The Stata Journal (2010)
10, Number 4, pp. 1–21

Making spatial analysis operational: Commands for generating spatial-effect variables in monadic and dyadic data

Eric Neumayer
Department of Geography and Environment
London School of Economics and Political Science
London, UK
Centre for the Study of Civil War
International Peace Research Institute
Oslo, Norway
e.neumayer@lse.ac.uk

Thomas Plümper
Department of Government
University of Essex
Colchester, UK
Centre for the Study of Civil War
International Peace Research Institute
Oslo, Norway
tpluem@essex.ac.uk

Abstract. Spatial dependence exists whenever the expected utility of one unit of analysis is affected by the decisions or behavior made by other units of analysis. If so, spatial dependence is ubiquitous in social relations and interactions. Yet, there are surprisingly few social-science studies accounting for spatial dependence. This holds true for settings in which researchers use monadic data, where the unit of analysis is the individual unit, agent, or actor, and even more true for dyadic data settings, where the unit of analysis is the pair or dyad representing an interaction or a relation between two individual units, agents, or actors. Dyadic data offer more complex ways of modeling spatial-effect variables than do monadic data. The commands described in this article facilitate spatial analysis by providing an easy tool for generating, with one command line, spatial-effect variables for monadic contagion as well as for all possible forms of contagion in dyadic data.

Keywords: st0001, spspc, spundir, spmon, spdir, spagg, spatial dependence, spatial analysis, contagion, spatial lag, spatial error, monadic data, dyadic data

1 Introduction

Do you avoid taking the car during rush hours? If so, you understand the concept of spatial dependence, which in this case means that your choice of a means of transport or the choice of your time of travel is partly a function of other individuals' choices.

More generally, spatial dependence exists whenever the expected utility of one unit of analysis is influenced by the choices of other units of analysis.

Spatial dependence is also of interest to biologists and other natural scientists, but it is for social scientists that its study and analysis is of greatest importance. To state that the social sciences are characterized by interdependence between the various units of analysis and thus by spatial dependence is almost a tautology. Social science is the study of social relations and interactions, so situations in which units are entirely unaffected by what other units do are likely to be rare.

Yet given the nature of its field of study, only a surprisingly small minority of social-science research either actively seeks to analyze spatial dependence or at least to control for its effect. Part of the reason is, of course, that spatial econometrics is still a fairly young subdiscipline (properly starting only with Anselin's [1988] monograph from some twenty years ago) and that it takes time for new methods and advice on specification issues to penetrate mainstream social-science research. Another reason is that many applied researchers may find it difficult, particularly for dyadic data, to create the spatial-effect variables required for modeling spatial dependence. It is here that the commands described in this article facilitate spatial analysis by providing an easy tool for the generation of spatial-effect variables in both monadic and dyadic data.

We start in section 2 by briefly discussing the importance of spatial dependence for the social sciences and contrasting this with the relatively minor role that relevant studies play in published research. In section 3, we provide an overview of the three types of spatial dependence and the appropriate models for analyzing them—namely, spatial lag (spatial autoregressive) models, spatial-x models, and spatial-error models. Whatever the model, spatial-effect variables need to be created.

The degrees of freedom open to researchers in specifying spatial-effect variables differ greatly between monadic and dyadic data. Spatial effects in monadic data (that is, where the unit of analysis is a single unit, actor, or agent) are discussed in section 4. In monadic data, spatial dependence always emanates from other units. A more detailed discussion is given in section 5 for the more complex specification of spatial-effect variables in dyadic data (that is, where the unit of analysis is a dyad or pair representing an interaction or a relation between two units, actors, or agents). Here spatial dependence can emanate from all other dyads, but also from merely one part of other dyads and from either their aggregate behavior relating to all dyads or their specific behavior relating to only the dyad under observation. There are thus many more modeling options available in dyadic data.

In section 6, we describe a technique for generating spatial-effect variables for dyadic data. It allows researchers to work from a standard dyadic dataset, obliterating the need to construct a 4-adic dataset that would connect dyads with dyads. Section 7 provides detailed information on the Stata commands that generate the various spatial-effect variables in monadic and dyadic data.

2 Spatial dependence in the social sciences

Spatial dependence is a common, albeit often neglected part of social interaction. From a theoretical perspective, spatial dependence can result from coercion, competition, externalities, learning, or emulation (Simmons and Elkins 2004; Elkins and Simmons 2005; Franzese and Hays 2010). Units of analysis—call them agents—change their behavior because others pressurize them (Levi-Faur 2005), because they need to find a competitive advantage (Basinger and Hallerberg 2004), because the strategies carried out by other agents affect the payoffs they generate from their own behavior (Genschel and Plümper 1997; Simmons and Elkins 2004; Franzese and Hays 2006; Plümper and Troeger 2008), because agents learn that other strategies proved to be more successful (Mooney 2001; Meseguer 2005), or because they want to mimic the behavior of others (Weyland 2005). As a consequence, all social-science studies in which agents' strategies are partly dependent on the strategies chosen by other agents need to account for spatial dependence.

Existing analyses of spatial dependence are usually motivated by studying one or more of the mechanisms mentioned above that cause dependence among agents. It is important to note, however, that spatial dependence is also likely to exist when researchers do not have a direct theoretical interest in analyzing it. Not controlling for existing spatial effects causes omitted variable bias just as it is caused by the exclusion of any other variable that is correlated with at least one regressor and the dependent variable (Franzese and Hays 2010). Empirical analyses in the social sciences should therefore control for spatial dependence almost as frequently as social scientists nowadays control for temporal dependence—that is, for the impact that the prior behavior of a unit of analysis has on its present behavior.

Surprisingly, however, the number of articles referenced in the *Social Sciences Citation Index* with either the term "spatial analysis" or "spatial dependence" in the title is very small, albeit slightly increasing over time; see figure 1.

(*Continued on next page*)

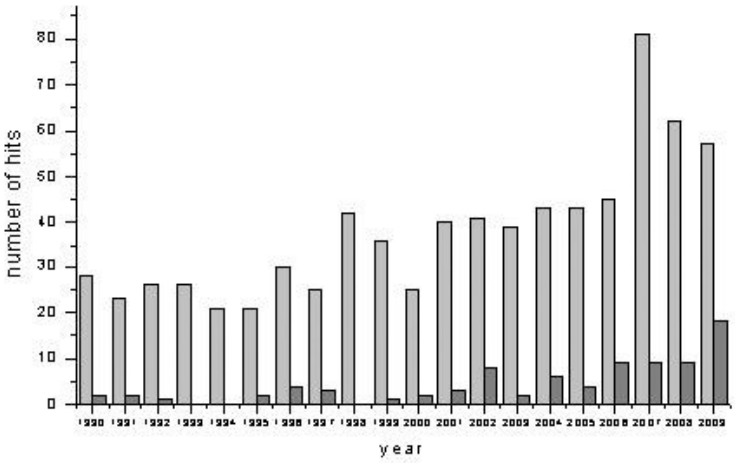


Figure 1. The number of articles in *Social Sciences Citation Index* journals with "spatial analysis" (light gray) or "spatial dependence" (dark gray) in the title in the years 1990–2009

Naturally, there will be many studies that study spatial dependence but do not include either "spatial analysis" or "spatial dependence" in the title. On the other hand, there will be some studies containing either term in the title without actually analyzing or modeling spatial dependence as further defined below. For example, there will be some studies dealing merely with the detection of spatial association and correlation in the data with the help of Moran's I statistic or similar. Such measurement error notwithstanding, the general picture certainly holds true: As yet, spatial analyses are still confined to a small minority. In addition, these spatial analyses are concentrated in only a handful of areas of the social sciences: demography,[1] health science[2] (especially epidemiology[3]), and geographic information system–based research[4] in geography. We also find a few articles in political science[5], political economy,[6] economics,[7] and geography.[8] Spatial analyses may have become more common over the last years, but given the underlying logic of social science, it seems fair to say that they are not yet common enough. The commands presented here facilitate the generation of spatial-effect variables, thus rendering it easier for researchers to study or at least control for spatial dependence.

1. See, for example, Schmertmann, Potter, and Cavenaghi 2008; Chi and Zhu 2008; and Crews and Peralvo 2008.
2. For example, Crighton et al. 2007, and Kandala and Ghilagaber 2006.
3. For example, Atanaka-Santos, Souza-Santos, and Czeresnia 2007.
4. For example, Alix-Garcia 2007, and Gray and Shadbegian 2007.
5. For example, Neumayer and Plümper 2010a.
6. For example, Plümper, Troeger, and Winner 2009; Hays 2009; and Garrett, Wagner, and Wheelock 2005.
7. For example, Kosfeld and Dreger 2006, and Rice, Venables, and Patacchini 2006.
8. For example, Perkins and Neumayer 2010, and Perkins and Neumayer forthcoming.

3 Types of spatial dependence

One can distinguish three types of spatial dependence that call for three types of spatial models. In the first type of spatial dependence, the dependent variable in other units of analysis exerts an influence on the dependent variable in the unit under observation. For example, active labor-market policies in other countries (negatively) influence active labor-market policies in the country under observation because such policies generate positive externalities not captured by the country implementing the policy (Franzese and Hays 2006). The estimation model required to deal with this effect is commonly called a spatial lag model (Franzese and Hays 2007) or a spatial autoregressive model (Anselin 1988). In such models, the spatial-effect variable consists of the weighted values of the dependent variable in other units—that is, on the spatially lagged dependent variable. In scalar notation, the spatial lag model or spatial autoregressive model is formally specified in its simplest form and for monadic data as follows:

$$y_{it} = \alpha + \rho \sum_{k} w_{ikt} y_{kt} + \beta X_{it} + \epsilon_{it} \tag{1}$$

where $i, k = 1, 2, \ldots, N$ denotes the (monadic) unit of observation; $t = 1, 2, \ldots, T$ is time; X_{it} is a set of explanatory variables that may include the temporally lagged dependent variable, unit fixed effects, and period-specific time dummies; and ϵ_{it} is an independent and identically distributed error term. The spatial autoregression parameter ρ gives the impact of the spatial-effect variable, the spatial lag $\sum_{k} w_{ikt} y_{kt}$, on the dependent variable y_{it}. The spatial lag itself is the product of two elements. The first element, an $N \times N \times T$ block-diagonal spatial weighting matrix, measures the relative connectivity between N number of units i and N number of units k in T number of time periods in the off-diagonal cells of the matrix.[9] The second element is an $N \times T$ matrix of the value of the dependent variable.[10]

In the second, rarely analyzed type of spatial dependence, some independent variable of other units affects the dependent variable in the unit under observation. For example, support of terrorist groups by other countries can affect the foreign policy (for example, military spending, alliance formation, and so on) of the country under observation. We call the estimation model required to analyze this type of dependence a spatial-x model. In such models, the spatial-effect variable consists of the (weighted) values of one or more independent explanatory variables in all other units:

$$y_{it} = \alpha + \rho \sum_{k} w_{ikt} x_{kt} + \beta X_{it} + \epsilon_{it}$$

Finally, there is a third type of spatial dependence, in which the error processes are systematically correlated across units of observation. To some extent, this type of dependence will simply be the consequence of failing to adequately model one or both of

9. The diagonal of the matrix has values of zero because $i = k$ and units cannot spatially depend on themselves.
10. The spatial lag could also be temporally lagged.

the other types of dependence: If, say, the dependent variable of other units affects the dependent variable of the unit under observation and this fact is not accounted for, then the error processes will be systematically correlated across units of observation. In fact, researchers will sometimes relegate spatial dependence to the error term for the sake of convenience, despite knowing that the correlated errors are the result of failing to model spatial dependence in the dependent and independent variables. However, there are also factors that can genuinely lead to this third type of spatial dependence. For example, Galton (1889) famously argued that common behavioral patterns across tribes and societies may well be the result of common descent, not of emulation or learning, which would suggest that the spatial correlation in the residuals is best modeled via a spatial-error model.

Spatial-error models account for spatial dependence in the error term, which consists of at least two parts: one is an independent and identically distributed spatially uncorrelated component ϵ_{it}, and the other is a spatial component $\rho \sum_k w_{ikt} u_{kt}$. The model to be fit is thus

$$y_{it} = \alpha + \beta X_{it} + \epsilon_{it} + \rho \sum_k w_{ikt} u_{kt}$$

Not controlling for correlated errors violates the Gauss–Markov assumptions and thus leads to spatial heteroskedasticity. As a consequence, estimates are inconsistent.

The models of spatial dependence can also be combined. Combining the spatial lag model or spatial autoregressive model with the spatial-x model leads to what Anselin (1988, 111) and LeSage and Pace (2009, 32) call a spatial Durbin model. Combining the spatial lag model or spatial autoregressive model with the spatial-error model leads to what Anselin (1988, 36) calls a mixed-regressive spatial autoregressive model with a spatial autoregressive disturbance.

When fitting spatial lag models or spatial autoregressive models, researchers have to deal with an obvious endogeneity problem: When units k affect unit i, the odds are that unit i also affects units k; thus $y_k \to y_i \to y_k \to \cdots$, where the arrows represent an influence.[11] In spatial-x models, endogeneity may also occur if there is feedback from the dependent variable on the spatially lagged independent variable. In this case, $x_j \to y_i \to x_i \to y_j \to x_j \to \cdots$. Franzese and Hays (2007) show that fitting such models with simple ordinary least squares, what they call spatial-ordinary least squares, does not suffer much from simultaneity bias if the strength of interdependence, ρ, remains modest. In all other cases, researchers need to appropriately account for the endogeneity in the variance–covariance matrix. They can do so by either instrumenting the endogenous spatial-effect variable, which Kelejian and Prucha (1998) and Franzese and Hays (2007) call spatial two-stage least squares (2SLS), or by using spatial maximum-likelihood (spatial ML) models. Maximum likelihood models and appropriate software

11. Endogeneity will be absent only if units exclusively depend on other units on which they do not exert an effect in turn, but this constellation is likely to represent the exception rather than the rule. Endogeneity is thus likely to be present in the vast majority of spatial lag models.

exist now for an increasing number of estimators (see, for example, Ward and Gleditsch [2008, appendix A] and LeSage and Pace [2009]).

4 Spatial-effect variables in monadic data

In all three variants of spatial analyses, researchers need to create a spatial-effect variable that consists of the weighted values of the dependent, independent, or error-term variable of other units of observation. Before we come to describe the commands that generate such variables, we first need to explain the multiple forms of modeling this spatial-effect variable. From now on, we will focus on spatial lag models or spatial autoregressive models, because these are very popular in applied research. We will keep in mind that everything we say carries over to spatial-x models and spatial-error models, as well. We start with monadic data, discussing the more complex case of spatial effects in dyadic data in more detail in the next section.

In monadic data, spatial dependence always emanates from all other units, weighted by the connectivity variable. The spatial weighting matrix represents the degree to which unit i is connected to units k, if at all. It can be a dichotomous variable such as geographical contiguity between two units, or it can measure a nonspatial relationship such as trade or investment links (Beck, Gleditsch, and Beardsley 2006). Theory should decide which is the appropriate variable and how exactly it is defined and operationalized. For example, contiguity can be defined in different ways, and trade flows can enter in levels, in logged form, or in other functional forms.

The variable used for the weighting matrix can be undirected as in the case of contiguity or directed as in the case of, say, exports. With directed connectivity variables, researchers must choose whether the weighting matrix measures connectivity from unit i to units k as in (1) above or measures connectivity from units k to unit i as in the following specification:

$$y_{it} = \alpha + \rho \sum_{k} w_{kit} y_{kt} + \beta X_{it} + \epsilon_{it} \tag{2}$$

For, say, exports as the connectivity variable, the weighting matrix in (1) measures exports from i to k, whereas in (2) it measures exports from k to i. Which weighting matrix is appropriate will depend on the specific research context and must also be justified on theoretical grounds.

The weighting matrix is often row-standardized, which means that each cell of the matrix is divided by the row-sum of cells. For example, if the nonstandardized weighting matrix consists of absolute foreign direct investment flows, then the row-standardized weighting matrix consists of shares of foreign direct investment flows. Plümper and Neumayer (2010) argue that researchers must always consider whether row-standardization of the weighting matrix is appropriate for their research design because it changes the substantive meaning of the connectivity variable. One should therefore justify one's decision on theoretical grounds rather than take row-standardization as the unquestioned norm.

5 Spatial-effect variables in dyadic data

In monadic data, spatial dependence weighted by the connectivity variable always emanates from all other units. As discussed above, the only freedoms researchers have lie in the choice of a connectivity variable and its functional form (for example, in levels or in logged form); whether to row-standardize the weighting matrix; and, in case the connectivity variable is directed, the decision whether connectivity should be directed from unit i to units k or the reverse.

In contrast, a dyadic estimation dataset offers more freedoms with respect to the channels through which spatial dependence can be modeled, leading to different types of contagion, and with respect to the specification of the weighting matrix. Most importantly, with dyadic data one can distinguish directed and undirected dyads. In directed dyads, the interaction between two dyad members ij initiates with i and is directed toward j. In the directed dyad ij, unit i is called the source, while unit j is called the target of the interaction. It is different from the directed dyad ji where, in contrast, unit j is the source and unit i is the target.

In contrast, in undirected dyadic data, whilst one can distinguish unit i from unit j, it is either not possible to distinguish between the dyad ij and the dyad ji or researchers do not want to make such a distinction. For example, if the dependent variable measures the presence or absence of militarized conflict between country i and country j, then it may not be clear which country started the conflict or this question may be irrelevant because researchers may merely be interested in whether a conflict exists, not who initiated it. As a consequence, the dependent variables of dyads ij and ji are identical in undirected dyadic data.[12]

Undirected dyadic datasets are most similar to the monadic setting, because spatial dependence always emanates from other dyads. With directed dyadic data, spatial dependence can also emanate from other dyads, but there are more options to be discussed further below. When spatial dependence comes from other dyads, the only choice is whether one wishes to allow dyads that either unit i or unit j form with other units to also exert an influence on the spatial effect variable. For example, a spatial lag model or a spatial autoregressive model of what Neumayer and Plümper (2010b) name inclusive dyad contagion will be specified as follows:

$$y_{it} = \alpha + \rho \sum_{km \neq ij} \omega_{pq} y_{km} + \cdots + \epsilon_{ij} \tag{3}$$

with "$\cdots$" representing other explanatory variables. For ease of exposition, (3) assumes a time-invariant research setting, but a time dimension can be easily added to all variables. Consider military alliances between two countries as an example. Whether country i and country j form an alliance may partly depend on what other alliances

12. Both directed and undirected dyadic data settings are no less likely to be subject to spatial dependence than monadic data settings. What one unit does in relation to another unit with which it forms a dyad will often influence as well as be influenced by the relations of other dyads. Yet in Neumayer and Plümper (2010b), we could identify only three prior studies analyzing spatial effects in a dyadic data setting.

exist between countries in the world, including those that either country i or country j have concluded with countries besides each other. Which other dyads are relevant for the spatial effect—and if so, to what extent—will be specified by the weighting matrix, with

$$\omega_{pq} \in \{w_{ik}, w_{ki}, w_{jm}, w_{mj}, w_{ik+jm}, w_{ki+mj}, w_{ik\times jm}, w_{ki\times mj}\}$$

as eight possible specifications of the (potentially directed) weighting matrix.[13] In words, the weighting matrix can either link (source) units i and k (w_{ik}, w_{ki}) or (target) units j and m (w_{jm}, w_{mj}) or the sum (w_{ik+jm}, w_{ki+mj}) or the product ($w_{ik\times jm}, w_{ki\times mj}$) of the two units.[14]

In contrast, a spatial lag model or a spatial autoregressive model of exclusive dyad contagion disallows all dyads that contain either unit i or unit j from exerting an influence on the spatial-effect variable and is modeled as

$$y_{ij} = \alpha + \rho \sum_{\substack{k\neq i,j \\ m\neq i,j}} \omega_{pq} y_{km} + \cdots + \epsilon_{ij}$$

with the same set of options available for the weighting matrix. On the decision whether to form an alliance between countries i and j, this specification would exclude alliances that countries i and j have with other countries.

As alluded to already, directed dyadic datasets offer more modeling flexibility than just dyadic contagion. The reason is that in such datasets, it is possible to distinguish the source i of a dyadic interaction from its target j. This opens the possibility that spatial dependence only derives from other sources or from other targets, instead of from all other dyads. Moreover, contagion may stem from either the aggregate behavior of other sources or targets or from their specific behavior with respect to the dyad ij under observation.

What Neumayer and Plümper (2010b) coin aggregate source contagion consists of spatial dependence coming from the aggregate behavior of other sources k—that is, from their relationship with any target m, not just the specific target j under observation:

$$y_{ij} = \alpha + \rho \sum_{k\neq i} \sum_{m} \omega_{pq} y_{km} + \cdots + \epsilon_{ij}$$

Our previous example of military alliances could be a directed dyadic relationship if it were possible to distinguish the source (initiator) from the target (recipient) of the interaction, but it is perhaps more likely to be an undirected dyadic relationship. We therefore switch to international terrorism as an example, where the dyadic relationship

13. The list is not exhaustive, and links can be combined with each other (see Neumayer and Plümper [2010b]). Even if the variable that is to be spatially lagged is an undirected dyadic variable, the weighting matrix can still be a directed dyadic variable.

14. For undirected dyad contagion, in which it is not possible to distinguish sources from targets, simply read this sentence, omitting the words source and target. Taking the sum of two weighting matrices implies that they are substitutes for each other (the lack of one link can be compensated by the presence of the other), whereas taking the product implies that they are complements.

between perpetrator and victim is more clearly directed. With aggregate source contagion, the likelihood that terrorists from country i attack victims from country j may partly depend on the aggregate overall propensity of terrorists from other countries k to attack victims from any other country m.

If, instead, only the relationship of other sources k with the specific target j matters for spatial dependence, then the situation calls for modeling specific source contagion:

$$y_{ij} = \alpha + \rho \sum_{k \neq i} \omega_{pq} y_{kj} + \cdots + \epsilon_{ij}$$

With specific source contagion, the aggregate overall propensity of terrorists from other countries k no longer matters for whether terrorists from country i are more likely to attack victims from country j. Instead, only the propensity of terrorists from other countries k to attack victims from this specific country j matters.[15] In both aggregate and specific source contagion, the basic set of link functions is $\omega_{pq} \in \{w_{ik}, w_{ki}\}$; that is, the weighting matrix links source units i and k with each other, either from i to k or from k to i, if it is a directed variable.[16]

The two forms of target contagion function very similarly, only this time it is the aggregate or specific behavior of other targets m from which the spatial effect emanates. For aggregate target contagion, in which the aggregate behavior of other targets m with any source k (not just the specific source i under observation) matters,

$$y_{ij} = \alpha + \rho \sum_{k} \sum_{m \neq j} \omega_{pq} y_{km} + \cdots + \epsilon_{ij} \tag{4}$$

For the example of international terrorism, the propensity of terrorists from country i to attack victims from country j may partly depend on how much terrorism victims from other countries m experience, independently of who the terrorists are.

Specific target contagion, in which only interactions of other targets m with the specific source i matter, is modeled as

$$y_{it} = \alpha + \rho \sum_{m \neq j} \omega_{pq} y_{im} + \cdots + \epsilon_{ij} \tag{5}$$

Here the propensity of terrorists from country i to attack victims from country j partly depends on how much terrorism terrorists from this country i inflict on victims from other countries m. In both forms of target contagion, the set of basic link functions comprises $\omega_{pq} \in \{w_{jm}, w_{mj}\}$; that is, the weighting matrix links target units j and m with each other, either from j to m or from m to j, if it is a directed variable.

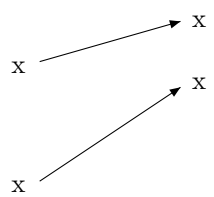
15. For example, Neumayer and Plümper (2010a) use the civilizational affiliation of countries of (potential) terrorists as a connectivity variable to test whether there is evidence for international terrorism spreading along civilizational lines in the form of specific source contagion, as predicted by Huntington (1996).
16. As with dyadic contagion, further link functions are possible. The same applies to the forms of target contagion of (4) and (5).

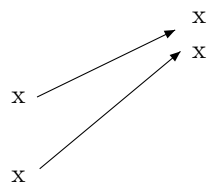
6 Parsing through a virtual 4-adic dataset

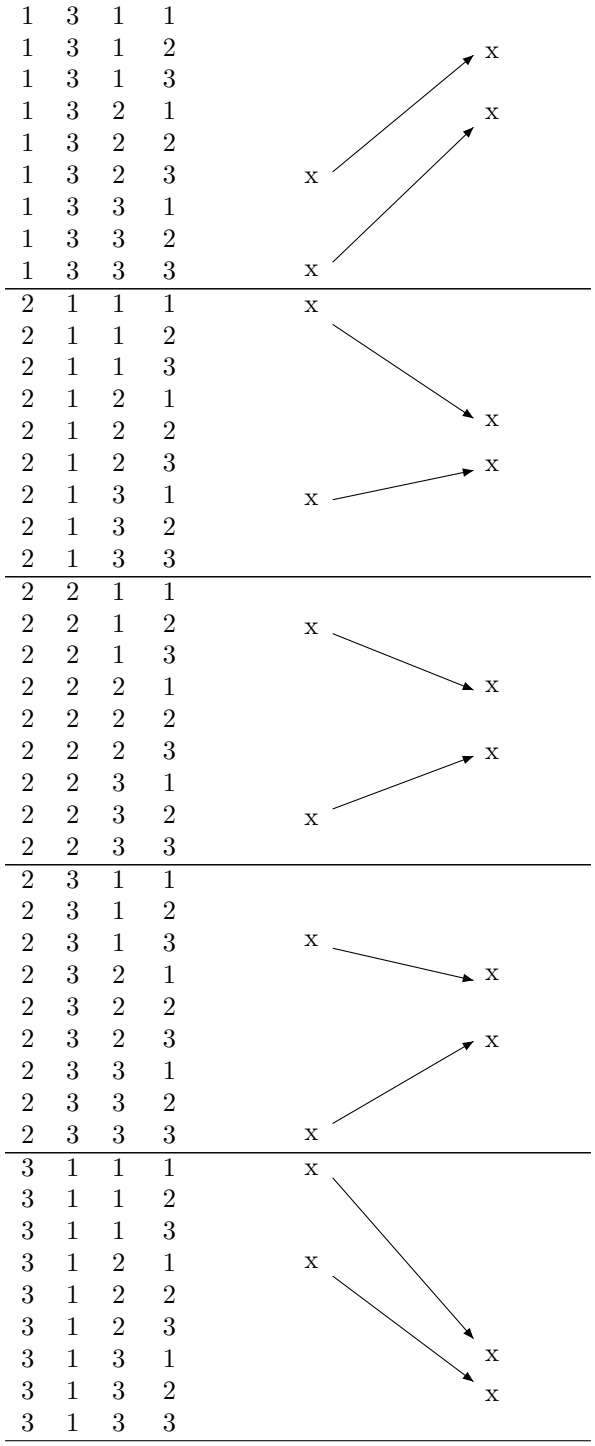
In principle, because the weighting matrix is of one dimension above the dimension of the estimation dataset, one needs a dataset of one dimension higher than the estimation dataset to generate the spatial-effect variable. So, for example, to create a spatial-effect variable for a monadic dataset of dimension $N \times T$, one needs a dataset connecting monadic units with each other—that is, a dyadic dataset of dimension $N \times N \times T$. To generate a spatial-effect variable for dyadic data, one would normally need a so-called 4-adic dataset of dimension $(N_i \times N_j) \times (N_i \times N_j) \times T$—that is, a dataset that connects dyads with dyads. Table 1 displays a very simple directed 4-adic dataset for the case of $N_i, N_j = 3$ with $i, j, k, m \in \{1, 2, 3\}$ and $T = 1$; that is, the dataset is time-invariant.

Table 1. The parsing technique and matching of spatial-effect variable components for specific source contagion and w_{ik} as connectivity

i	j	k	m	Relevant dyads for spatially lagged variable	Relevant dyads for connectivity
1	1	1	1		
1	1	1	2		
1	1	1	3		
1	1	2	1		
1	1	2	2		
1	1	2	3		
1	1	3	1		
1	1	3	2		
1	1	3	3		
1	2	1	1		
1	2	1	2		
1	2	1	3		
1	2	2	1		
1	2	2	2		
1	2	2	3		
1	2	3	1		
1	2	3	2		
1	2	3	3		






1 3 1 1
1 3 1 2
1 3 1 3
1 3 2 1
1 3 2 2
1 3 2 3
1 3 3 1
1 3 3 2
1 3 3 3
2 1 1 1
2 1 1 2
2 1 1 3
2 1 2 1
2 1 2 2
2 1 2 3
2 1 3 1
2 1 3 2
2 1 3 3
2 2 1 1
2 2 1 2
2 2 1 3
2 2 2 1
2 2 2 2
2 2 2 3
2 2 3 1
2 2 3 2
2 2 3 3
2 3 1 1
2 3 1 2
2 3 1 3
2 3 2 1
2 3 2 2
2 3 2 3
2 3 3 1
2 3 3 2
2 3 3 3
3 1 1 1
3 1 1 2
3 1 1 3
3 1 2 1
3 1 2 2
3 1 2 3
3 1 3 1
3 1 3 2
3 1 3 3
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x

3	2	1	1	
3	2	1	2	x
3	2	1	3	
3	2	2	1	
3	2	2	2	x
3	2	2	3	
3	2	3	1	x
3	2	3	2	x
3	2	3	3	
3	3	1	1	
3	3	1	2	
3	3	1	3	x
3	3	2	1	
3	3	2	2	
3	3	2	3	x
3	3	3	1	x
3	3	3	2	x
3	3	3	3	

Note: Arrows indicate which observations from the two separate spatial-effect variable components are merged with each other.

The dataset shown in table 1 is very small, but in many actual research contexts with i and j of medium to large size and multiple time periods, such a 4-adic dataset will be far too large for the memory of standard personal computers (PCs). The commands discussed in this article circumvent this problem by parsing through a virtual 4-adic dataset. Thus, rather than generating an actual full sized 4-adic dataset of dimension $(N_i \times N_j) \times (N_i \times N_j) \times T$, the commands exploit the fact that for any one specific dyad ij, say, dyad 1–1 in table 1, the dyadic dataset of dimension $N_i \times N_j \times T$ highlighted in a light gray color contains both the full set of dyads from which spatial dependence can possibly derive and the full set of dyads that are potentially relevant for the weighting matrix. The commands therefore loop through the full set of dyads, and for any one specific dyad ij, they save in temporary files the dyads km that are relevant for whichever type of contagion is created, as well as another set of dyads km that are relevant for the weighting matrix that is dependent on the type of connectivity chosen. Table 1 shows which dyads km are relevant for the example of specific source contagion and w_{ik} as the chosen connectivity. For the ij dyad of 1–1, the km dyads of 2–1 and 3–1 are relevant for the spatially lagged variable because with specific source contagion it is dyads of the other sources 2 and 3 with the specific target 1 that matter. The km dyads of 1–2 and 1–3 are relevant for measuring connectivity from source 1 to source 2 and from source 1 to source 3.

Once all the necessary components for creating the spatial-effect variable have been saved in temporary files, the commands then combine all the components with each

other by merging the relevant dyads for the spatially lagged variable with the relevant dyads for the connectivity variable to create the spatial-effect variable. The arrows in table 1 show which dyads of the variable that is to be spatially lagged are merged with which dyads of the variable representing connectivity. Finally, unless the `nomerge` option is specified, the resulting spatial-effect variable is created and saved in the current working directory, as well as merged into the original dyadic dataset.

This parsing technique has two main advantages. The one already mentioned is that it makes generating spatial-effect variables possible without having to create a large 4-adic dataset. The second advantage is that users need not worry about creating a connectivity variable that links sources or targets with each other or with additive or multiplicative combinations of the two, depending on the type of connectivity required. By looping through each possible dyad ij and saving only the dyads km in temporary files that are relevant for the specific connectivity variable chosen by the user, all that is required is a connectivity variable that links unit i to unit j. The commands virtually transform this connectivity variable to generate the actual weighting matrix chosen by the user according to the link options available.

Unfortunately, this parsing technique also comes with two disadvantages. First, depending on the size of the dyadic dataset, it can take from seconds to several minutes, hours, or even days to generate the spatial-effect variable on standard PCs. As a general rule, the commands that generate aggregate source or target contagion are fast,[17] the commands that generate specific source or target contagion are considerably slower, and the ones that generate undirected or directed dyad contagion are the slowest. However, creating an actual 4-adic dataset is unlikely to represent a superior alternative to the parsing technique. When its size is moderate enough that so it could be handled by standard PC memory size, the commands employing the parsing technique also work relatively fast. Processing the commands is time-consuming only when creating an actual 4-adic dataset is difficult or impossible.

The second disadvantage of the parsing technique is that because no actual 4-adic weighting matrix is constructed, researchers cannot apply spatial ML methods because doing so would require using the 4-adic weighting matrix. Instead, researchers need to rely on the instrumental-variable technique of spatial-2SLS (Kelejian and Prucha 1998; Franzese and Hays 2007) to account for the simultaneity bias introduced by the spatial-effect variable. Of course, the commands were written specifically for cases in which the 4-adic weighting matrix is simply too large to be handled by standard PCs. In samples for which the 4-adic weighting matrix is not too large, spatial ML can be used—but in such situations researchers do not need the commands described here anyway because they can simply create the spatial-effect variables by hand if they can construct the entire actual 4-adic weighting matrix.

Sparse matrix modeling represents another alternative that is potentially superior to the parsing technique in some contexts (Ward and Gleditsch 2008). However, this technique makes sense only where a large share of zeros (or some other specific con-

17. The same is true for the command that generates spatial effects for monadic data, but `spmon` does not rely on the parsing technique, anyway.

stant number) is in the weighting matrix, as is typically the case for using contiguity or similar as the connectivity variable. If the share of zeros is small—for example when researchers weight by distance, exports, or some other continuous connectivity variable—sparse matrix modeling does not provide any advantage. We contend that connectivity variables with no zeros or a small share of zeros will become much more popular in the future because theories will often predict spatial dependence working via more complicated links than simple dichotomous weights, such as contiguity.

7 Commands for generating spatial-effect variables

7.1 Syntax

Monadic contagion

`spmon` *lagvar* [*if*] [*in*], `i(`*varname*`) k(`*varname*`) weightvar(`*varname*`)`
[`reverse_W` *std_options*]

Undirected dyad contagion

`spundir` *lagvar* [*if*] [*in*], `i(`*varname*`) j(`*varname*`) weightvar(`*varname*`)`
`link(`*link_fcn*`)` [`exclusive` *std_options*]

Directed dyad contagion

`spdir` *lagvar* [*if*] [*in*], `source(`*varname*`) target(`*varname*`)`
`weightvar(`*varname*`) link(`*link_fcn*`)` [`exclusive` *std_options*]

Aggregate source or target contagion

`spagg` *lagvar* [*if*] [*in*], `source(`*varname*`) target(`*varname*`)`
`weightvar(`*varname*`) form(source|target)` [`reverse_W` *std_options*]

Specific source or target contagion

`spspc` *lagvar* [*if*] [*in*], `source(`*varname*`) target(`*varname*`)`
`weightvar(`*varname*`) form(source|target)` [`reverse_W` *std_options*]

For *std_options*, see the *Standard options* subsection in section 7.2.

(*Continued on next page*)

7.2 Description of commands and options

Because for dyadic datasets the parsing technique described in section 6 obliterates the need for a 4-adic dataset, both `spmon`, which generates spatial-effect variables for monadic data, and the set of commands that generate spatial-effect variables for dyadic data (`spagg`, `spspc`, `spdir`, and `spundir`) all merely require a dyadic dataset. This dataset must contain at least four variables.

First, the dataset must contain the variable to be spatially lagged (*lagvar*), the name of which is stated right after each command. For `spmon`, this variable must be the same for all dyads of a specific unit k with various combinations of unit i (for any given time period), whereas for `spagg`, `spspc`, `spdir`, and `spundir`, this variable will typically differ from dyad to dyad. For example, in spatial lag (spatial autoregressive) models, this variable will simply be the dependent variable of other dyads.

Second, the dataset must contain a variable identifying unit i, which is stated in `i(`*varname*`)` in `spmon` and `spundir` and in `source(`*varname*`)` in the commands that generate spatial-effect variables for directed dyadic data. The difference is purely notational. All that matters is that the variable listed in `i(`*varname*`)` or `source(`*varname*`)` identifies unit i. It can be a numeric or string variable.

The third variable must identify a second unit, which is stated in `k(`*varname*`)` for `spmon`, in `j(`*varname*`)` for `spundir`, and in `target(`*varname*`)` for the commands for directed dyadic data. Again the difference is purely notational. What matters is that this numeric or string variable identifies unit k or unit j and that together ik or ij uniquely identify a specific dyad.

The fourth and final variable that a dataset must contain for the commands described here to work is the weighting or connectivity variable, which is always listed in `weightvar(`*varname*`)`. It connects unit i with units k in case of `spmon` and unit i with units j in case of the commands that generate spatial-effect variables in dyadic data. This variable will typically be different for each dyad of a specific unit i with various combinations of units k (j). Also it may or may not be directed. If the spatial-effect variable is to be time-variant, then one additionally needs a fifth (optional) variable in the dataset that identifies time; see `time(`*varname*`)` in the *Standard options* subsection of section 7.2.

For `spagg` and `spspc`, you must also specify whether the spatial effect arises from sources or targets in other directed dyads. Use `form(source)` if the spatial effect stems from other sources, or `form(target)` if the spatial effect derives from other targets.

The commands `spmon`, `spagg`, and `spspc` each allow only two basic link functions such that, for simplicity, the function linking unit i to units k is the default option for `spmon` and for the source contagion forms of `spagg` and `spspc`, while the function linking unit j to units m is the default option for the target contagion forms of `spagg` and `spspc`. In each of these cases, specifying the `reverse_W` option reverses the direction of the connectivity variable such that the weighting matrix represents connectivity from, respectively, units k to unit i and to connectivity from units m to unit j, instead.

Naturally, this option makes sense only if the connectivity variable is in fact a directed variable. Otherwise, both the default and the `reverse_W` option will lead to the same generated spatial-effect variable.

Both `spundir` and `spdir` can create spatial-effect variables with a variety of specified connectivities, such that the required `link(`*link_fcn*`)` option prompts users to choose one of eight possible link functions: `ik`, `ki`, `jm`, `mj`, `ik+jm`, `ki+mj`, `ik*jm`, or `ki*mj`. The `ik` link function requests that the virtually transformed weighting variable listed in `weightvar(`*varname*`)` is to represent connectivity from unit i to other units k. The `ki` link function requests connectivity from other units k to unit i, instead. The `jm` link function requests connectivity from unit j to other units m. The `mj` link function requests connectivity from other units m to unit j, instead. The `ik+jm` link function requests that the virtually transformed weighting variable represents the sum of connectivities invoked by `ik` and `jm`. The `ki+mj` link function does the same, but for the sum of connectivities invoked by `ki` and `mj`. The `ik*jm` option requests that the virtually transformed weighting variable represent the product of connectivities invoked by `ik` and `jm`. The `ki*mj` option does the same, but for the product of connectivities invoked by `ki` and `mj`.

Both `spundir` and `spdir` can also create either inclusive dyad contagion (the default option) or exclusive dyad contagion, to be requested by invoking the `exclusive` option.

Standard options

std_options can be any of the following:

`time(`*varname*`)`	contains the numeric time variable
`sename(`*name*`)`	names the created spatial-effect variable
`labelname(`*label*`)`	names the label given to the spatial-effect variable
`filename(`*filename*`)`	names the file to which the spatial-effect variable is saved
`norowst`	specifies that the spatial-effect variable not be row-standardized
`nomerge`	specifies no automatic merge of spatial-effect variable into the original dataset

All commands allow restricting the relevant sample with `if` and `in` conditions. As mentioned already, there is an optional time-variable identifier, `time(`*varname*`)`, which is needed if the spatial-effect variable is to be time-variant. All commands also allow users to name the created spatial-effect variable by specifying the `sename(`*name*`)` option, to give the created spatial-effect variable a specific label by specifying the `labelname(`*label*`)` option, and to save a dataset containing the generated spatial-effect variable in the current directory under the name specified in the `filename(`*filename*`)` option. Without these options, the generated spatial-effect variable and files are given pre-

defined names.[18] Each command allows deviating from generating a row-standardized spatial-effect variable (the default option) by specifying the `norowst` option. Each command will normally automatically merge the generated spatial-effect variable into the original dataset used for generating it, but this can be prevented by specifying the `nomerge` option. This option is particularly relevant if one uses two separate datasets—one for the creation of the spatial-effect variable and another one that is the actual estimation dataset into which the spatial-effect variable created from the other dataset then needs to be merged by hand. The automatic merge default option is most suitable when the dataset used for the creation of the spatial-effect variable is also the estimation dataset. For the analysis of spatial dependence in monadic data, users must always have two datasets because the estimation dataset is monadic, whereas the dataset used for the creation of the spatial-effect variable must be dyadic.[19]

7.3 A note on the format of the dyadic dataset required for spundir

Often, undirected dyadic datasets are organized such that if dyad ij is contained in the dataset, then dyad ji is excluded, and vice versa. The reason is that one of the dyads contains redundant information given that the value of the dependent variable for ij equals that of ji. If the dataset is in this nonsymmetric format, then it must be the case that the dataset contains only those dyads for which i is numerically smaller or equal to j and excludes all dyads for which i is larger than j, which follows common practice.[20] Thus, for example, if i and j both run from 1 to 4, then the dataset would contain the dyads 1–1, 1–2, 1–3, 1–4, 2–2, 2–3, 2–4, 3–3, 3–4, and 4–4, but it would exclude dyads 2–1, 3–1, 3–2, 4–1, 4–2, and 4–3. (Dyads 1–1, 2–2, 3–3, and 4–4 may also be excluded if a dyadic relationship of a unit with itself is impossible, which depends on the research context, namely, the type of relationship studied.)

It is, however, possible and often convenient for users that an undirected dyadic dataset is organized such that it contains both dyad ij and dyad ji, despite the fact that the value of the dependent variable for these two dyads must be the same. For `spundir` to work, it does not matter whether the dataset is kept in the nonsymmetric or symmetric format. Users must, however, organize their data in symmetric format if the weighting variable is to be directed, because a directed, dyadic weighting variable requires a fully symmetric dyadic dataset.

8 Conclusion

Spatial dependence is a common phenomenon in social relations. Social-science research is therefore particularly in need of modeling or at least controlling for spatial depen-

18. Consult the help files for each command for information on these default names.
19. For `spmon`, if the spatial-effect variable is merged into the original dyadic dataset used for the creation of the variable, then it will have the same value for all units of i in any given time period.
20. If i and j are string variables, then this condition requires that the dataset contain only those dyads for which i is alphabetically prior or equal to j and excludes all dyads for which i is alphabetically subsequent to j.

dence. The rapid improvements in both computing power and spatial estimation techniques as well as mounting advice on specification issues are bound to make more scholars interested in spatial analysis (Anselin 1988; Beck, Gleditsch, and Beardsley 2006; Darmofal 2006; Franzese and Hays 2007, 2010; Ward and Gleditsch 2008; Neumayer and Plümper 2010b; Plümper and Neumayer 2010). The purpose of the commands described here is to render such analysis easier by allowing users the generation of all types of spatial-effect variables with one command line, and in the case of dyadic data, allowing users to do so without the need for constructing a large 4-adic dataset.

9 References

Alix-Garcia, J. 2007. A spatial analysis of common property deforestation. *Journal of Environmental Economics and Management* 53: 141–157.

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

Atanaka-Santos, M., R. Souza-Santos, and D. Czeresnia. 2007. Spatial analysis for stratification of priority malaria control areas, Mato Grosso State, Brazil. *Cadernos de Saúde Pública* 23: 1099–1112.

Basinger, S. J., and M. Hallerberg. 2004. Remodeling the competition for capital: How domestic politics erases the race to the bottom. *American Political Science Review* 98: 261–276.

Beck, N., K. S. Gleditsch, and K. Beardsley. 2006. Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly* 50: 27–44.

Chi, G., and J. Zhu. 2008. Spatial regression models for demographic analysis. *Population Research and Policy Review* 27: 17–42.

Crews, K. A., and M. F. Peralvo. 2008. Segregation and fragmentation: Extending landscape ecology and pattern metrics analysis to spatial demography. *Population Research and Policy Review* 27: 65–88.

Crighton, E. J., S. J. Elliott, R. Moineddin, P. Kanaroglou, and R. Upshur. 2007. A spatial analysis of the determinants of pneumonia and influenza hospitalizations in Ontario (1992–2001). *Social Science and Medicine* 64: 1636–1650.

Darmofal, D. 2006. Spatial econometrics and political science. Paper presented at the Annual Meeting of Southern Political Science Association, Atlanta, GA, January 5–7.

Elkins, Z., and B. A. Simmons. 2005. On waves, clusters, and diffusion: A conceptual framework. *Annals of the American Academy of Political and Social Sciences* 598: 33–51.

Franzese, R. J., Jr., and J. C. Hays. 2006. Strategic interaction among EU governments in active labor market policy-making: Subsidiarity and policy coordination under the European employment strategy. *European Union Politics* 7: 167–189.

———. 2007. Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis* 15: 140–164.

———. 2010. Spatial-Econometric Models of Interdependence. (Book prospectus.) PDF file.

Galton, F. 1889. Discussion of "On a method of investigating the development of institutions; applied to laws of marriage and descent", by Edward B. Tylor. *Journal of the Anthropological Institute of Great Britain and Ireland* 18: 270–272.

Garrett, T. A., G. A. Wagner, and D. C. Wheelock. 2005. A spatial analysis of state banking regulation. *Papers in Regional Science* 84: 575–595.

Genschel, P., and T. Plümper. 1997. Regulatory competition and international co-operation. *Journal of European Public Policy* 4: 626–642.

Gray, W. B., and R. J. Shadbegian. 2007. The environmental performance of polluting plants: A spatial analysis. *Journal of Regional Science* 47: 63–84.

Hays, J. C. 2009. *Globalization and the New Politics of Embedded Liberalism*. Oxford: Oxford University Press.

Huntington, S. 1996. *The Clash of Civilizations and the Remaking of World Order*. New York: Simon & Schuster.

Kandala, N.-B., and G. Ghilagaber. 2006. A geo-additive Bayesian discrete-time survival model and its application to spatial analysis of childhood mortality in Malawi. *Quality and Quantity* 40: 935–957.

Kelejian, H. H., and I. R. Prucha. 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17: 99–221.

Kosfeld, R., and C. Dreger. 2006. Thresholds for employment and unemployment: A spatial analysis of German regional labour markets, 1992–2000. *Papers in Regional Science* 85: 523–542.

LeSage, J., and R. K. Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton: Chapman & Hall/CRC.

Levi-Faur, D. 2005. The global diffusion of regulatory capitalism. *Annals of the American Academy of Political and Social Sciences* 598: 12–32.

Meseguer, C. 2005. Policy learning, policy diffusion and the making of a new order. *Annals of the American Academy of Political and Social Sciences* 598: 67–82.

Mooney, C. Z. 2001. Modeling regional effects on state policy diffusion. *Political Research Quarterly* 54: 103–124.

Neumayer, E., and T. Plümper. 2010a. Galton's problem and the spread of international terrorism along civilizational lines. *Conflict Management and Peace Science* 27: 308–325.

———. 2010b. Spatial effects in dyadic data. *International Organization* 64: 145–166.

Perkins, R., and E. Neumayer. 2010. Geographic variations in the early diffusion of corporate voluntary standards: Comparing ISO14001 and the global compact. *Environment and Planning A* 42: 347–365.

———. Forthcoming. Transnational spatial dependencies in the geography of nonresident patent filings. *Journal of Economic Geography*.

Plümper, T., and E. Neumayer. 2010. Model specification in the analysis of spatial dependence. *European Journal of Political Research* 49: 418–442.

Plümper, T., and V. E. Troeger. 2008. Fear of floating and the external effects of currency unions. *American Journal of Political Science* 52: 656–676.

Plümper, T., V. E. Troeger, and H. Winner. 2009. Why is there no race to the bottom in capital taxation? *International Studies Quarterly* 53: 761–786.

Rice, P., A. J. Venables, and E. Patacchini. 2006. Spatial determinants of productivity: Analysis for the regions of Great Britain. *Regional Science and Urban Economics* 36: 727–752.

Schmertmann, C. P., J. E. Potter, and S. M. Cavenaghi. 2008. Exploratory analysis of spatial patterns in Brazil's fertility transition. *Population Research and Policy Review* 27: 1–15.

Simmons, B. A., and Z. Elkins. 2004. The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science Review* 98: 171–189.

Ward, M. D., and K. S. Gleditsch. 2008. *Spatial Regression Models*. London: Sage.

Weyland, K. G. 2005. Theories of policy diffusion. Lessons from Latin American pension reform. *World Politics* 57: 262–295.

About the authors

Eric Neumayer is a professor and head of department in the Department of Geography and Environment at the London School of Economics and Political Science. He works on international political economy topics.

Thomas Plümper is a professor of government and Director of the Essex Summer School in Social Science Data Analysis at the University of Essex. He works on political economy and methodology.