shocks. The project will add consideration of spatial diffusion among the **X** also.

Using Spatial Lags as Instruments

With endogenous variables among the right-hand-side terms in a regression equation, analysts might be able to leverage existing spatial interdependence to achieve better estimates of the effects of these variables. In particular, the endogenous variables from *other* cross-sectional units might provide valid instruments for the endogenous variables in each country's right-hand-side. As noted above, political economists and other social scientists occasionally apply just such expedients, and, more frequently, simply assume the exogeneity of *external conditions* when using them directly as controls or substantive right-hand-side variables. Indeed, with spatially orthogonal (independent) disturbances, and with the spatial correlation in **X** attributable to common shocks, the average value of the independent variable in other spatial units each time period works quite well as an instrument because it correlates with the exogenous component of $x_{n,t}$ only. In other words, an *automatically available* and *perfectly exogenous* spatial instrument exists in such cases! Table 3 illustrates, comparing OLS (Analysis I) with 2SLS using $\bar{x}_{\sim n,t}$ to instrument for X (Analysis II). Because $x_{n,t}$ correlates positively with the disturbances, the OLS $\hat{\beta}$ overestimates β substantially. Conversely, the 2SLS $\hat{\beta}$ estimate, which leverages $\bar{x}_{\sim n,t}$ for identification, is unbiased and within one-half a standard error from the true β.

Table 3: Spatially Correlated and Endogenous Regressors (Simulated Data)

	(True Model: $y_{n,t} = 0.8 * x_{n,t} + u_{n,t}$, $x_{n,t} = h_t + .9 * u_{n,t}$)	
Variable	OLS	2SLS (Instrument: $\bar{x}_{\sim n,t}$)
X	1.0833**	0.8168**
	(.01961)	(.03436)
Constant	0.0800*	.06658
	(.03610)	(.04125)
No. of Observations	600	600

**p-value < .01, *p-value < .05

Unfortunately, we expect this simple and readily available instrument far more often to be imperfect, *i.e.*, to produce biased and inconsistent estimates because it too is endogenous, *i.e.*, to provide at best a *quasi-instrument* in Bartels' (1991) terms. The instrument in our example was perfect because the source of spatial correlation in **X**, which is what bought the instrument power in the first stage, was solely a spatially shared and *exogenous* shock, which did not therefore jeopardize the identification leverage of the instrument in the second stage. More commonly, however, **X** will correlate spatially for both common-shock and diffusion reasons, just as does **Y**, especially in cases where one is already entertaining the notion that **X** and **Y** have the kind of conceptual symmetry implied by the definition of endogeneity: **X**? **Y**. In fact, one may well suspect what is sadly usual regarding instruments: that the better $\bar{x}_{\sim n,t}$ predicts $x_{n,t}$, the more likely it is more strongly endogenous. Therefore, here as elsewhere, researchers must face Bartels' dilemma: a tradeoff between bias and efficiency whose terms depend on the ratio of estimable correlation of instruments with that for which they instrument to inestimable correlation of the true stochastic component with the instruments. In some cases the analyst may achieve smaller MSE estimates *via* OLS than with weak or insufficiently exogenous instruments. Recognizing that simple spatial-instruments such as the one suggested above will far more likely be imperfect *quasi-instruments* than perfect ones, our project will seek to characterize the pattern of spatial and other causal arrows, **X**? **Y**. $\mathbf{X}_i$? $\mathbf{X}_j$, and $\mathbf{Y}_i$? $\mathbf{Y}_j$, that determine the terms of Bartel's tradeoff in this case. Finally, once more, analysts who more-directly substantively interested in spatial relationships may prefer more-sophisticated techniques for modeling the potentially complex patterns of spatial dependence and the endogeneity in their data. We will explore several such techniques, including several developed in the endogenous choice / quasi-experimental literature (Heckman 1978, Achen 1986, etc.).

5. Proposed Research

As noted to begin the proposal, this project addresses both social-science researchers directly interested in spatial relationships (*spatial substance*) and those primarily concerned to make optimal inferences regarding other substantive relationships given spatially dependent data (*spatial nuisance*). Our approach, broadly speaking, is to build from analogies to similar, better-explored issues arising from temporal dependence, and through analytic derivation and Monte Carlo experimentation. We will, as noted, tackle six tasks in this way:

(1) detail the conditions under which failing to model spatial dependence or relegating its role to standard-error adjustment biases other coefficient estimates or *merely* induces inefficiency, exploring the magnitudes

of these biases and inefficiencies under differing degrees and natures of spatial dependence;

(2) distinguish spatial diffusion from spatially correlated responses to omitted spatially correlated factors conceptually, and evaluate alternative approaches to making such distinctions empirically;

(3) compare the properties of simple proxies for fuller models of spatial diffusion or spatially correlated omitted-factors—*e.g.*, spatial dummies or symmetric spatial-lags comprised of averages of other cross-section units' dependent variables each time-period—to each other, to PCSE's alone, and to alternative techniques for estimating the full model;

(4) explore the properties of leveraging spatial dependence to aid identification of endogenous systems by using spatial lags as *quasi-instruments* (Bartels 1991) under differing magnitudes of the relevant spatial correlations and causal arrows: **X**? **Y**, $\mathbf{X_i}$? $\mathbf{X_j}$, and $\mathbf{Y_i}$? $\mathbf{Y_j}$.

(5) develop and explore the properties of several parametric, semi-, and non-parametric methods of testing for spatially correlated disturbances, in the presence or absence of spatial lags in the model;

(6) create and publish statistical-software algorithms in widely used software packages, *e.g.*, Stata, to implement the techniques explored and, where potentially useful, pedagogical modules to help teach them.

Tests for and Measures of Spatial Correlation

We have discussed the first four extensively above, and the last is self-explanatory. Regarding tests (*i.e.*, statistical measures) for spatial correlation, notice that so far we have assumed that analysts working with time-series-cross-section data know their data correlates spatially. Such data certainly will typically exhibit spatial correlation, but analysts will still want to gauge the magnitude of (*i.e.*, diagnose) this problem before treating it, especially if the sort of bias-efficiency-complexity tradeoffs we discussed manifest as frequently and prominently as we expect. For example, we expect that, just as including temporally lagged dependent variables as regressors induces bias if residuals retain serial correlation, including spatially lagged dependent variables as regressors induces bias if residuals retain spatial correlation. Accordingly, this project will develop several parametric, semi-parametric, and non-parametric tests for (gauges of) spatial correlation and provide (analytic and) experimental results documenting their (large and) small sample properties. We expect to include among these:

(a) a Lagrange multiplier test, derived by analogy to White's test for heteroscedasticity, which would regress $e_{i,t} \cdot e_{j,t}$ rather than White's $e^2_{i,t}$ on the corresponding products and, as feasible, cross products of $\mathbf{X_{i,t}}$ and $\mathbf{X_{j,t}}$ rather than White's products and, as feasible, cross products of the elements of $\mathbf{X_{i,t}}$. We have confidence in the analogy's soundness: as with heteroskedasticity, the sample pattern of which $e^2_{i,t}$ estimates, the spatial correlation, the sample pattern of which $e_{i,t} \cdot e_{j,t}$ estimates, that most imperils standard regression statistics will have a form whose pattern relates to the corresponding products and cross-products of the independent-variable matrix. Moreover, we expect this test to retain validity with or without spatial lags in the model. However, the very large number of "corresponding products and cross products of $\mathbf{X_{i,t}}$ and $\mathbf{X_{j,t}}$ (*1xK* vectors)" may severely limit feasibility (even ignoring cross-products, at some cost, to limit this concern), and the test statistic's small-sample (large-sample) properties may (will) prove (very) difficult to simulate (to derive).

(b) a Lagrange multiplier test derived by analogy to the standard LM test for serial correlation, which would regress $e_{i,t}$ on $e_{j,t}$, rather than the LM's $e_{i,t}$ on $e_{i,t-1}$, controlling for the right-hand side of the model. This test should retain validity with or without spatial lags in the model, as its temporal analogue does; it should prove more-universally feasible; its large- and small-sample properties should prove easier to derive and to simulate; but we do not expect this LM test to distinguish the more perilous forms of spatial correlation where the spatial-dependence pattern correlates with the moments of the **X** matrix from the lesser forms.

(c) the likelihood-ratio test described in Greene (1997, p. 661), which compares the sum of the logged diagonal elements of the residual variance-covariance matrix under a restriction holding all off-diagonal elements to zero with the log of the unrestricted variance-covariance matrix's determinant, to test the null hypothesis that the unrestricted matrix's off-diagonals are all zero. The determinants of diagonal matrices are simply the products of their diagonal elements, so, if the restricted and unrestricted variance-covariance matrices are equal, the sum of the logged diagonal elements and the log of the determinant will be equal. The sum of the restricted matrix's logged diagonals will therefore exceed the log of the unrestricted matrix's determinant by more the larger are the absolute values of the off-diagonal elements in the unrestricted matrix. Under the null hypothesis, T times this difference is distributed chi-squared with n(n-1)/2 degrees of freedom.

(d) some semi- and non-parametric tests and measures based on groupings of residuals, including one based on the familiar Durbin-Watson statistic, various spatial versions of serial autocorrelation and partial-autocorrelation correlograms, and Ljeung-Box Q-tests. Some of these will not be robust to the presence of spatial lags in the model, and their generally weaker structural assumptions will typically weaken them as tests for specific patterns of spatial dependence, such as the one termed most *perilous* above. However, we expect that, conversely, this characteristically looser structure might broaden the range of patterns that might register with the test or measure.

Adding the development and evaluation of these tests and measures to our other five tasks, our proposal is, in sum, to undertake six research projects and sets of research questions, addressing each (a) by extending analogies to similar, more-studied issues in temporal dependence and by analytic derivation and Monte Carlo experimentation, (b) for each of three alternative types of spatial interdependence—disturbances only; disturbances and right-hand-side variables; and disturbances and right-hand-side variables with endogeneity—and (c) for each of two types of audiences—those for whom spatial dependence is solely nuisance potentially jeopardizing estimates of other substantive quantities that hold their primary interest and those for whom the spatial dependence itself holds more-central substantive interest.

6. Results of Prior Support

The PI has received prior support in connection with NSF award # S137-245-01 (Amount: $174,687; Dates: 1/10/02-4/30-04) for "Empirical Implications of Theoretical Models (EITM) Summer Training (Political Science Program: EITM Competition IIIa." The PI of this proposal is co-PI on NSF award # S137-245-01, which funds a series of four summer training institutes (2002-2004) in EITM, hosted at Harvard, Michigan, Duke, and Berkeley, respectively, for about 25 advanced graduate students and early junior faculty. To be exact, the PI served as one of about fifteen faculty lecturers at the first institute and is serving as the institute director this summer at Michigan. Only in the latter role, representing approximately one-fourth of the total grant and only just begun the end of 2002, does he receive and manage any support directly. Thus, *results of funding* here will refer to the first year of the grant, in which, again, the PI played a supporting but not a directing role.

In the five months from funding to the start of the first institute, roughly fifteen faculty lecturers were recruited, a very large number of students applied from which the PI's selected 25 participants. Instructors lectured, led discussion, and designed projects, all of which produced about five-and-a-half days of full time engagement for the participants for four weeks from mid-June through mid-July. The participant evaluations enthusiastically agreed on the overall quality of the experience, which concluded with a coast-to-coast (participants in Cambridge; several lecturers in Seattle) web-simulcast of student project proposals. All lecture, assignment, and related material has been recorded and will be made publicly available in various forms.

The second summer session of this EITM institute at Michigan is in planning stages, with most logistics arranged, nine of the roughly faculty lecturers having committed, and applications due within one month of this writing.