THE FUTURE OF SPATIAL ECONOMETRICS*

JORIS PINKSE†
CAPCP
DEPARTMENT OF ECONOMICS
THE PENNSYLVANIA STATE UNIVERSITY

MARGARET E. SLADE‡
DEPARTMENT OF ECONOMICS
THE UNIVERSITY OF BRITISH COLUMBIA

June 2009

Prepared for the 50th anniversary symposium of the Journal of Regional Science held at the New York Federal Reserve

Forthcoming in the Journal of Regional Science

The purpose of this paper is threefold. First, we give an overview of the general direction the spatial econometrics literature has taken without attempting to provide a representative survey of all interesting work that has appeared. Second, we identify a number of problems in spatial econometrics that are as yet unresolved. Finally, we provide advocacy for the notion that new spatial econometric theory should be inspired by actual empirical applications as opposed to being directed by what appears to be the most obvious extension of what is currently available.

*Margaret Slade's research is supported by the Social Sciences and Humanities Research Council of Canada. Joris Pinkse thanks the Human Capital Foundation for their support of CAPCP. We thank Gilles Duranton, Christian Redfearn, Giorgio Topa, and an anonymous referee for valuable comments. Joris Pinkse is an extramural fellow of Tilburg University.

†(corresponding author) Center for the study of Auctions, Procurements, and Competition Policy, Department of Economics, The Pennsylvania State University, 608 Kern Graduate Building, University Park 16802, joris@psu.edu

‡m.slade@mac.com

1. INTRODUCTION

Spatial econometrics has been in existence for decades, and the number and diversity of applications has grown at a rapid rate in recent years. These include problems such as price competition among firms located in geographic space (Pinkse, Slade, and Brett, 2002), demand for differentiated products located in product–characteristic space (Pinkse and Slade, 2004), and spillovers among firms whose R&D activities are located in product, technology, and geographic spaces (Lychagin, Pinkse, Slade, and Van Reenen, 2009). Nevertheless, the theory is in many ways in its infancy relative to the complexity of many applications.[1] In this paper, we sketch some of the problems that spatial econometricians face and, in some cases, suggest possible solutions and directions for future research.

To illustrate the problems that confront spatial econometricians, and to emphasize the differences between spatial and time–series analysis, we begin with two economic examples. The first — resource plays in mineral exploration — involves geographic space, whereas the second — estimating the demand for differentiated products — involves product–characteristic space.

Exploration for many minerals can result in rushes or plays in which success breeds increased activity. Gold is the most notable example, but other fuel and nonfuel minerals, such as silver, copper, and petroleum, are subject to exploratory rushes, plays, or bubbles. To make things simple, we use drilling for petroleum as our example. It is clear that a large find in a particular area or basin generates additional exploratory activity. Indeed, a discovery changes the expectations that petroleum geologists have about the profitability of the region. One might, however, be interested in determining whether decision makers overreact, generating bubbles that subsequently collapse, or whether the exploratory effort that is devoted to the area is 'optimal' from a cost/benefit point of view. In an attempt to

[1] This situation is in sharp contrast to time–series econometrics, where the theory is well developed.

answer this question, the econometrician might collect data on drilling activity in a region. The data thus generated would be located in geographic space as well as time. The analysis of such data is not simple. First, the timing and location of drilling is endogenous and not determined by the analyst. In other words, geologists choose where they will drill based on their expectations of profitability. Second, the number of observations is also endogenous. Moreover, the sample and the population can be one and the same. It is therefore obvious that the analyst is not choosing a representative sample from an underlying distribution. Third, drilling locations are not spaced on a grid that overlays the region but are instead apt to cluster in particular areas. None of these problems is apt to surface in a time–series context. Finally, the process that generates the data is dynamic. This feature, unlike the first three, is shared by time–series analysis.

Our second example involves the demand for a differentiated product. Suppose that our data consist of a large number of brands that are sold in different markets, where a market is a location/time pair. The econometrician might be interested, for example, in determining the effect of a merger between two producers of that product. The answer to that question clearly depends on whether the brands of the merging firms are 'close' substitutes. Indeed, if they are (are not) the merger is apt (not apt) to result in increased prices. Space in this example is not geographic. Instead, an observation or brand is located in product–characteristic space. All of the afore–mentioned problems surface here as well. In addition, however, a new problem emerges — it is not clear how to measure 'closeness' in product–characteristic space. To illustrate, if the product is beer, we expect ales to be closer substitutes for one another than for lagers. Furthermore, it is natural to ask if closeness is a zero/one distinction — same or different class — or if it decays gradually with distance. Finally, there are many other characteristics, such as alcohol content, along which brands can be close or distant.

These are challenging questions that the spatial econometrician must answer. Moreover, they illustrate that we need tools that are not simply extensions of familiar time–series techniques to multiple dimensions. Furthermore, it should be obvious that a 'one size fits all' approach is unapt to be fruitful. Instead, the econometrics should be tailored to the economic question and the characteristics of the data.

For a more formal analysis, consider the following spatial problem. In a nutshell, the objective of spatial econometrics is to learn about the nature of a function m_n for which

$$m_n(\boldsymbol{A}) = \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{A}$ is an $n \times d$ matrix whose i–th row contains the available data pertaining to observation i and $\boldsymbol{u}$ is an n–dimensional independent and identically distributed (i.i.d.) vector of errors. For the sake of convenience we will refer to i as a location here, and one can think of (1) as a cross–sectional problem, but i could equally well be a (location, time) pair.[2]

There is no hope of estimating m_n without making simplifying assumptions. Aside from the fact that it is unclear what estimating a function that changes with the sample size would mean, we would essentially be trying to estimate an n–dimensional function with $n \times d$ arguments on the basis of a single draw $\boldsymbol{A}$. We thus need to restrict the function m_n in some manner.

There are many ways of restricting m_n and which restrictions are plausible depends on the nature of the application. A large fraction of the theoretical literature is dedicated to highly parsimonious fully parametric specifications such as the first order spatial autoregressive model (SAR(1)) with regressors

$$\boldsymbol{y} = \psi_0 W \boldsymbol{y} + \boldsymbol{X}\beta_0 + \boldsymbol{u}. \tag{2}$$

[2] It would be more precise to denote the location of observation i by a vector $\boldsymbol{\ell}_i$.

With (2), $\boldsymbol{A} = [\boldsymbol{y}|\boldsymbol{X}]$ and the entire vector $\boldsymbol{u}$ is often assumed to be independent of the entire matrix $\boldsymbol{X}$, the spatial weight matrix W is assumed known, and the errors $\boldsymbol{u}$ are assumed to be i.i.d. normal (or possibly have some simple spatial dependence relationship). The SAR(1) model is taken as an example here, but the criticism below applies equally to other low order spatial ARMA processes, including ones in panel data settings.

It is certainly true that estimation of the unknown coefficients in (2) is both straightforward and efficient, provided of course that the model is correctly specified, the weight matrix W satisfies appropriate regularity conditions, and spatial dependence is sufficiently weak. It is equally true that there are often interesting features to the generally careful, rigorous and sometimes elegant theoretical work in this area; a good example is Bao and Ullah (2007); see Anselin (1988) for a comprehensive but outdated list of work in this area. And yes, simple models like (2) can be the most that some limited data sets will bear. But the most one will get out of the SAR(1) model and its brethren is some idea of the sign and strength of the spatial dependence among the elements of $\boldsymbol{y}$, something that can be discovered equally well, and usually better, with a test of spatial dependence.

The limitations of the SAR(1) model are endless. These include: i) the implausible and unnecessary normality assumption, ii) the fact that if $\boldsymbol{y_i}$ depends on spatially lagged $\boldsymbol{y}$'s, it may also depend on spatially lagged $\boldsymbol{x}$'s, which potentially generates *reflection–problem* endogeneity concerns (Manski, 1993), iii) the fact that the relationship may not be linear, and iv) the rather likely possibility that $\boldsymbol{u}$ and $\boldsymbol{X}$ are dependent because of e.g. endogeneity and/or heteroskedasticity.

Even if one were to leave aside all of these concerns, there remains the laughable notion that one can somehow know the entire spatial dependence structure up to a single unknown multiplicative coefficient ψ_0. The comparison to the stationary time–series case, on which the SAR model is based, does not apply. Indeed, for stationary time series, a low–dimensional parametric formulation is often appropriate. But with spatial data, stationarity is unlikely; data are not equally spaced; missing observations can generate endogeneity; spatial observations are themselves often spatial aggregates; it is unclear whether space grows, the density of observations increases, or both; the dependence structure can change as new data are added; and the very locations can themselves be endogenous.

There is a strand of the literature that removes some of the rough edges of models like the SAR(1) by doing away with the normality assumption (e.g. Kelejian and Prucha, 1999), replacing independence assumptions by conditional moment conditions, allowing for some dependence between $\boldsymbol{u}$ and $\boldsymbol{X}$ and between different elements of $\boldsymbol{u}$ (Brett and Pinkse, 2000), and indeed allowing for nonlinear parametric specifications (e.g. Conley, 1999; Lee, 2007; Pinkse, Slade, and Shen, 2006). Such procedures typically require the estimation of an asymptotic variance using a procedure that accounts for the spatial dependence (e.g. Kelejian and Prucha, 2007; Pinkse, Slade, and Shen, 2006), of which the new and attractive procedure of Bester, Conley, Hansen, and Vogelsang (2009) is both the most ambitious and requires the strongest assumptions. One can even achieve the semiparametric efficiency bound (Robinson, 2009b) and improve the higher order properties of estimators in such models (Iglesias and Phillips, 2008), much like in the case of i.i.d. data, e.g. Robinson (1987) and Newey and Smith (2004), respectively. Robinson (2009a) contains results for nonparametric regression estimation subject to spatial dependence; McMillen (2009) advocates the use of such methods. Some of these procedures rely on asymptotic theory based on the assumption of exogenous locations (e.g. Jenish and Prucha, 2009), others on abstract assumptions about the ability to group data (Pinkse, Shen, and Slade, 2007).

None of the above methods solves the basic problem of having to choose which restrictions to impose on m_n. There is no guarantee, indeed few reassurances, that the restrictions imposed by any of the existing theoretical methods is suitable for a given application.

We believe that the best way of extending spatial econometric theory in empirically relevant directions is not to see how we can create *ad hoc* extensions to existing theory or to simply translate existing time–series methods to the spatial case, but to shape the theory to suit particular classes of applications. Indeed, most of our work has taken this approach (e.g., Pinkse, Slade, and Brett (2002), Pinkse and Slade (2004), and Pinkse, Slade, and Shen (2006)). It is unrealistic to expect to be able to conduct an empirical exercise with spatial data that is beyond criticism. In particular, finding fault with any empirical work, no matter how carefully done, is easy. But letting applications guide the theory does allow one to remove the serious sources of misspecification, especially ones due to endogeneity.

The discussion above, and indeed the rest of the paper, highlights problems arising from the analysis of spatial data. Perhaps it is therefore not surprising that in much applied work the presence of spatial dependence is ignored altogether. But aside from providing a theoretically interesting challenge and being empirically relevant, spatial data are also easier than i.i.d. data in some important respects. The most salient of these is the availability of instruments. Indeed, if a given instrument, say $\boldsymbol{z}_i$, is orthogonal to the error $\boldsymbol{u}_i$ and correlated with $\boldsymbol{x}_i$, then it is often arguably also uncorrelated with error $\boldsymbol{u}_j$ and correlated with $\boldsymbol{x}_j$ at a location j near location i. For example, when modeling the demand for differentiated products, it is common to use the characteristics of rival products j that are 'close' in characteristic space as instruments for the price of product i. This means that, although endogeneity problems are often more severe, we tend to have more instruments at our disposal and thus better methods to deal with those problems.

In what follows we highlight some specific challenges that arise in spatial applications. Many of these are still waiting for good solutions. Where possible, we illustrate problems in the context of a simple linear spatial model, but sometimes we need more complicated models to make our point. The examples are heavily biased towards our own work and, since this is not intended as a survey, we do not come close to citing all interesting articles that have appeared in the spatial literature.

The remainder of the paper is laid out as follows. The next section deals with dependence structure and strength, some related identification issues, aggregation, and distinguishing between dependence and independence. Section 3 discusses the general issue of endogeneity and some of its causes, section 4, which suggests new directions, highlights discrete choice and partial identification, and finally, section 5 concludes.

2. Dependence Structure

2.1. **Linear Spatial Dependence.** The problem that is formalized in equation (1) is essentially one of modeling spatial dependence. Unfortunately, however, modeling the entire dependence structure of a spatial data set accurately is a near impossible task. But suppose that we are willing to assume that the spatial dependence relationship is in fact linear in the sense that we are willing to write something like

$$\boldsymbol{y} = G(\psi_0)\boldsymbol{y} + \boldsymbol{X}\beta_0 + \boldsymbol{u}, \tag{3}$$

where $G(\psi_0)$ is a matrix to be modelled and $\boldsymbol{u}$ satisfies a suitable conditional moment condition. To simplify the discussion we ignore the possibility that $\boldsymbol{y}$ is also spatially dependent on $\boldsymbol{X}$ as well as any endogeneity concerns.

Before discussing equation (3) formally, it is helpful to think about when it might, or might not, be appropriate for particular economic applications. There are natural ways in which such models can arise that we illustrate with an example.

If the economic context is a game among firms, and if their profits are quadratic in their choice variables, (3) can be the set of first–order conditions or reaction functions that arise out of the firms' optimization problems. In particular, under the above assumptions, a single player's profits are maximized, conditional on rival choices, by choosing $\boldsymbol{y}_i$ as a linear function of $\boldsymbol{y}_{-i}$, exogenous observables, and unobservables. Furthermore, although a quadratic specification for profits is not general, it provides a second–order approximation to an arbitrary specification.

More generally, if the objective functions of decision makers can be approximated by quadratic equations, their decision rules will be linear, and if their choices are simultaneous and related, the economic problem will be spatial.

Returning to equation (3), in order to get anywhere, some restrictions must be placed on $G(\psi_0)$. One possibility that we have used is to let G be a matrix with zeroes on the diagonal and whose off–diagonal elements are a function of the distance δ_{ij} between observations i and j in some metric. In other words, the (i, j) element for $i \neq j$ is $g(\delta_{ij}, \psi_0)$.[3] Furthermore, 'distance' can consist of a vector of measures and need not be symmetric in the sense that δ_{ij} and δ_{ji} need not be the same.

There are limitations to restricting G in the way described above. First, it is conceivable that the function g itself depends on n; this problem is comparatively straightforward to address; see Pinkse, Slade, and Brett (2002). More importantly, however, one could imagine that the strength of dependence between observations i and j depends not only on δ_{ij}, but also on the distance between i (or j) and other observations. It may be possible to incorporate some of this by redefining δ_{ij} as in Pinkse, Slade, and Brett (2002).[4]

It is often reasonable, and it can be necessary, to impose some parametric form on g. The SAR(1) model assumes among other things that $g(\delta_{ij}, \psi_0) = \psi_0 w(\delta_{ij})$ for

[3] See, e.g., Pinkse, Slade, and Brett (2002) and Pinkse and Slade (2004).

[4] In that paper, the notion of, for example, sharing a boundary or being the closest neighbor depends on relationships with all other observations.

some known function w. The only situation we can think of in which such an assumption makes some modicum of sense is if δ_{ij} is a binary measure, e.g. whether (1) or not (0) two counties have a common border. Even in that example, however, one can question the relevance of arbitrary administrative decisions pertaining to the allocation of land to counties made a very long time ago to economic dependence relationships today. Furthermore, there are issues relating to aggregation and choice of location that are likely to generate endogeneity problems; see section 3.

An alternative possibility is to allow g to be nonparametric. The most straightforward way to estimate g is to use a *series expansion*

$$g(\delta) = \sum_{\jmath=0}^{\infty} \psi_{0\jmath} \epsilon_\jmath(\delta), \tag{4}$$

where the $\epsilon_\jmath$–functionals are chosen by the econometrician and form a basis for the function space that g belongs to. Substituting (4) into (3) yields

$$\boldsymbol{y} = \sum_{\jmath=0}^{\infty} \psi_{0\jmath} W_\jmath \boldsymbol{y} + \boldsymbol{X}\beta_0 + \boldsymbol{u}.$$

As is typical with series estimation, one estimates only the first $\mathscr{J}_n$ ψ–coefficients, where $\mathscr{J}_n$ increases to infinity with the sample size, but more slowly. See Pinkse, Slade, and Brett (2002) for a set of theoretical results and Pinkse, Slade, and Brett (2002); Pinkse and Slade (2004); Pofahl (2007) for applications.

2.2. **Identification.** Endogeneity, to be discussed in section 3, raises complicated identification problems. Even without endogeneity, however, identification can be a thorny issue in spatial models due to the *reflection problem*; see Manski (1993). The reflection problem arises when a researcher tries to infer whether the average behavior of a group influences the behavior of individuals that belong to that group. This problem is especially problematic in models of *social interactions* (e.g.

Manski, 2000), where, for example, one might ask if the average educational attainment of an ethnic group influences the performance of individuals of that ethnic origin. It also has implications, however, for spatial regression models more generally where, for example, one might ask if the prices posted by individual sellers in a local market are influenced by the average price in that market.

Treating location as random, the argument in Manski (1993) in the current context is essentially that in an SAR(1) model for observation i we have

$$y_i = \psi_0 \sum_{j \neq i} w_{ij} y_j + x_i' \beta_0 + u_i, \quad i = 1, \ldots, n, \tag{5}$$

that $\sum_{j \neq i} w_{ij} y_j$ resembles a nonparametric estimate of $\mathbb{E}[y_i | \ell_i]$, and that the coefficients in

$$y_i = \psi_0 \mathbb{E}[y_i | \ell_i] + x_i' \beta_0 + u_i, \tag{6}$$

are not identified if $\mathbb{E}[y_i | \ell_i]$ and x_i are collinear.

The situation is even more problematic if the x's are also spatially lagged, leading to something like

$$y_i = \psi_0 \mathbb{E}[y_i | \ell_i] + x_i' \beta_0 + \mathbb{E}[x_i' | \ell_i] \gamma_0 + u_i. \tag{7}$$

Assuming $\mathbb{E}[u_i | \ell_i] = 0$ a.s., it follows from (7) that

$$\mathbb{E}[y_i | \ell_i] = \mathbb{E}[x_i' | \ell_i] \frac{\beta_0 + \gamma_0}{1 - \psi_0},$$

which in turn implies that the regressors in (7) are collinear.

The reflection problem is important, but there are several issues that mitigate the problem in a typical spatial application. First, the model of interest in spatial econometrics is typically not (6) but (5), i.e. $\sum_{j \neq i} w_{ij} y_j$ is not an estimate of $\mathbb{E}[y_i | \ell_i]$ but the actual intended regressor. This distinction is important because ψ_0 and β_0 in (5) are identified if

$$\mathbb{E}[y | X, \ell] = (I - \psi_0 W)^{-1} X \beta \text{ a.s.} \Leftrightarrow (\beta, \psi) = (\beta_0, \psi_0). \tag{8}$$

Absent further assumptions (e.g. about the dependence structure of $\boldsymbol{u}$), whether or not β_0, ψ_0 are identified depends on the sample size n.[5] Nonidentification can occur but is unlikely in most applications.

A more likely and interesting possibility is that of *weak identification* (Staiger and Stock, 1997), a situation in which identification strength deteriorates with the sample size to preclude consistent estimation. To see this, consider a contrived example that has the off–diagonal elements of W equal to $1/(n-1)$, i.e. $W = (\iota\iota' - I)/n$ where ι is a vector of ones. Then some minor mathematical manipulations yield

$$\mathbb{E}[\boldsymbol{y_i}|\boldsymbol{X}, \boldsymbol{\ell}] = \boldsymbol{x_i'}\beta_0 + \frac{\psi_0}{1+\psi_0}\bar{\boldsymbol{x}}'_{-\boldsymbol{i}}\beta_0 \text{ a.s.,} \tag{9}$$

where $\bar{\boldsymbol{x}}_{-\boldsymbol{i}}$ is the sample mean of the $\boldsymbol{x_j}$'s, excluding $\boldsymbol{x_i}$ itself. If the slope coefficients in β_0 are nonzero and there is variation in $\boldsymbol{x_i}$ across observations, both β_0 and ψ_0 are identified in any sample of finite size. In the limit, however, the right hand side in (9) becomes $\boldsymbol{x_i'}\beta_0 + \psi_0\mu_x'\beta_0/(1+\psi_0)$, such that neither the intercept coefficient nor ψ_0 is identified. With spatially lagged regressors, more serious examples arise.

This is the only context that we are aware of in which weak identification is not just an artificial theoretical construct but can in fact occur in practice. Unfortunately, we are not aware of any work on weak identification for spatial data.

2.3. **Dependence Strength.** A secondary problem is that of the strength of spatial dependence. In a time series one can have e.g. a unit (or greater) root without much consequence; the series simplify diverges. However, due to the 'feedback' with spatial data (dependence is multidirectional), too much dependence can cause problems. Indeed, it can lead to self–contradictory or unstable models. To illustrate, consider a game in which $\boldsymbol{y_i}$ is a strategic choice and the spatial model is interpreted as a vector of first–order conditions or reaction functions that can be solved to find the equilibrium of that game, see section 2.1. In this context, with too

[5]For instance, if W is nonrandom then identification depends on the rank of $(I - \psi_0 W)$ which varies with the sample size.

much dependence there might be no equilibrium, an equilibrium might exist but not be unique, or the addition of new observations might cause the equilibrium to change radically.

In SAR(1) models the typical assumption is that the weight matrix W has eigenvalues not exceeding one (often imposed by *row standardization*) and requiring ψ_0 to be less than one in absolute value. These conditions are sufficient but not always necessary.

In the more general model (3), the situation is more complex. Among other things, strength depends on the number of observations for which $g(\delta_{ij})$ is nonzero, the dimension of the space, and whether the space grows (*increasing domain asymptotics*) or only becomes more densely populated (*infill asymptotics*). With increasing domain asymptotics, having an exponentially decreasing g–function with suitably bounded maximum (as in Lychagin, Pinkse, Slade, and Van Reenen (2009)) usually suffices. Alternatively, having no more than a fixed number of elements in any row of G nonzero and the g–elements (strictly) bounded by one over that number likewise suffices.

2.4. **Aggregation.** With the game theoretic example, decision makers are individual firms (or the managers of those firms). However, in many applications, units of observation are aggregates such as industries. It is therefore natural to investigate the circumstances under which we can we treat such aggregates as decision makers. If the industry is competitive and there are no constraints on choices (e.g., no capacity constraints), a consistent aggregate exists and a collection of firms can be treated like a single decision maker (see e.g. Bliss, 1975). However, when complications such as imperfect competition or quasi–fixed factors are introduced, this is no longer the case. In particular, except under very special circumstances, an aggregate profit function does not exist and estimates of the aggregate coefficients imply nothing about the individual relationships.

In earlier work (Pinkse and Slade, 2004), we dealt with aggregation over consumers in the context of the British beer market by assuming a functional form for demand for which aggregation does not depend on the distribution of consumer heterogeneity or of income.[6] As we discuss there, however, the simplifying assumptions that must be made for this approach to be valid are not always acceptable. Moreover, although similar assumptions can be used in other applications, this is not a 'one size fits all' type of problem; plausible assumptions are generally determined by the nature of the application.

2.5. **Estimation versus Testing.** Many of the above caveats only apply to estimation. For testing, especially for testing a null hypothesis of independence against an alternative of spatial dependence, a complete and correct specification of the spatial relationship is not generally necessary.

It is true that a correct specification yields a powerful consistent test, but even tests against misspecified alternatives generally pick up some of the spatial dependence, albeit with a possibly significant loss of power. An alternative to such parametric tests (e.g. Baltagi, Song, and Koh, 2003; Kelejian and Prucha, 2001; Pinkse, 1999; Robinson, 2008, 2009c) are fully nonparametric tests (e.g. Brett and Pinkse, 1997) which are consistent but have less power than parametric tests for which the dependence structure under the alternative is correctly specified.

3. ENDOGENEITY

3.1. **General Comments.** In all models with spatially lagged dependent variables, including (3), endogeneity is implicit in the model. Such endogeneity issues can be readily addressed by using GMM with one of the consistent covariance matrix estimators mentioned in the introduction. Furthermore, with (2) a natural vector

[6] Note that, since consumers face budget constraints, there are no simple aggregation results for consumers comparable to those for unconstrained competitive firms. The restrictions that must be satisfied for consistent aggregation over consumers or constrained competitive firms can be found in Gorman (1953).

of instruments for $\sum_{j\neq i} w_{ij}\boldsymbol{y_j}$ is $\sum_{j\neq i} w_{ij}\boldsymbol{x_j}$. In the more general model (3), finding good instruments is only marginally more complicated.

Like other models, spatial regression models can feature additional endogenous regressors for the usual amalgam of reasons. Moreover, aside from the inclusion of spatially lagged dependent variables and the aggregation issue mentioned in section 2.4, there are other potentially serious sources of endogeneity. Two such reasons are discussed below.

3.2. **Missing Data.** If the true model is (3), but some data are missing, we have a problem since we cannot construct the $G(\psi)\boldsymbol{y}$–term for most values of ψ. Indeed, in the SAR(1) model, we could only construct this term for the trivial case in which $\psi = 0$.

There is not much work offering a serious solution to this problem. Lee (2007) has shown that if data are missing for exogenous reasons in the SAR(1) model, then the problem can be solved by using two stage least squares. What to do in more general models and especially if data are missing for endogenous reasons (e.g. resulting from an unwillingness to release unfavorable information) is largely an open question.

3.3. **Choice of Location.** The trickiest, most interesting, hardest to solve, most ignored, and arguably most important cause of endogeneity in spatial regression models, however, is that of the endogeneity arising from choices of location.

The most intuitive example is the case in which the unit of observation is a product and space is product–characteristic space. Presumably a firm chooses product characteristics to maximize profit. Hence location is endogenous and consequently so are all distances. This is problematic since it can be difficult to instrument for distances; see Pinkse, Slade, and Brett (2002) for the only attempt that we know of to do so. Alternatively, one can argue that product characteristics are difficult to change compared with e.g. prices, making locations 'relatively' exogenous (see Pinkse and Slade (2004)).

Although the product space example is the most intuitive, as we discuss in the introduction, the endogeneity of location problem arises equally in geographic space. Economists have studied the location choices of individuals (e.g. Kennan and Walker (2009)) and of firms (e.g., Ellison and Glaeser (1997)), but generally treat the characteristics of locales as given. The purpose of much spatial work, however, is to uncover the interaction among (authorities of) geographic units, who choose e.g. tax rates to attract firms or social services to attract households (Brett and Pinkse, 2000). An ideal model would marry the two; it would provide a model explaining both individuals' location decisions and the actions of, say, local authorities.

Many generic large sample results treat locations as both exogenous and fixed and assume that they are observations at particular locations of an underlying spatial process. This makes little sense in many economic applications. Allowing the characteristics to vary with the sample size (as in Jenish and Prucha (2009)) is a start, but is insufficient. Indeed, such results do not accommodate endogeneity of locations including the possibility that products are taken off the market or that their characteristics are changed in response to the introduction of new products.

Our preference is to make explicit, possibly strong, assumptions about the economic relationships that suit one's application and then to match those assumptions to an abstract generic limit result such as is done in Pinkse, Shen, and Slade (2007). This can admittedly be challenging.

4. New Directions

4.1. **Discrete Choice.** Nonlinearities in the spatial dependence structure are treacherous in general, but this is particularly true if the dependent variable y_i is *discrete*, as is the case, for example, when a firm chooses from a limited set of standardized contracts. Often this choice is binary, e.g., enter or do not enter a market (for a discussion of the binary discrete–choice spatial problem, see McMillen (1992)). Even if all regressors are exogenous and spatial dependence is only present in the error terms, the spatial dependence structure can lead to heteroskedasticity, which causes standard probit estimates to be inconsistent (see Pinkse and Slade (1998)).

If some of the regressors are endogenous but continuously distributed, it may be possible to resolve the endogeneity problem along the lines of Rivers and Vuong (1988). But if spatially lagged $\boldsymbol{y}$ belong in the linear spatial model (3), then presumably there are circumstances in which this would be equally true in a spatial regression model with binary dependent variable, such as the spatial probit model. This problem arises, for example, when firm i's preferred contract type depends on the contract choices that are made by rival firms j. It also arises when n firms simultaneously decide whether to enter a new market. In that setting, the profit that each firm derives from entering the market depends on how many and which of its competitors decides to enter, and even with two players there can be multiple equilibria depending on covariate values (see e.g. Tamer (2003) and Xu (2009)).

There is now a large literature on the estimation of coefficients in discrete game–theoretic models where the same small number of players play the same game in a large number of different markets; see e.g. Bresnahan and Reiss (1991). We are unaware, however, of any work on models with multiple equilibria in which the number of markets is fixed but the number of players is allowed to grow.

The increasing number of players case is a very different, and we think that it is both more interesting and much more challenging from an econometrics perspective than the increasing number of markets case, even in a static environment. The dynamic case, which would involve panel data, is an order of magnitude more complicated. There is a substantial literature on spatial panel data models (e.g. Anselin, Gallo, and Jayet, 2008; Baltagi, Song, and Koh, 2003; Druska and Horrace, 2004; Kapoor, Kelejian, and Prucha, 2007), both for fixed and for random effects. Extensions of dynamic panel data models from the case with independence in the cross–sections dimension (e.g. the work cited in Arellano and Honoré, 2001; Arellano and Hahn, 2006) would also be of theoretical interest. However, such techniques are descriptive and cannot be used to compute equilibria such as steady–state distributions of firms. The importance of a structural approach is eloquently expressed by Holmes (2009) in this issue. Alternatively, one could use calibration, which has the unattractive feature that results from calibration studies are essentially untestable

In short, there are many problems here which are both interesting from an econometric theory perspective and are important for empirical work.

4.2. **Partial Identification.** One of the main areas of current interest in econometrics is that of partial identification, in which the vector of parameters of interest is not 'point–identified,' but in which we can only identify a set that it belongs to, see Rosen (2007, 2008) for a game–theoretic example. Such models can be challenging to estimate and the econometric theory justifying them can be complicated.

Having theoretical results that allow partial identification methods to be used in a spatial context would be helpful, because many relationships in game–theoretic models can be expressed as inequalities rather than equalities of moments. Since many spatial models can be thought of as games, such theoretical results would be especially welcome.

5. Conclusion

As is evident from the preceeding text, we believe that the best way to generate the most valuable new methodology in spatial econometrics is to start from concrete empirical problems. We have highlighted several important and interesting areas of spatial econometrics that have not yet been satisfactorily addressed, including the possibility of weak identification, the treatment of spatial models as games with e.g. the potential for a multiplicity of equilibria, the possibilities of missing data and endogenous locations, and the potential problem that the parameter of interest is only *set–identified*.

REFERENCES CITED

ANSELIN, L. (1988): *Spatial econometrics*. Kluwer Academic Publishers Dordrecht.

ANSELIN, L., J. L. GALLO, AND H. JAYET (2008): *Spatial panel econometrics*. Springer.

ARELLANO, M., AND J. HAHN (2006): "A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects," Discussion paper, CEMFI.

ARELLANO, M., AND B. HONORÉ (2001): "Panel data models: some recent developments," in *Handbook of econometrics*, vol. 5, pp. 3229–3296. North-Holland.

BALTAGI, B., S. SONG, AND W. KOH (2003): "Testing panel data regression models with spatial error correlation," *Journal of Econometrics*, 117(1), 123–150.

BAO, Y., AND A. ULLAH (2007): "Finite sample properties of maximum likelihood estimator in spatial models," *Journal of Econometrics*, 137(2), 396–413.

BESTER, C., T. CONLEY, C. HANSEN, AND T. VOGELSANG (2009): "Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators," Discussion paper, University of Chicago.

BLISS, C. (1975): *Capital theory and the distribution of income*. North-Holland.

BRESNAHAN, T., AND P. REISS (1991): "Entry and competition in concentrated markets," *Journal of Political Economy*, 99(5), 977–1009.

BRETT, C., AND J. PINKSE (1997): "Those taxes are all over the map! A test for spatial independence of municipal tax rates in British Columbia," *International Regional Science Review*, 20(1-2), 131.

——— (2000): "The determinants of municipal tax rates in British Columbia," *Canadian Journal of Economics*, 33(3), 695–714.

CONLEY, T. (1999): "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 92(1), 1–45.

DRUSKA, V., AND W. HORRACE (2004): "Generalized moments estimation for spatial panel data: Indonesian rice farming," *American Journal of Agricultural Economics*, 86(1), 185–198.

ELLISON, G., AND E. GLAESER (1997): "Geographic concentration in US manufacturing industries: A dartboard approach," *Journal of Political Economy*, 105(5), 889–927.

GORMAN, W. (1953): "Community preference fields," *Econometrica*, 21(1), 63–80.

HOLMES, T. J. (2009): "Structural, experimentalist, and descriptive approaches to empirical work in regional economics," *Journal of Regional Science*, 50th anniversary edition (forthcoming).

IGLESIAS, E., AND G. PHILLIPS (2008): "Asymptotic bias of GMM and GEL under possible nonstationary spatial dependence," *Economics Letters*, 99(2), 393–397.

JENISH, N., AND I. PRUCHA (2009): "Central limit theorems and uniform laws of large numbers for arrays of random fields," *forthcoming in the Journal of Econometrics*.

KAPOOR, M., H. KELEJIAN, AND I. PRUCHA (2007): "Panel data models with spatially correlated error components," *Journal of Econometrics*, 140(1), 97–130.

KELEJIAN, H., AND I. PRUCHA (1999): "A generalized moments estimator for the autoregressive parameter in a spatial model," *International Economic Review*, 40(2), 509–533.

——— (2001): "On the asymptotic distribution of the Moran I test statistic with applications," *Journal of Econometrics*, 104(2), 219–257.

——— (2007): "HAC estimation in a spatial framework," *Journal of Econometrics*, 140(1), 131–154.

KENNAN, J., AND J. WALKER (2009): "The effect of expected income on individual migration decisions," Discussion paper, University of Wisconsin.

LEE, L. (2007): "GMM and 2SLS estimation of mixed regressive, spatial autoregressive models," *Journal of Econometrics*, 137(2), 489–514.

LYCHAGIN, S., J. PINKSE, M. SLADE, AND J. VAN REENEN (2009): "Spillovers in space," Discussion paper, Pennsylvania State University.

MANSKI, C. (1993): "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 60(3), 531–542.

——— (2000): "Economic analysis of social interactions," *The Journal of Economic Perspectives*, 14(3), 115–136.

MCMILLEN, D. (1992): "Probit with spatial autocorrelation," *Journal of Regional Science*, 32(3), 335–348.

——— (2009): "Spatial data analysis: specification testing with unknown functional form and spatially correlated missing variables," *Journal of Regional Science*, 50th anniversary edition (forthcoming).

NEWEY, W., AND R. SMITH (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72(1), 219–255.

PINKSE, J. (1999): "Asymptotic properties of Moran and related tests and testing for spatial correlation in probit models," Discussion paper, University of British Columbia.

PINKSE, J., L. SHEN, AND M. SLADE (2007): "A central limit theorem for endogenous locations and complex spatial interactions," *Journal of Econometrics*, 140(1), 215–225.

PINKSE, J., AND M. SLADE (1998): "Contracting in space: An application of spatial statistics to discrete-choice models," *Journal of Econometrics*, 85(1), 125–154.

——— (2004): "Mergers, brand competition, and the price of a pint," *European Economic Review*, 48(3), 617–643.

PINKSE, J., M. SLADE, AND C. BRETT (2002): "Spatial price competition: A semiparametric approach," *Econometrica*, 70(3), 1111–1153.

PINKSE, J., M. SLADE, AND L. SHEN (2006): "Dynamic spatial discrete choice using one-step GMM: An application to mine operating decisions," *Spatial Economic Analysis*, 1(1), 53–99.

POFAHL, G. (2007): "Essays on horizontal merger simulation: the curse of dimensionality, retail price discrimination, and supply channel stage-games," *Texas A&M Ph.D. thesis*.

RIVERS, D., AND Q. VUONG (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39(3), 347–66.

ROBINSON, P. (1987): "Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form," *Econometrica*, pp. 875–891.

——— (2008): "Correlation testing in time series, spatial and cross–sectional data," *Journal of Econometrics*, 147(1), 5–16.

——— (2009a): "Asymptotic theory for nonparametric regression with spatial data," *forthcoming in the Journal of Econometrics*.

——— (2009b): "Efficient estimation of the semiparametric spatial autoregressive model," *forthcoming in the Journal of Econometrics*.

——— (2009c): "Large sample inference on spatial dependence," *forthcoming in the Econometrics Journal*.

ROSEN, A. (2007): "Identification and estimation of firms' marginal cost functions with incomplete knowledge of strategic behavior," Discussion paper, CEMMAP.

ROSEN, A. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146(1), 107–117.

STAIGER, D., AND J. STOCK (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65(3), 557–586.

TAMER, E. (2003): "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 70(1), 147–165.

XU, H. (2009): "Estimation of Bayesian Nash equilibria in two–player games with correlated private signals," Discussion paper, Pennsylvania State University.