Spatial Analysis for Political Scientists

How does space matter in our analysis? Is space just about geography? How can we evaluate diffusion or interdependence between units? How biased can be our analysis if we do not consider spatial clustering? All above questions are critical theoretical and empirical issues for political scientists belonging to several subfields from Electoral Studies, passing to Comparative Politics, and also for International Relations. In this special issue on methods, our article introduces political scientists to conceptualizing interdependence between units and how to empirically model these interdependencies using spatial regression. First, the article presents the building blocks of any feature of spatial data (points, polygons, raster) and the task of georeferencing. Second, the article discusses what is an interdependence matrix (*W*), its importance and variations. Third, the article introduces how to investigate spatial clustering trough visualizations (e.g., maps) but also statistical tests (e.g., Moran's Index). Finally, the article explains how to analyse data with geographic interdependences, but also non-geographic spatial interdependencies, using spatial error and spatial lags models. We conclude inviting researchers to carefully consider space in their analysis and reflect on the need (and lack of thereof) for spatial models.

1. Introduction

The metaphor of contagion has been used by social scientists to describe a variety of phenomena that diffuse in space.[1] Technological innovations, policy adoption, norms and ideas, political regimes, conflict, criminal behaviours and political preferences are some of the numerous examples of issues that spread across geographic space and networks. In its conciseness, this list illustrates how spatial interdependence is ubiquitous to social phenomena, but also how spatial diffusion can occur among different subjects, or nodes in a network. Policies diffuse across countries and innovations across firms, norms are spread via international organizations or non-governmental organizations, crime and violence travel across regions, and political opinions spread through personal networks. It is important to point out that diffusion processes are not simple *clustering* of similarity (or dissimilarity) among proximate units; rather, *interdependence* is the key feature of all diffusion processes. As we will see, this is the fundamental difference between spatial autocorrelation and spatial interdependence.

Suppose, for example, that in a survey a group of individuals reports similar levels of religiosity. It turns out that these individuals are friends. We could start wondering whether their religiosity is independent from each others' or there is a peer-group effect at play. One could conclude that the clustering of preference within this group is the result of a diffusion process, where each individual's preference has influenced other friends' preferences, ultimately leading to convergent attitudes toward religion. However, alternative explanations, that *would not* imply interdependent decisions, are equally plausible. For example, individuals can self-select into groups that share similar views *a priori*. Alternatively, friends may share similar views because they had been exposed to similar external stimuli (e.g., they all had religious parents), or they just grew up in the same neighbourhood or village where all had the same religious' attitude. Homophily and common exposure would then explain the observed clustering, rather than any spatial interdependence underpinning diffusion processes.

[1] Diffusion and contagion are often used interchangeably, and we do the same in this article. However, some have used them to refer to slightly different processes. See, for example, Midlarsky et al (1980) and Koktsidis (2014).

But let's suppose what we see is true interdependence and friends' attitudes are simultaneously affecting each other's. Even if we were able to show that the similarity of attitudes is not due merely by clustering but to diffusion, we do not know yet what mechanism lead to diffusion. Hence, we could move this even further and start wondering whether the mechanism behind this diffusion process is one of emulation or learning. Are individuals changing their views in response to each other to intentionally become closer to their friends? Or are they adapting their attitudes because, after observing their friends, they see that religiosity brings some benefits, for instance going to the church every Sunday reinforces ties with community. Finally, and more importantly, the same clustering could be detected among citizens living in the same neighbourhood, beyond the group of friends we initially focused on. All above questions and necessary reflections would apply to this case as well. Notice, however, that one key difference would concern how individuals are connected spatially. In the case of friends, connections are based, by definition, on existing friendship ties. In the case of neighbours, connected individuals are geographically proximate and live in the same neighbourhood, without being necessarily part of the same friends' network.

This introduction illustrates how complex spatial interdependence can be, and why it is not just a simple inferential problem that needs to be fixed in a statistical model. Spatial interdependence offers nuance and insights to the theories we formulate to explain social phenomena. First, not all spatial clustering is spatial interdependence. Spatial interdependence implies that an outcome in a unit directly affects the same outcome in another unit. Second, geographic proximity is only one way to define the web of connections through which diffusion processes unfold. Third, once spatial interdependence is detected, there are different plausible underlying mechanisms that would explain such interdependence, such as spill-over, mimicry or learning.[2] This article cannot comprehensively discuss all three issues, and only focuses on first two. It begins discussing what spatial data is and the broad notion of proximity, which is

[2] We will not cover the specific mechanisms of diffusion, but for an example of learning leading to spatial autocorrelation see Di Salvatore (2018).

used to define connections among units (i.e., the spatial weights matrix). We then discuss spatial autocorrelation and the distinction between clustering and interdependence. We present a step-wise strategy based on commonly used statistical tests (and software)[3] to distinguish between the two different data generating processes. Finally, we present the main modelling strategies in presence of spatial interdependence. Overall this article should be understood as a teaser more than an introduction (Darmofal 2015; Ward and Gleditsch 2018).[4] It puts forward a way to think about spatial interdependence as a norm rather than an exception. This does not mean all our models should also be spatial; on the contrary, it invites a more thorough reflection about the theoretical needs behind spatial models that presume a non-trivial knowledge of the data generating process. In its attempt to provide an overview of types of spatial data, exploratory spatial analysis and some spatial models, this article - in a special issue on methods in Political science- differs from other excellent work on space in social sciences[5] . This article should be understood as a practical guide for political scientists who want to get a sense of where to get started with spatial data, which is the frontier of spatial analysis and the methodological opportunities that it can offer. Hence, this is a short introduction for non (yet) expert in spatial analysis and, we hope, it will provide a steppingstone for future researches based on spatial analysis.

2. What are the core elements for spatial analysis?

Political scientists work with two main data structure, namely cross-sectional and time-series cross-sectional observational data. Hence, our datasets often focus on several units of observation (e.g., countries), which sometimes we can observe at time intervals (e.g., yearly). In most introductory courses to quantitative methods, dealing with temporal dependencies (i.e., serial autocorrelation) is a standard learning

[3] We will refer to ArcGIS, R, and Stata in this article. ArcGIS is a good choice for processing, creating, and mapping spatial data; Stata also allows some simple mapping, but is preferred for spatial statistical analysis. R allows users to perform both data management and data analysis tasks.

[4] Maps, like all other types of data visualization, can be misleading or simply inefficient and uninformative. For a recent guidance on visualization principles and general guidelines (in R), see for example Healy (2018).

[5] See for e.g. Gleditsch and Weidmann (2012) and Harbers and Ingram (2019).

objective. However, rarely the relational nature of the observations across space is addressed, possibly with the exception of considering each unit within a larger spatial unit (e.g., province, regions, continents). In other words, at best, spatial relations and interdependence are model as nuisance and left to corrections of the standard errors (Franzese and Hays 2008). In order to start approaching the possible spatial interdependence in our data, we first need to present the building blocks of spatial data and spatial relationships: first, we discuss different spatial data formats, and second, we explain how to define and declare the spatial interdependence between observations using a so-called spatial matrix (also referred to as W matrix).

Spatial data comes in two main formats: vector and raster data. The key difference between vector and raster data is that the former is used for discrete representations of spatial data (e.g., population in a country), while the latter is used for continuous representations (e.g., terrain elevation or surface temperature). Vector data includes three types of features: points, lines, or polygons. These features are part of so-called shapefile. Notice that a shapefile can only contain one of these three types of feature at a time. While the unit of a vector can be a point, a polygon or a line, in raster data a unit is represented by pixels. As we will discuss later, neither vector features or raster pixels are necessary the observational units of our analysis, but it is necessary to know their existence and differences in order to link this spatial data to our observational unit.

In general, event data usually come as point shapefiles, while information on administrative units (e.g., the population of a district or the GDP of a country) are available as polygon shapefiles. Starting from vectors, points are the most basic form of spatial data, and often these types of data are those we tend to deal the most. Points are defined by pairs of coordinates, x and y, and can represent events (e.g., protests, conflict events), buildings (e.g., military barracks, pooling stations), individuals (e.g., respondents in a survey), towns or any other discrete object defined in space. To provide a more concreate example, if you want to study the local onset of conflict and the presence of natural resources such as diamonds, you will need data points on where the conflicts are (Eck 2012) but also the location of diamonds' mines (Gilmore et al. 2005). Hence these data will come with data points and relative coordinates but also with information, for instance, on the intensity of the conflict (e.g., number of killed soldiers) or types of diamonds in that specific location. Notice that also a temporal information is often attached to spatial data point, especially when we deal with data events (Ruggeri, Gizelis, and Dorussen 2011). Important to note here is that spatial points are often available as list of coordinates (i.e., in Excel formats rather than shapefiles), that can be easily imported in statistical software and transformed into actual spatial data.[6]

Lines and polygons can be thought of as sequences of connected points, where the first point is the same as the last for the polygons, whereas lines are open polygons where the sequence of points does not result in a closed shape with a defined area. Being based on points, lines and polygons also are ultimately based on groups of coordinates. Lines may represent roads (Zhukov 2012), railroads (Ferwerda and Miller 2014; Kocher and Monteiro 2016), rivers (Toset, Gleditsch, and Hegre 2000) or other physical barriers (e.g., checkpoints). Line data will be useful to provide, for instance, information on infrastructures or demarcations among our analytical units. Polygons are a widely used spatial feature as well. Countries in the world can be represented spatially as polygon or, geographic space can be represented with fixed polygon units. For instance, the PRIO-grid project (Tollefsen, Strand, and Buhaug 2012) provides a standardized spatial grid structure with global coverage at a resolution of 0.5 x 0.5 decimal degrees (approximately 50x50 km at the equator).

Finally, the raster data represent continuous surfaces, such as forests or mountains. Raster data are created dividing an entire space into equal-sized cells,

[6] One fundamental point to make in this regard is that *geographical coordinates and projected coordinates are two different things*. In the simplest terms, one's position on a 3D model of the Earth (e.g., a globe) is a geographical coordinate and it usually comes in decimal degrees; one's position on a 2D model of the Earth (e.g. a map) is a projected coordinate and it usually comes in linear units (e.g. feet or kilometers). There are many geographic coordinate systems and many projected coordinate systems. There is not direct conversion from geographical to projected coordinate systems as the move from the 3D to the 2D model of the Earth can be achieved in different ways. It is of utmost importance that researchers make sure that when combining different spatial data these have the same coordinate system, be it geographic or projected. We recommend researchers interested in spatial data to familiarize with the basics of Geographic Information System (GIS). This is essential for those who aim to work with spatial data. For beginners, we strongly recommend Gleditsch and Weidmann (2012).

often represented by pixels, and the value of the variable of interest is measured for each of these cells. An increasingly popular raster data among political scientist is nightlight emissions, gathered via satellite imagery and used as a proxy to measure state reach or local-level economic activity (Harbers 2015; Weidmann and Schutte 2017). Interestingly, it is possible to transform digital images into raster data. Software such as ArcGIS allows to import images and *georeference* them by assigning it geographical information (i.e., linking points in the image to their actual position on the globe). This function is particularly useful to extract information from historical maps that are usually not already available as spatial data formats.[7]

To clarify an issue that often arises among those who are in their first steps of spatial analysis, namely the fact that spatial data points are rarely the unit of observation of our dataset. Similarly, lines and raster data are spatial data but are rarely our unit of analysis. Indeed, among geographers the data points analysis is common (Bivand et al. 2008), but political scientists more often merge information of the data points into their analytical units (e.g. provinces, countries) spatially. For example, it is possible to assign the characteristics of a data point (e.g., a conflict event) to a country in the world if the point is spatially contained in the polygon defining the country.[8] This operation is called spatial merge (or also spatial join). The issue about the match between analytical versus geographic units would deserve a longer discussion (see for example Arjona 2019; Harbers and Ingram 2017), and we cannot fully address it in this article. Ultimately, the answer, is not methodological but theoretical. In most cases, researchers need to combine different types of data and transform it to make it consistent with the most theoretically meaningful unit of analysis. For example,

[7] Notice that georeferencing and geocoding are different tasks. Geocoding involves assigning *coordinates* (e.g. latitude and longitude) to a location (e.g. a village) or events; Costalli and Ruggeri (2015b) provide an example of events' geocoding about violent clashes between partisans and Nazi forces during the Italian civil war. Georeferencing involves assigning *coordinate systems* (e.g. WGS1984) to an image, raster or vector. Event data are usually geocoded; Di Salvatore (2016) provides an example of georeferencing of a census map. She uses digital maps of settlements in Bosnia-Herzegovina to extract information on the ethnic composition of villages where most violence occurred.

[8] It is also important the level of precision of the geocoded event data. If events are geocoded at the country level, this usually implies that countries centroids have been used to assign coordinates. This means such data should not be spatially joined to subnational units. Most geocoded data indicate the level of precision of the assigned coordinates.

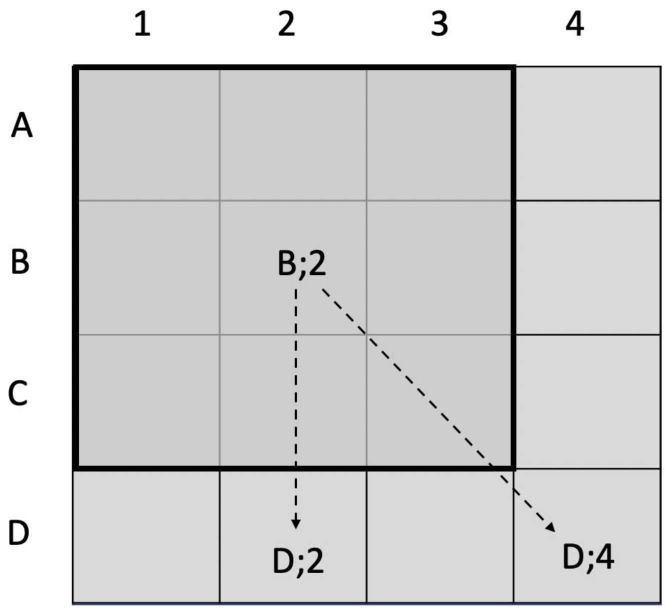
Daxecker, Di Salvatore, and Ruggeri (2019) were interested in individuals' perceptions of fraudulent elections; hence, they used respondents as unit of analysis that were spatially represented as data points within Nigerian states. Then, they linked state's characteristics to each respondent spatially and calculated the distance from each individual to polling stations reporting electoral irregularities. In another study, Costalli and Ruggeri (2015) opted for geographically fixed units (e.g. grid cells) to study the effect of violence on voting preferences because administrative units changed over time and, more importantly, were possible endogenous to the previous regime and conflict. Exactly because grid-cells may be too arbitrary, the authors also evaluated the robustness of their findings using grid-cells with alternative sizes.[9]
We now move on to the discussion of one of the most critical choices we make when we analyse spatial data: which units are close (or connected) to other units?

3. How are things connected?

Once our data contains spatial information, we can start thinking about how units in our data are connected with each other. The final output of this is a spatial connectivity matrix that usually is defined as a weight matrix *W*. The spatial weight matrix *W* is an NxN matrix representing the connections between all units in the data and is used for any type of spatial analysis, from exploratory analysis to statistical modelling and tests of spatial autocorrelation. Anselin et al. (2008, 627) define spatial autocorrelation as being present "whenever correlation across cross-sectional units is non-zero, and the pattern of non-zero correlations follows a certain spatial ordering". Of course, the first intuition about the structure of the spatial connections in our data is geographic and physical space. In fact, as the Tobler's first law of geography suggests, "all places are related but nearby places are more related than distant places" (Tobler 1970, 236). However, we will discuss how cross-units connections can be defined in different ways and, moreover, spatial linkages are not uniquely a geographic matter (Beck, Gleditsch, and Beardsley 2006).

[9] In fact, related to this, another issue that is important to mention, but not fully discuss, is the "modifiable areal unit problem" (MAUP). It is a source of statistical bias that can radically affect the results of statistical hypothesis tests. It affects results when point-based measures of spatial phenomena are aggregated into districts and, therefore, also the size or type of spatial unit can affect the results (Fotheringham and Wong 1991).

Figure 1: Contiguities and Distances



So, how do we define a unit's neighbourhood? How many neighbours do we need to include, and should this be based on distance? Neighbourhoods can be defined in a number of ways, such as contiguity (i.e., units sharing a boundary), distance or on a certain number of closest units (k-nearest neighbours' criteria).

In Figure 1, we show different ways to represent proximity and, therefore, possible channels of spatial interdependence. Let's suppose that we are analysing an area, and this is composed by 16 squares (they could be polygons of different shapes, of course). Each square can be defined by a combination of a letter and a number. Let's focus on square {B; 2}. If we ask to which squares is {B; 2} connected to, our answer depends on how we defined proximity. Suppose we want to use the contiguity criteria. Contiguity-wise, we could think about *first-order proximity*, that is the eight squares immediately surrounding {B; 2} will be its neighbours. This is usually defined as Queen-based contiguity, because in the game of chess the Queen can move in all direction. A rook or bishop-based contiguity criteria will identify only four neighbours, only those squares that share a full side. But we could also decide that we need to account for *second-order proximity* as well. In other words, not just those squares directly at the borders, but also the immediate neighbours' neighbours. In this case {B; 2} would be connected to all squares in Figure 1.

Connectivity matrices' elements are usually binary (two units are either connected or not). However, we can define the connectivity matrix based on distance rather than binary contiguity. For instance, for {B; 2} the squares {D; 2} and {D; 4} if proximity was defined in term of first-order proximity are not connected, but if we consider *second-order* they are equally connected to {B; 2}. However, if we were using as connectivity rule the distance from the centre of {B; 2} to the other squares, then {D; 2} would be closer than {D; 4}. Distance-based connectives can be also transformed into binary values by defining a distance threshold (e.g., 100 km) beyond which units are not considered as neighbouring (Ward and Gleditsch 2018).

Figure 2: Example of binary spatial matrix W.

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	D1	D2	D3	D4
A1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
A2	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
A3	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0
A4	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0
B1	1	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0
B2	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	0
B3	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0
B4	0	0	1	1	0	0	1	0	0	0	1	1	0	0	0	0
C1	0	0	0	0	1	1	0	0	0	1	0	0	1	1	0	0
C2	0	0	0	0	1	1	1	0	1	0	1	0	1	1	1	0
C3	0	0	0	0	0	1	1	1	0	1	0	1	0	1	1	1
C4	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1
D1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
D2	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0
D3	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1
D4	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0

As it is clear, defining a connectivity matrix is not a trivial task and it involves several operational decisions.[10] Once we have decided what is our definition of proximity, how does a *W* matrix look like? Figure 2 provides an example of a connectivity matrix based on our Figure 1. In this case, *W* is defined as a *Queen-based first-order binary* contiguity matrix. As there are 16 units in the data, *W* is a 16x16 matrix based on the possible connection of the 16 units. Since it is binary and based on contiguity, if a square is connected to another - for instance {B;2} and {A;2} – the value is 1. Notice that the diagonal elements of *W* (from top right to bottom left) of the matrix is always a zero diagonal. This is because the spatial matrix excludes the connection of a unit with itself. On top of all the possible choices to define *W*, row-standardization should also be considered. A row-standardized *W* transforms the element in the matrix so that each row sums up to a 1 (this in not the case of our example in Figure 2). If a unit has two contiguous neighbours, a row-standardized matrix will assign 0.5 to each rather than 1. Intuitively, this may make sense if we believe all units in the neighbourhood *j* have the same influence over the unit *i*. However, it has been argued that this assumption may be contradicting most theories of spatial dependence: "unless homogenous exposure is theoretically warranted, W should not be row standardized. If researchers are uncertain, the assumption of homogenous exposure can be tested against the assumption of heterogeneous exposure" (Neumayer and Plümper 2016, 191).

Moreover, as put by Beck and co-authors space is more than geography. If we believe that the outcomes in a certain unit are affected by outcomes in other units, those other units may be the geographic neighbours – or something else. Interconnected units may be units with dense trade relationships, with shared membership in regional organizations or even units with similar political institutions. Beck, Gleditsch, and Beardsley (2006) look into how trade interdependencies could define a *W* rather than a mere *W* based on geographic distance. Another example can be found in Böhmelt and co-authors (2017), where they show how leaders adjust their anti-coup polices

[10] It is advisable to consider alternative connectivity matrices to ensure results are not based on arbitrary definitions of *W*.

based on what other countries with similar coup's history have implemented, hence conceptualizing and operationalizing the *W* as shared history rather than shared geography.

Concluding this section on the *W* matrix, we want to recall some central questions a researcher should consider when defining and creating a *W*: i. Is the main criterion to define the W merely geographic? Or another form of interdependence should be considered? ii. Is the proximity defined on contiguity or distance? iii. What is the order of proximity? Should the neighbour's neighbours be considered?
Moreover, Neumayer and Plümper (2016), in their article succinctly titled "W", elaborate some further points that should be considered: first, *W* needs to capture the *causal mechanism* through which spatial dependence works. Hence, theory is the essence to define connectivity. Second, *W* determines whether total *exposure* to spatial dependence is specified as homogeneous or not (thinking critically therefore about row-standardization). Third, researchers specify the *dimensionality* of spatial dependence, whether there is a unique causal channel or multiple ones, hence thinking whether there are different *Ws* at play.

4. Do we need to model spatial feature?

The growing availability of spatial data and increasing awareness of the risks that interdependence poses for statistical inferences, has pushed researchers to test and model spatial interdependence. In some cases, spatial interdependence is an empirical nuisance researcher want to account for or get rid of. For example, Griffith's eigenvector spatial filtering (Griffith, 2003) can be used to remove the spatially interdependent component of a variable by splitting it into a spatial and a non-spatial component; the latter is filtered out of the original variable by regressing some linear combination of different connectivity matrices on the variable (Griffith 2003; Thayn 2017). However, in many other cases, we do want to model spatial interdependence: we do want to see whether it explains variation in the outcome of interest and how. Any theoretical account that involves interdependent actors is also interested in empirically modelling and estimating that interdependence. It can be the case, however, that the spatial dynamics of some social phenomenon are negligible and ultimately do not require sophisticated modelling nor filtering. How do we know whether we need to model spatial interdependence? What if our data present some hotspot (i.e., localized clustering) but overall do not exhibit spatial interdependence? In other words, how do we know that spatial clustering is due to interdependence rather than exposure to common shocks?

In this section, we illustrate a stepwise approach to spatial analysis. We first highlight the value of mapping spatial data, and then introduce the two main statistics for global and local for spatial autocorrelation, i.e., Moran's *I* and Geary's *G*. While these tests can tell whether a variable is spatially autocorrelated, they do not imply we should model that interdependence. To this aim, we show two ways of detecting interdependence, namely checking for autocorrelation in the residual of a simple OLS and the use of the LaGrange Multiplier tests.

4.1 Detecting Spatial Autocorrelation: Map it and test it

In their seminal article on "Contagion or Confusion?", Buhaug and Gledtisch (2008) start from a very simple observation. If you map armed conflict around the world (they focus on the years 2001-2005), you will see how violence clusters in space. Countries experiencing armed conflict are often surrounded by other countries having the same experience. It suffices to think about countries in the Sahel region and the Middle East. This pattern of similarity among neighbouring countries would suggest that conflict spreads across national boundaries; in other words, violence is contagious. Buhaug and Gleditsch, however, also show a map of GDP per capita in 2000, which seems to describe a very similar pattern in the same regions. Now, the association between civil war onset and poverty is probably one of the most robust findings of the civil war literature, and the two maps certainly corroborate this finding. This preliminary mapping exercise is very helpful, of course, as most of us unable to mentally map data and identify clusters of values. Mapping gives us the first quick glance into the spatial structure of the data we use as we can see immediately the geographic variation. Notably, mapping data to detect clusters implies that we expect the spatial structure to be geographic. This is often the most intuitive structure that country-level data exhibit, though as we have discussed, it is by no means the only one. An alternative mapping of a non-geographical structure would require a network, where connections can be defined in different ways. For simplicity, we will assume we are sure that the only relevant spatial connection in our data is geographic.

The maps presented by Buhaug and Gleditsch raise two important questions. The first one concerns our own ability to spot clusters. It is clear that some countries with armed conflict are close to each other in some regions, but those this mean that the data is *globally spatially autocorrelated*? Second, if conflict and GDP both clusters in the same regions and GDP is a plausible cause of armed conflict, is GDP explaining the spatial autocorrelation rather than conflict spreading across countries? Put differently, is y_i (i.e. conflict in a country i) caused by y_j (i.e. conflict in neighbouring countries) or by x_i (i.e. GDP in country i, independent of j) which just happen to cluster in space? Notice that in the first instance, we have spatial interdependence, thus an issue of endogeneity; in the second instance, the spatial autocorrelation does not introduce interdependence in the observations and is usually solved by controlling for x_i. But we now begin addressing the first question, namely, how to test for spatial autocorrelation.

Tests of spatial autocorrelation aim at detecting non-randomness in the spatial distribution of a variable. Data may exhibit positive or negative spatial autocorrelation. Positive spatial autocorrelation implies that similar values cluster in space. For example, rich countries are close to rich countries, while poor countries are close to poor countries. Notice that positive autocorrelation is only concerned with the similarity of neighbouring units and can be driven by either *high-high* or *low-low* clustering. Negative spatial autocorrelation, on the other hand, describes a pattern in the data where nearby values are systematically dissimilar, such as a democratic country surrounded by mostly autocratic countries. In this case, the pattern is of *high-low* or *low-high*. While we can spot these patterns by looking at maps, we are bad at assessing how serious autocorrelation is; and sometimes, we might simply not be able to spot any autocorrelation even when it is present. There are two types of indicators that can help us to identify statistically significant spatial autocorrelation, namely global indicators and local indicators. The main difference between these two is that global indicators provide one single statistics for spatial autocorrelation across observations; local indicators will produce one score for each observation, allowing us to identify exactly where the spatial clustering occurs.

The most used global test for spatial autocorrelation is Moran's I. Intuitively, Moran's I calculates the correlation between the values of a variable in unit i and the values of the variable in all other locations. Notice that, however, Moran's I is not exactly a correlation coefficient, as we will see. The formula of Moran's I is:

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} \ (y_i - \bar{Y})(y_j - \bar{Y})}{\sum_{i=1}^{N} (y_i - \bar{Y})^2}$$

The numerator is the covariance between values in y_i and y_j. Notice that y_j is the "neighborhood" of y_i as defined by the spatial matrix $w_{i,j}$. S_0 is the sum of all $w_{i,j}$. It is sometimes erroneously believed that Moran's I ranges from -1 to 1, as other correlation coefficients. In fact, this is a key difference as I does not equal 0 under the null hypothesis. The null hypothesis of the Moran's I test is that in absence of spatial autocorrelation the expected value $I_0 = \frac{-1}{(N-1)}$. If the observed value of $\hat{I} > I_0$, the data exhibit positive spatial autocorrelation; on the other hand, an observed $\hat{I} < I_0$ indicates negative spatial autocorrelation. Statistical software (R and Stata) report the expected and observed value I, the correspondent z-score and its p-value that allows us to reject (or not) the null hypothesis of random spatial distribution. There are two important points to keep in mind when using Moran's test. First, the test will report different results depending on how we define the spatial matrix w_{ij}; as w_{ij} changes, y_i will be compared to different y_j. It is useful to explore how Moran's I changes as the number of spatial lags or distance vary. Correlograms are helpful for this purpose as they plot the estimated Moran's score as a function of distance or lags on the x-axis.

Second, the Moran's test work under the assumption that the variable of interest is normally distributed. This assumption usually holds when continuous variables are used. When dealing with count variables, it is often possible to transform them in ratios, for example by dividing the count by population or area. In turn, such transformed variable can be used for the Moran's test. When the variable under scrutiny is not continuous, however, spatial autocorrelation can be tested using join counts. In presence of a binary variable $0/1$, y_i and y_j can either take values (0,0), (1,1) or (0,1). Clustering of similar values (0,0 or 1,1) would indicate positive autocorrelation, while a higher number if dissimilar combinations (0,1) would indicate negative autocorrelation.

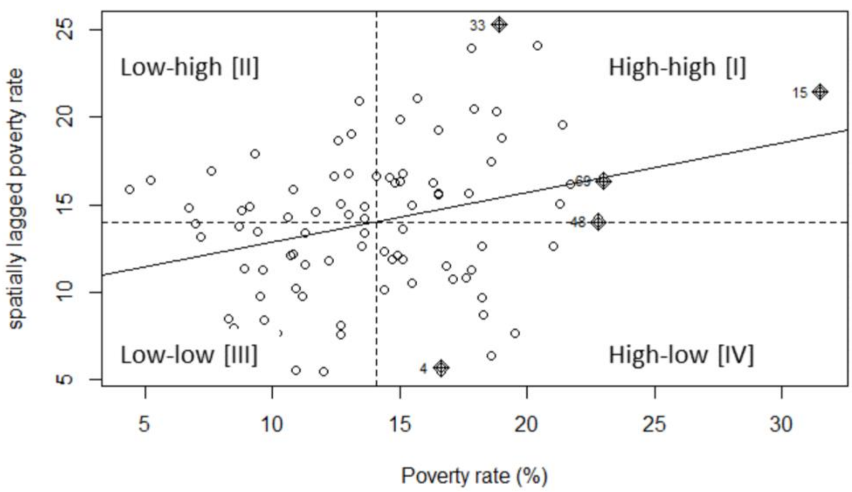
Suppose that we have performed a Moran's test for global spatial autocorrelation and have failed to reject the null hypothesis, i.e., there is no evidence of global spatial interdependence in our variable of interest. Does this mean that we can be confident that there is no spatial interdependence in our data? It is possible, in fact, that the Moran's test does not report a significant test statistic even if some local clustering exists in the data. This is because such clustering may be not significant enough to be picked up by the global statistics. As pointed out by Anselin "it is quite possible that the local pattern is an aberration that the global indicator would not pick up, or it may be that a few local patterns run in the opposite direction of the global spatial trend" (Anselin 1995, 97). Local indexes of spatial autocorrelations (LISAs) can be used to detect such clusters of similar or dissimilar values. Notice that LISAs are worth exploring not only when the Moran's global statistics is not significant; even when we do detect global autocorrelation, LISAs help detect exactly where the correlation may be occurring among our units.

As with global statistics, there are several available tests. Here, we will keep our focus on Moran's statistics and illustrate the local version of Moran's I. The formula for the Local Moran's I can be written as:

$$I_i = \frac{(y_i - \bar{Y}) \sum_{j=1}^{N} w_{i,j} \left(y_j - \bar{Y}\right)}{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2}$$

Local Moran's statistics is estimated for each unit i, in contrast with the global statistics that returns one single statistics for the entire sample. Thus, for each unit we can obtain the observed local statistics of spatial autocorrelation, its expected value, its z-score and its p-value. In fact, it is also possible to map the z-scores for each unit to visualize the location of clusters and hotspot that are statistically significant. The interpretation is not dissimilar to the global test, so that an observed statistic greater than the expected statistic in a unit is suggestive of local clustering (i.e., the unit is surrounded by similar units), and vice versa for a smaller observed statistic. The Moran's scatterplot is an extremely useful visualization tool summarizing both global and local Moran's tests. Figure 3 shows the Moran's scatterplot using data on poverty rates in Ohio's counties in 2015. The scatterplot visualizes the poverty rate in each county against poverty rate in the county's neighbourhood. In the first and third quadrants (high-high and low-low), there are counties that are surrounded by counties with similar levels of poverty rates; in quadrants two and four (low-high and high-low), on the other hand, there are counties that are dissimilar from their neighbours. Most observations fall in quadrants one and three, thus suggesting overall a positive spatial autocorrelation summarizes by the positive slope of the linear fit. In fact, the slope of the linear trend corresponds to the global Moran's I with row-standardized spatial matrix.

Figure 3. Moran's Scatterplot on poverty rate (source: (Wu and Kemp 2019)*)*



Importantly, neither global indicators or LISAs can ascertain whether the spatial association is due to clustering of other factors associated with y_i or to the interdependence between y_i and y_j. Once we find evidence for spatial autocorrelation, we can use two main strategies to assess whether such autocorrelation is likely interdependence that would make an OLS estimate biased and inconsistent. A first simple check is to run an OLS with all relevant control variables and then test for spatial autocorrelation in the residuals using global and local tests. If the spatial clustering in the outcome is the result of covariates' clustering, likely controlling for them should remove the spatial pattern and we should see not spatial autocorrelation in the residuals. The LaGrange Multiplier test (LM test) is an alternative approach that compares our OLS with two other models, namely the spatial lag and the spatial error models. Rejecting the null hypothesis of the LM test means that the alternative model (spatial lag or spatial error) is preferred to the non-spatial OLS. The LM test compares the OLS to each alternative spatial specification but does not allow to pick between the two when both have a significant test statistic. When this happens, one can estimate a Robust LM test to identify the most appropriate spatial model.[11]

5. How to model spatial interdependence?

Suppose that you have used the tests for spatial autocorrelation and found that there seem to be local and/or global autocorrelation in your data, and the LaGrange Multiplier tests also suggests this is due to spatial interdependence. More specifically, the test would likely indicate that a spatial lag model or a spatial error models are superior to a simple OLS model. What is the difference between these two alternatives? Which other spatial model can be used to estimate interdependence among observation?

Elhorst (2014) outlines the relationship between different spatial models (and non-spatial OLS) starting from what he calls the General Nesting Spatial model (GNS). The

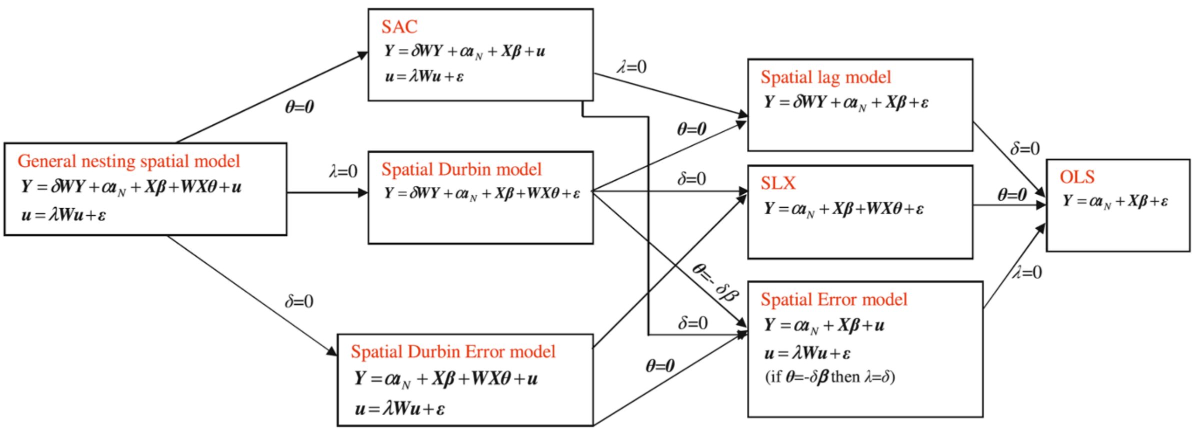
[11] See (Darmofal 2006, chap. 6) discussion of the differences between robust and non-robust LM tests and the decision rule proposed by Anselin (2005).

GNS includes, as the OLS, a parameter $\boldsymbol{\beta}$ that is the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ and a constant term $\boldsymbol{\alpha}$, but also:

- the spatial autoregressive coefficient $\boldsymbol{\delta}$ that is the effect of $\boldsymbol{Y}$ in neighbouring units (defined by the spatial matrix $\boldsymbol{W}$). This parameter is often also indicated as $\boldsymbol{\rho}$.
- the spatial spillover parameter $\boldsymbol{\theta}$ that is the effect of $\boldsymbol{X}$ in neighbouring units (defined by the spatial matrix $\boldsymbol{W}$)
- the spatial autocorrelation coefficient $\boldsymbol{\lambda}$ in the error term $\boldsymbol{u}$

Each of these parameters has a very clear practical interpretation. In the first case, we are assuming that, for example, when estimating the likelihood of a civil war in a country, we need to account for civil war onset in neighbouring units as well as conflict may be "contagious". In the second case, we are assuming that, for example, regime types (democracies vs autocracies) may influence the likelihood of civil wars, but it is likely that regime types cluster in space. Hence, we may want to control not only for the type of regime in each country, but also in its neighbourhood. Notice that in both examples the spatial clustering we observe is due to *observable* factors and their spatial structure (either in the outcome or in the covariates) (Cook, Hays, and Franzese 2015). In the third case, the clustering is due to *unobservable* factors that are spatially interdependent.

Figure 4. Spatial Models classification presented by (Elhorst 2014)



There are two additional important things to note here. First, the spatial autoregressive coefficient is what is commonly (though sometimes inaccurately) referred to as a spatial lag. This should not be confused with the spatial autocorrelation coefficient, which is also sometimes referred to the spatial error term. The former is autoregressive because values of a variable in unit i depends on values of the same variable on units j (and vice versa); the latter concerns autocorrelation in the error terms for observations linked in the spatial matrix $\boldsymbol{W}$, similarly to what serial autocorrelation is in time-series models. Finally, notice that parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are (assumed to be) both exogenous. We will see that this implies less estimation problems for models where $\boldsymbol{\delta} = 0$, as reverberation effects due to the global spillover need to be accounted for. In the GSN model, all parameters are non-zero; Elhorst classification is very useful as it shows how the combination of each coefficient results in a specific spatial model. It follows that when parameters $\boldsymbol{\delta}, \boldsymbol{\theta},$ and $\boldsymbol{\lambda}$ are zero, our model is just an OLS.

In the Spatial Autoregressive Combined model (SAC), both the outcome variable $\boldsymbol{Y}$ and the error term $\boldsymbol{u}$ exhibit spatial interdependence. The Spatial Lag model (or Spatial Autoregressive, SAR) is a nested version of the SAC model where the spatial autocorrelation parameter $\boldsymbol{\lambda}$ is zero. In the Spatial Durbin model (SD), both $\boldsymbol{Y}$ and $\boldsymbol{X}$ are spatially correlated and the model includes the terms $\boldsymbol{\theta W X}$ for the spillover effect and $\boldsymbol{\delta W Y}$ for the spatial intercedence. When spatial autocorrelation is only in the vector of $\boldsymbol{X}$, the model becomes a Spatial Lag of X model (SLX). The Spatial Durbin Error model (SDEM), intuitively, includes again the spatial lag of $\boldsymbol{X}$ but instead of the spatial lag of $\boldsymbol{Y}$, it includes the spatial autocorrelation in the error term. An SDEM without the spatial lag of $\boldsymbol{X}$ is a Spatial Error model (SEM).

As mentioned, the LaGrange Multiplier test, particularly its robust version, helps us to identify which spatial model we should use in presence of spatial interdependence. But the spatial lag and the spatial error model are two out of the six possible spatial models illustrated above and in Elhorst's classification. As Cook and co-authors put it (2015, 10) put it, these models assume "that the spatial heterogeneity in the outcomes arises from a single source, constraining the other possibilities to zero". Ideally, theory should help researchers identifying the type of model that more closely allows to test an expectation. In their review of the literature, Wimpy and co-authors (2020) claim that most researchers have focused on SAR models, even if this particular model does not match the theoretical expectations regarding the existing spatial relationships among the units. Cook and co-authors (2015) provide a useful guiding principle according to which researchers should first of all identify the sources of spatial clustering as observable and unobservable. Moreover, they add:

> "where researchers are principally interested in obtaining unbiased estimates of the non-spatial parameters, the spatial Durbin model should be preferred. This should provide the most insurance against possible omitted variable bias by explicitly introducing both forms of observable spillovers into the systematic component of the model. However, where researchers are explicitly interested evaluating spatial theories, we believe one of the other two-source models (SAC or SDEM) are best. Each frees one parameter to capture spillovers in observables (either δ or θ) while accounting for spatial effects in the unobservables (λ)." (Cook, Hays, and Franzese 2015, 16)

At this point, one might wonder why we should care about spatial interdependence in a statistical model, and how ignoring it affects our inferences. Suppose we want to establish the relationship between civil wars and regime type, more specifically whether autocratic regimes are more likely to experience civil unrest. We can think of this as a simple OLS model with no spatial component in it. In this non-spatial OLS we are assuming that y_i is independent of y_j ($i \neq j$). However, this assumption does not hold if the probability of conflict in a country i is affected by the probability that conflict also erupts in country j. Ignoring this spatial interdependence when it exists means the OLS will exhibit omitted variable bias, which means it will produce inefficient and biased coefficients (Franzese and Hays 2008). More specifically, as reported by Franzese and Hays, the OLS will overestimate the impact of the non-spatial covariates. We could address this omitted variable problem by including the spatial lag of the dependent variable in the right-hand side of the equation as in a spatial OLS. This, however, introduces an endogeneity problem due to the simultaneity between y_i and y_j, which will be affecting each other at the same time and thus make the estimates inconsistent. In this case, the spatial lag is likely to be an overestimation of the actual spatial interdependence, and coefficient of non-spatial covariates will likely exhibit a downward bias[12]. At higher levels of spatial interdependence, we need to model spatial relationships more accurately than the OLS can do. Franzese and Hays here suggest two options. First, one could estimate a spatial two-stage least square where the spatial lag of other non-spatial covariates ($\boldsymbol{WX}$) is used as instrument for the spatial lag of the outcome variable $\boldsymbol{WY}$. Second, one could estimate a spatial lag model (or SAR model) with a maximum likelihood estimator. Franzese and Hays (2008) find the latter to weakly dominate the former. [13] In the remaining of this section, we discuss the challenges in estimating effects from spatial models. More specifically, we first discuss the (relatively) simpler SEM, SDEM and SLX models and then move to spatial models with an autoregressive component (SAR, SDM and SAC models). The challenges stemming from the latter should push researchers to think about the nature of the expected spatial effects and the extent to which models with more assumptions and less straightforward interpretations (SAR, SDM and SAC) are necessary.

5.1 SEM, SDEM and SLX models

As probably the second most popular spatial model, we begin presenting the Spatial Error Model (SEM). A SEM can be written as:

$$y = \boldsymbol{X}\beta + u$$

$$\text{where } u = \lambda \boldsymbol{W} u + \varepsilon$$

So, the SEM can also be re-written as:

$$y = \boldsymbol{X}\beta + \lambda \boldsymbol{W} u + \varepsilon$$

[12] It should be noted, though, that Franzese and Hays conclude that the simultaneity bias is less concerning than the omitted variable bias when the magnitude of the spatial interdependence is below 0.3 (Franzese and Hays 2008).

[13] Stata 16 now allows to estimate spatial models using both the instrumental variable approach and the maximum likelihood estimator. The R package *spdep* includes the *lagsarlm* function that uses the maximum likelihood approach.

As mentioned, the SEM allows to estimate the spatial dependence of the error for observations that are connected via the $\boldsymbol{W}$ matrix. Ward and Gleditsch (2018) point out that the presence of spatial autocorrelation in the disturbances may well be the result of some mispecifications in the model, such as the omission of variable that clusters in space. One can also imagine the case of model that incorporate a spatial lag, but the connectivity matrix is defined in a way that does not reflect the actual spatial connections among units. This type of misspecification may result in the LaGrange Multiplier test to support the use of a SEM.

Compared to the models we discuss in the next section, presenting the effects in an SEM model is straightforward because there is no feedback effect to account for. If we are interested in the impact of a covariate on the outcome, we can simply interpret the coefficient from the regression table and ignore the parameter of the spatial autocorrelation λ. The effect of a covariate on an outcome does not travel in space because the only spatial component of the equation is in the error term. Hence, the interpretation of coefficient in a SEM model is ultimately the same as in a non-spatial model. The SDEM has the same parameters of the SEM, plus the spatial lag of covariates ($\boldsymbol{\theta W X}$). This additional term captures the so-called *local spillovers*, namely the effect of covariates in neighbouring units *j* on a specific unit *i*. This effect, however, does not further spillover to other neighbouring units in a cascade effect; it stops at unit *i*. What this imply is that the spatial lag of covariates does not introduce feedback effects that, as we will see, are responsible for the *global spillovers* in the SAR, SAC and SDM models and require spatial econometric techniques. For the same reasons, the estimates of a SLX model do not present particular challenges and coefficients can be interpreted as in classic linear models.

5.2 The problem of Global Spillovers: Spatial Effects in SAR, SAC and SDM

We begin this discussion with the SAR model, also known as the spatial lag model. This is one of the most used spatial models among those available to researchers. As indicated in Figure 4, the SAR model can be written as

$$y = \boldsymbol{X}\beta + \delta \boldsymbol{W} y + \varepsilon$$

In turn, this can be rewritten as:

$$(\boldsymbol{I} - \delta \boldsymbol{W})y = \boldsymbol{X}\beta + \varepsilon$$

This means that the expected value of y is

$$E(y) = (\boldsymbol{I} - \delta \boldsymbol{W})^{-1} \boldsymbol{X}\beta$$

In absence of spatial interdependence, the expected outcome would be predicted by the exogenous covariates and their estimated coefficient; in presence of spatial interdependence, however, the term $\boldsymbol{X}\beta$ is multiplied by $(\boldsymbol{I} - \delta \boldsymbol{W})^{-1}$. Indeed, the term $(\boldsymbol{I} - \delta \boldsymbol{W})^{-1}$ is also known as the spatial multiplier and it "tells us how much of the change in x_i will spill over onto other states *j* and in turn affect y_i through the impact of y in the spatial lag" (Ward and Gleditsch 2018, 59). In other words, this implies that a change in a covariate in one single observation will affect the outcome in other units, depending on the degree of connectivity defined by $\boldsymbol{W}$. If the change occurs in a unit that has no neighbour (so-called islands), there is no spill over. Conversely, a small change in a very connected unit is likely to reverberated throughout the cluster of neighbours. Because of this reverberation, the spatial effects of the SAR model are also called *global spillovers* because any change in any unit will affect other units. It follows that the impact of the change will depend on the specific unit where the change itself takes place. Now, researchers can estimate three main effects of interests:

1. Direct effect $(\Delta \boldsymbol{X})\beta$: that is the effect of a change in x_i on y_i. Notice the average direct effect concerns changes occurring in the same unit *i,* so this is intuitively similar to the standard interpretation of β. However, notice that in a spatial model the effect of x_i on y_i and then on y_j which feedbacks to y_i is accounted for.
2. Indirect effect of x_j on y_i $[(\boldsymbol{I} - \delta \boldsymbol{W})^{-1} - \boldsymbol{I}]$ $(\Delta \boldsymbol{X})\beta$, or in other words the average impact of a change of a covariate in the neighbourhood *j* on the outcome in *i*.
3. Total effect $(\boldsymbol{I} - \delta \boldsymbol{W})^{-1}$ $(\Delta \boldsymbol{X})\beta$, which is the sum of the average direct and indirect effects.

Statistical software nowadays allows researchers to easily estimate these effects, but it is worth reminding that these spatial effects are not straightforward to interpret from standard regression tables. Also, differently from effects in non-spatial models, spatial effects are different for each unit of the sample, simply because the composition of the $\boldsymbol{W}$ matrix changes for each unit (i.e., each unit has different neighbours). This is the reason why LeSage and Pace suggest to report the three spatial effects above as averages (LeSage and Pace 2014), as some statistical software already do. It is also possible, however, to explore how a change in one specific unit (which may be of particular interest for theoretical reason) affects other units. Or, to rank different units in terms of their impact on another specific unit of interest. For interested readers, Ward and Gleditsch (2018) illustrate a step-by-step procedure to present these effects in R.

Now that we have discussed the problem of how to interpret and report direct and indirect effects in SAR due to the spatial multiplier, one can easily understand how the same issues apply to the case of the SAC and the SDM. Both models include a spatial lag of the dependent variable, hence suggesting that outcomes across the sample will affect each other, and so will changes in their covariates. The direct and indirect effects for the SAC and SAR have the same interpretation, as the inclusion of the spatial interdependence in the disturbances of the SAC ($\lambda \boldsymbol{W} u$) does not introduce additional feedbacks to account for in the model as we also discussed in the previous section. It is slightly different for the SDM case, which includes the spatial lag of the dependent variable y and the spatial lag of x ($\theta \boldsymbol{W} X$) at the same time. This inclusion means that there are not only global spillovers due to the spatially lagged outcome, but also local spillovers. Again, while global spillovers affect all units' outcomes via other units, local spillovers only affect the immediate neighbours of the unit where the change occurs. This means that the total average effect (see above) in the SDM will need to include another term, that is $\theta \boldsymbol{W}(\Delta \boldsymbol{X})$.

6. Conclusions

In this article we have introduced some intuitions on spatial interdependence and how political scientists could start thinking about modelling it rather than filtering it out and treating as a residuals' nuisance.

Concluding and going back to our warning about the theoretical need for some spatial models, the initial fascination with sophisticated spatial econometrics seem to have been replaced by more careful evaluation of the appropriateness of such models. Elhorst and Vega (2013) suggests a "revision" in the way researchers select statistical models that does not uniquely rely on statistical tests (Moran's, LaGrange, etc) but more on theory and context. One interesting conclusion they draw is that "global spillover specifications, unless theoretically motivated, are difficult to justify or have been overused in applied studies" (Elhorst & Vega 2013, p.11). In a more recent study, Wimpy et al (2020) review applied research in political science that has used spatial models has over-relied on the SAR when, in fact, this was not the most accurate spatial model for the given theoretical account. In line with Elhorst and Vega's call to reconsider the SLX model as a starting point, Wimpy et al show "even if the true GDP is SAR, the SLX performs quite well at detecting spatial relationships; […] the same cannot be said of the SAE when the true DGP is the SLX" (Wimpy et al 2020, p.33).

As we have stressed earlier this brief article is just a teaser, we strongly advise to read the books by Darmofal (2015) and Ward and Gleditsch (2018) because written specifically for political scientists. Most likely the major and most demanding part of the learning curve will be on dealing with spatial data, their spatial merging and defining the W matrixes. Of course, also learning how to run the apt tests and estimators will be a central component of the learning curve, but we believe it will be quite rewarding.

References

Anselin, Luc. 1995. “Local Indicators of Spatial Association-LISA.” *Geographical Analysis* 27(2): 93–115.

———. 2005. “Exploring Spatial Data with GeoDaTM: A Workbook.”

Anselin, Luc, Julie Le Gallo, and Hubert Jayet. 2008. “Spatial Panel Econometrics.” In *The Econometrics of Panel Data*, Springer, 625–60.

Arjona, Ana. 2019. “Subnational Units, the Locus of Choice, and Concept Formation.” *Inside Countries: Subnational Research in Comparative Politics*: 214.

Beck, Nathaniel, Kristian Skrede Gleditsch, and Kyle Beardsley. 2006. “Space Is More than Geography: Using Spatial Econometrics in the Study of Political Economy.” *International Studies Quarterly* 50(1): 27–44.

Bivand, Roger S, Edzer J Pebesma, Virgilio Gomez-Rubio, and Edzer Jan Pebesma. 2008. 747248717 *Applied Spatial Data Analysis with R*. Springer.

Böhmelt, Tobias, Andrea Ruggeri, and Ulrich Pilster. 2017. “Counterbalancing, Spatial Dependence, and Peer Group Effects.” *Political Science Research and Methods* 5(2): 221–39.

Buhaug, Halvard, and Kristian Skrede Gleditsch. 2008. “Contagion or Confusion? Why Conflicts Cluster in Space.” *International Studies Quarterly* 52: 215–33.

Cook, Scott J., Jude C. Hays, and Robert Franzese. 2015. APSA Conference Paper *Model Specification and Spatial Interdependence*.

Costalli, Stefano, and Andrea Ruggeri. 2015a. “Forging Political Entrepreneurs: Civil War Effects on Post-Conflict Politics in Italy.” *Political Geography* 44: 40–49.

———. 2015b. “Indignation, Ideologies, and Armed Mobilization: Civil War in Italy, 1943–45.” *International Security* 40(2): 119–57.

Darmofal, David. 2006. “Spatial Econometrics and Political Science.” *Society for Political Methodology*: 1–40.

———. 2015. *Spatial Analysis for the Social Sciences*. Cambridge University Press.

Daxecker, Ursula, Jessica Di Salvatore, and Andrea Ruggeri. 2019. “Fraud Is What People Make of It: Election Fraud, Perceived Fraud, and Protesting in Nigeria.” *Journal of Conflict Resolution* 63(9): 2098–2127.

Di Salvatore, Jessica. 2016. “Inherently Vulnerable? Ethnic Geography and the Intensity of Violence in the Bosnian Civil War.” *Political Geography* 51: 1–14.

———. 2018. “Does Criminal Violence Spread? Contagion and Counter-Contagion Mechanisms of Piracy.” *Political Geography* 66: 14–33.

Eck, Kristine. 2012. "In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets." *Cooperation and Conflict* 47(1): 124–41.

Elhorst, J P. 2014. *Spatial Econometrics - From Cross-Sectional Data to Spatial Panels*. Springer.

Elhorst, J.Paul, Solmaria Halleck Vega, J.Paul Elhorst, and Solmaria Halleck Vega. 2013. "On Spatial Econometric Models, Spillover Effects, and W."

Ferwerda, Jeremy, and Nicholas L Miller. 2014. "Political Devolution and Resistance to Foreign Rule: A Natural Experiment." *American Political Science Review* 108(3): 642–60.

Fotheringham, A Stewart, and David WS Wong. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and planning A* 23(7): 1025–44.

Franzese, Robert J., and Jude C. Hays. 2008. "Interdependence in Comparative Politics: Substance, Theory, Empirics, Substance." *Comparative Political Studies* 41(4–5): 742–80.

Gilmore, Elisabeth, Nils Petter Gleditsch, Päivi Lujala, and Jan Ketil Rod. 2005. "Conflict Diamonds: A New Dataset." *Conflict Management and Peace Science* 22(3): 257–72.

Gleditsch, Kristian Skrede, and Nils B Weidmann. 2012. "Richardson in the Information Age: Geographic Information Systems and Spatial Data in International Studies." *Annual Review of Political Science* 15: 461–81.

Griffith, Daniel a. 2003. *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*.

Harbers, Imke. 2015. "Taxation and the Unequal Reach of the State: Mapping State Capacity in E Cuador." *Governance* 28(3): 373–91.

Harbers, Imke, and Matthew C Ingram. 2017. "Geo-Nested Analysis: Mixed-Methods Research with Spatially Dependent Data." *Political Analysis*: 1–19.

Kocher, Matthew A, and Nuno P Monteiro. 2016. "Lines of Demarcation: Causation, Design-Based Inference, and Historical Research." *Perspectives on Politics* 14(4): 952–75.

Koktsidis, Pavlos I. 2014. "How Conflict Spreads: Opportunity Structures and the Diffusion of Conflict in the Republic of Macedonia." *Civil Wars* 16(2): 208–38.

LeSage, James P., and R. Kelley Pace. 2014. "Interpreting Spatial Econometric Models." In *Handbook of Regional Science*, Springer Berlin Heidelberg, 1535–52.

Midlarsky, Manus I, Martha Crenshaw, and Fumihiko Yoshida. 1980. "Why Violence Spreads: The Contagion of International Terrorism." *International Studies Quarterly* 24(2): 262–98.

Neumayer, Eric, and Thomas Plümper. 2016. "W." *Political Science Research and Methods* 4(1): 175–93.

Ruggeri, Andrea, Theodora-Ismene Gizelis, and Han Dorussen. 2011. "Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection, and Data Quality." *International Interactions* 37(3): 340–61.

Thayn, Jonathan B. 2017. "Eigenvector Spatial Filtering and Spatial Autoregression." In *Encyclopedia of GIS*, Springer International Publishing, 511–22.

Tobler, Waldo R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic geography* 46(sup1): 234–40.

Tollefsen, Andreas Forø, H\a avard Strand, and Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49(2): 363–74.

Toset, Hans Petter Wollebæk, Nils Petter Gleditsch, and H\aavard Hegre. 2000. "Shared Rivers and Interstate Conflict." *Political geography* 19(8): 971–96.

Ward, Michael D., and Kristian Skrede Gleditsch. 2018. *Spatial Regression Models*. 2nd ed. SAGE Publications, Inc.

Weidmann, Nils B, and Sebastian Schutte. 2017. "Using Night Light Emissions for the Prediction of Local Wealth." *Journal of Peace Research* 54(2): 125–40.

Wimpy, Cameron, Guy D Whitten, and Laron Williams. 2020. "X Marks the Spot: Unlocking the Treasure of Spatial-X Models." *Journal of Politics* Forthcomin.

Wu, A.-M., and K. K. Kemp. 2019. "Global Measures of Spatial Association ." In *The Geographic Information Science & Technology Body of Knowledge* , ed. John P. Wilson. Washington, DC: Association of American Geographers.

Zhukov, Yuri M. 2012. "Roads and the Diffusion of Insurgent Violence: The Logistics of Conflict in Russia's North Caucasus." *Political Geography* 31(3): 144–56.