CHAPTER 3. INTERDEPENDENCY AMONG OBSERVATIONS

Interaction and Social Science

In this chapter, we examine how insights from *spatial* analysis can help researchers take dependence between observations into account and deal with spatially clustered phenomena. The term *spatial* has a broad meaning in this context. On the one hand, space can refer to conventional forms of geographical distances. At the same time, one can also expand the concept of space to the myriad ways in which observations may be connected. This includes forms of social connectivities beyond networks defined purely by geographical proximity. The term *social* can be defined in terms of transactions, legacy, heritage, and many other aspects of social, economic, and political life.

In particular, we focus on two important regression models with (spatially) dependent observations. The first of these concerns situations in which there is a spatially lagged dependent variable, where the response for one observation has a direct impact on other connected observations. The second focuses on regression models in which errors are spatially correlated. We recognize that there is a much larger set of interesting spatial modeling perspectives. This monograph is not intended to provide an exhaustive survey of these, but rather serves to introduce models with spatially lagged dependent variables and those with spatially correlated error terms. Many empirical undertakings in social science may benefit from these approaches that until very recently have been widely ignored in much of the empirical social science literature.

These types of models allow us to examine the impact that one observation has on other, proximate observations. We believe this is important not only from first principles but also from the simple fact that many social phenomena are spatially "clustered."

In short, there are myriad studies across the gamut of the social sciences that employ data that are actually organized on a spatial template, whether the units are counties, cities, states, countries, firms, or individuals. It often turns out that the characteristics of these units are highly clustered in particular spatial regions. In many of these applications, it is plausible to assume that there may be dependencies across the observations. In practice, this clustering is generally ignored or treated

as a nuisance. Ignoring these dependencies imposes a substantial price on our ability to generate meaningful inferences about the processes we study. Spatial analysis provides one way of reducing that price and taking advantage of the information we have about how social processes are interconnected. We turn next to a simple example of how this works in an important area of social science, namely, the study of the diffusion of democratic institutions.

Democracy Around the World

To motivate our discussion, we use a simple example with data where observations are unlikely to be independent of one another. Social scientists have long been interested in possible explanations for why some countries are democracies and others not. An early and influential contribution by Lipset (1959) suggested that there were certain social requisites for democratic rule. One of these requisites was high levels of average income; Lipset noted that average wealth tends to be considerably higher for the more democratic countries (p. 75). This argument—which has served as a cornerstone of comparative analysis for more than four decades—suggested that societies with higher average income were more likely to have democratic institutions. Table 3.1 provides an abbreviated view of data on gross domestic product (GDP) per capita and level of democracy for most countries in the world in 2014–2015.

Our measure of democracy is the so-called Polity index, which classifies countries on a series of institutional criteria. The index ranges from −10 for the least democratic societies to 10 for the most democratic societies. Gleditsch and Ward (1997) provide further details on the construction of this index. We have sorted this table on GDP per capita and democracy so that it is easier to see simple patterns among the variables. As can be seen, some wealthy societies, such as Denmark, are indeed democratic, while low-income countries, such as Sierra Leone and North Korea, are autocracies. Interestingly, Lipset suggested that in 1959, Australia, Belgium, Canada, Denmark, Ireland, Luxembourg, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, and the United States made up a list of "stable democracies" in Europe and North and South America. The unstable democracies and dictatorships in 1959 included Austria, Finland, France, West Germany, Italy, and Spain. All these countries are now democracies and generally considered no less stable than the countries on the other list. Despite the fact that we can find some cases at either end of the per capita income

Table 3.1. Democracy Data (PITF: 2015) and Logged GDP per Capita (WB: 2014)

Country	Polity	ln GDP per Capita
Luxembourg	10	11.67
Norway	10	11.48
Qatar	−10	11.46
Switzerland	10	11.36
Australia	10	11.03
Denmark	10	11.02
Sweden	10	10.99
Singapore	−2	10.93
Ireland	10	10.92
United States	10	10.91
Burundi	−1	5.66
Central African Republic	−10	5.87
Malawi	6	5.89
Niger	6	6.07
Congo (DRC)	5	6.08
Gambia	−5	6.09
Madagascar	6	6.12
Liberia	6	6.13
Guinea	4	6.29
Somalia	5	6.29

Note. Top and bottom 10 countries in terms of logged GDP per capita are shown. GDP = gross domestic product.

spectrum that clearly are consistent with Lipset's claim, is there a strong general relationship between wealth and democracy? India is democratic in spite of low average national income, and although India has recently experienced high rates of growth, it remains far below the levels observed for Organisation for Economic Co-operation and Development (OECD) countries. At the same time, it is also hard to ignore the existence of many relatively high income autocracies situated in the Middle East, which seems to contradict the claim made by Lipset. To evaluate the relationship more generally, we turn to a systematic, comparative analysis.

Following the work of Lipset (1959) and many others since, it is common in empirical, comparative work on democracy to consider democracy as a linear function of the natural log of GDP per capita. We estimate the level of democracy in a country, measured by the Polity score, given

its GDP per capita using ordinary least squares (OLS) regression.

$$\text{Polity score} = \beta_0 + \beta_1 \ln \text{GDP per capita} + \epsilon. \tag{3.1}$$

The estimates for this linear regression of democracy on GDP per capita are shown in Table 3.2. The positive sign of the coefficient for ln GDP per capita illustrates the positive relationship between democracy and income, but the estimated substantive impact is actually relatively small when we take into account the metric of the variables.

Table 3.2. OLS Regression of Polity on Logged GDP per Capita

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	t Value	$\Pr(>\|t\|)$
Intercept	−5.70	2.88	−1.98	0.05
GDP	1.14	0.33	3.44	0.00

More specifically, this linear model predicts that a country with Burundi's GDP per capita ($287 in 2014) would have a democracy score of approximately 1. By contrast, for a country that has a level of GDP per capita income approximately twice that of Uzbekistan ($2038), the model predicts an associated democracy score of about 3. For most analysts, scores of 1 and 3 are considered to be similar on the Polity democracy index. Thus, there does not seem to be a large impact of even fairly dramatic differences in income on the predicted level of democracy, despite the statistical significance of the estimated coefficient for the log of GDP per capita.[1]

Figure 3.1 shows that despite the precision with which the linear effects are estimated, the estimated OLS equation predicts democracy levels that are generally not close to the actual values in the data. Only 1 in 20 of the actual observations fall within a standard error of the regression line. Nonetheless, the implied, estimated effect of wealth on democracy is not only small—more than doubling the GDP per capita has a small impact on democracy—for poor countries, such as Uzbekistan. Almost any standard analysis of these residuals will reflect the first impression given in this figure: They do not look "well behaved," in the sense that the mean prediction of the model is a full two points higher than the

[1] We recognize that there are different inferential frameworks, but in this primer we will stick mainly with a classical interpretation of estimated coefficients and empirical standard errors. The analysis in this paragraph does not consider the uncertainty around these estimates.

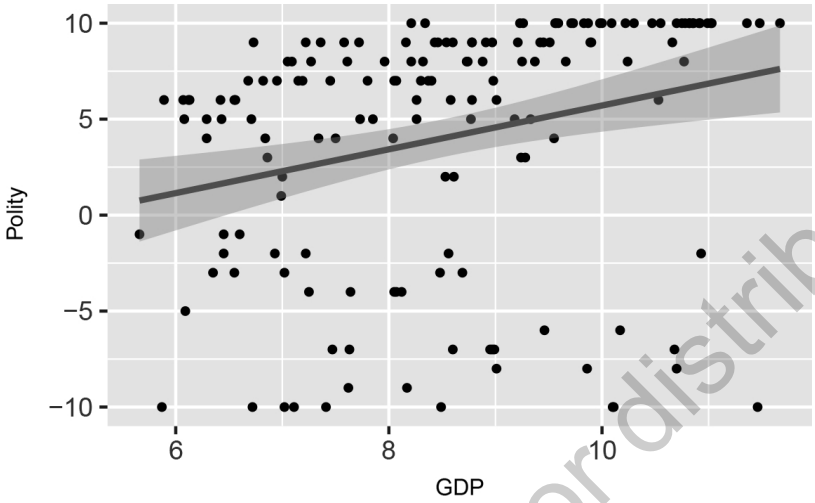

Figure 3.1. Polity Scores as a Function of Logged GDP per Capita, 2014–2015

Note. Ordinary least squares regression line with the regression standard error as a gray band.

Source: Created by Ward and Gleditsch.

mean of the actual data, suggesting that the model actually overpredicts the Polity score substantially.

Figure 3.2 provides a q-q plot of the quantiles of the observed residuals from the linear model in Table 3.2 against the expected quantiles under a normal distribution. As can be seen, there is substantial and patterned variation around the estimated regression line or general tendency. But are these residuals organized in a way that is dependent on the interdependencies of the observations? Both Figure 3.2 and the density plot in Figure 3.3 show convincingly that the residuals are not distributed normally. The normal distribution is shown as a solid line in Figure 3.3. The residuals have tails that are far too large, and the residuals are skewed to the right considerably. It is clear in this example that the distribution of the residuals from the OLS regression reported in Table 3.2 is problematic. Although linear regression is relatively robust to violations against nonnormality and significance tests can be adjusted by alternative estimates of the standard errors (Lumley et al., 2002), the basic lack of fit of the model raises problems about whether the model specification itself can be trusted. These residuals suggest that the underlying systematic model does not capture the relationship between democracy and

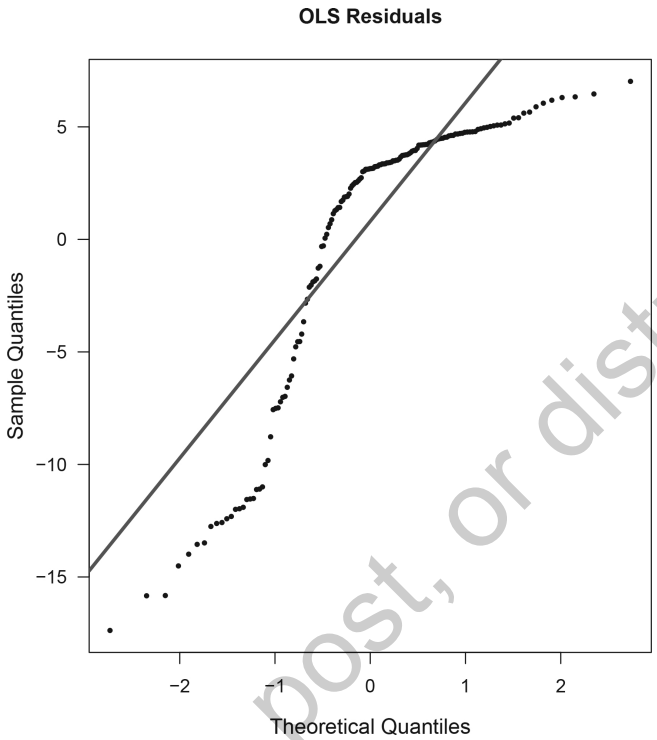

Figure 3.2. Q-Q Plot Examination of the Normality of the Residuals From the OLS Regression of GDP (Logged per Capita) on the Polity Score

Note. Normally distributed residuals would fall along the solid line. The actual obtained results are shown as dots, and these do not generally fall close to normal distribution.

Source: Created by Ward and Gleditsch.

economic output very well, potentially in part as a result of dependencies among the data—specifically clustering of similar values. For example, it may be that countries exert influence on each other beyond their individual income in ways that produce such results.

Introducing Spatial Dependence

One possibility for explaining these results is that, in addition to characteristics of individual countries, the prospects for democracy in one country are not independent of whether neighboring countries have

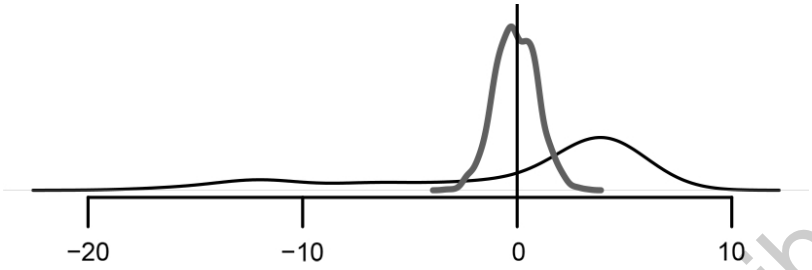

Figure 3.3. Density Plot of the Residuals From the OLS Regression of GDP (Logged per Capita) on the Polity Score

Source: Created by Ward and Gleditsch.

democratic institutions or not. During the Cold War, Soviet intervention enforced socialist rule in many states in Eastern Europe. Moreover, democratic transitions in many Latin American states appear to have been influenced by processes in other countries (Gleditsch and Ward, 2006). Looking at the data organized alphabetically, it would be hard to identify easily whether there are any pockets or regions of similar regimes beyond what we would expect from GDP per capita. Even with the information sorted on salient features for comparison, careful analytical study may be required to identify various kinds of patterns.

Exploratory examination of plausible spatial (and spatial-like) clustering may be important in a variety of situations, revealing aspects of social interaction that are missing from unconnected displays. Potentially unobserved clusters can influence our understanding of what is actually occurring in the part of the model we think we *do* understand. Before we turn to an examination of how to take spatial correlation into account, we explain a bit more about why it is important to do so.

Even if an analyst simply wants to compare means and construct classical statistical tests, such as difference of means tests, if the data are spatially correlated this becomes problematic. Consider a one-sample *t*-test on variable *y* defined as

$$t = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{\sigma}{\sqrt{n}}}.$$

If there is a correlation among observations that are near one another temporally or spatially (first-order serial correlation), then the actual standard error will be larger for positive values of serial correlation (and smaller for negative values). Researchers tend to be sensitive to the

problem of serially correlated observations over time but often neglect the fact that the same problem will apply for serial correlation across observations at the same point in time. Using the unadjusted estimate of the variance will result in having a t-value that is larger than warranted. This increases the chance of making a Type I error, even for situations in which there is only a small amount of spatial autocorrelation and abundant observations.

In short, because of serial, spatial correlation among the observations —for whatever reason—classical tests are biased in terms of accepting the hypothesized substantive account, even when it is untrue. Assuming that the data are spatially dependent such that the dependence is inversely proportional to the distance between observations, ρ represents the resultant, first-order spatial correlation. This correlation measures how similar neighbors are on some measured attribute. As a result of this correlation, the true standard error of the data is given approximately by

$$\sigma_{\overline{y}} \approx \sqrt{\frac{1+\rho}{1-\rho}} \frac{\sigma}{\sqrt{n}}.$$

A simple way to understand the impact of spatial correlation is to imagine a variable y observed on n observations: $y_1, y_2, \ldots, y_{n-1}, y_n$. In many situations, we think of these observations as being independent of one another and each identically distributed, typically from a normal distribution of unknown mean μ and variance σ^2. The typical estimator of μ is

$$\overline{y} = \sum_{i=1}^{n} y_i / n.$$

Since the observations are thought to arise from a normal distribution, inference depends on y and σ. The 95% confidence interval is given as $\overline{y} \pm 1.96\sigma/\sqrt{n}$. If there is spatial correlation among the y_i that is greater, the closer the observations y_i and y_j are to each other spatially, then as Cressie (1993, p. 14) shows, the covariance for positive values of ρ will be

$$\text{cov}\,(y_i, y_j) = \sigma^2 \times \rho^{|i-j|} \tag{3.2}$$

and the variance is

$$\text{var}\,(\overline{y}) = n^{-2} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}\,(y_i, y_j) \right\} \tag{3.3}$$

which expands to

$$= \{\frac{\sigma^2}{n}\}\left[1 + 2\{\frac{\rho}{1-\rho}\}\{1 - \frac{1}{n}\} - 2\{\frac{\rho}{1-\rho}\}^2\frac{1-\rho^{n-1}}{n}\right].$$

The factor $\left[1 + 2\{\frac{\rho}{1-\rho}\}\{1 - \frac{1}{n}\} - 2\{\frac{\rho}{1-\rho}\}^2\frac{1-\rho^{n-1}}{n}\right]$ essentially is the discount on the number of observations that is imposed by spatial correlation, which does not disappear in large samples.

If $n = 10$ and $\rho = 0.26$ (as in Cressie's example), then the discount is about 40%. This means that 10 spatially correlated observations have the same precision as about six independent observations. This in turn implies that ignoring the spatial correlation leads to a confidence interval that is far too small when there is positive spatial correlation among observations. In general, ignoring spatial dependence will tend to underestimate the real variance in the data. Thus, for a sample of 158 observations on GDP, the 95% confidence band under an assumption of normality would be $\frac{1.96\times\sigma}{\sqrt{n}}$, but if there were a spatial correlation of 0.65—the actual value of $\hat{\rho}$ for GDP from the above example—the correct confidence interval would be approximately 4.22 instead of 1.96, more than twice as large. In the case of the level of democracy, $\hat{\rho}$ is 0.47, which leads to a 95% confidence band that is $\frac{3.26\times\sigma}{\sqrt{n}}$, which is almost 70% wider.[2]

If there are different forms of spatial correlation, then different specific adjustments may be required, but the general point is that if there is positive spatial correlation, the sample mean will have less precision. As a result, the null hypothesis will frequently be rejected when it is true. It is unwise to rely on statistical tests that perform well in independent and identically distributed (*iid*) samples if the underlying data are spatially (inter)dependent. Schabenberger and Gotway (2005) illustrate this relative excess variability of the least squares estimator for different levels of autocorrelation in different sample sizes. For $\rho > 0$, this excess variability rises with n such that with $\rho = 0.9$, the excess variability is approximately 14 when the sample size approaches 50. The important point is that spatially correlated data will wreak considerable havoc with statistical tests designed for *iid* data, leading researchers to reject the null hypothesis because the standard tests underestimate the variability.

In Latin America, most states are democracies in 2015, despite large differences in their GDP per capita. By comparison, in the Middle East,

[2] Even for the mean, Grenander (1954) illustrates that the minimum unbiased estimator should not ignore the value of correlated observations: $\hat{\mu} = \frac{\left[y_1 + (1-\rho)\sum_{i=2}^{n-1} y_i + y_n\right]}{[n-(n-2)\rho]}$.

most states are autocratic, despite having GDP per capita levels that are consistently higher than the world average. Indeed, mapping these attributes suggests that both democracy and GDP per capita display spatial clustering (images not included). In many cases, visualization and mapping reveal structure in the data that is not readily available from looking at the data in tabular format.

The clustering of regime types around the world is evident in this display. There are low levels of democracy clustering in the Middle East and South and East Asia, but Latin America, Europe, and North America have higher levels of democracy.

Measuring Spatial Association and Correlation

Unfortunately, just as patterns may be ignored in a data matrix, humans are adept at seeing structure when there really is none. As such, it is useful to have more formalized ways of evaluating whether observations are spatially clustered or related across some forms of ties between observations. We turn to formal exploratory tools in the next section.

Exploring such associations, however, requires that we have some idea about which observations are likely to be related to one another. For a set of n units, each observation i can be potentially related to all the $(n - 1)$ possible units, but in practice, however, we can usually assume that some interactions or ties are more important than others. The network or structure between units that we are interested in must generally be specified *prior* to analysis of dependence between observations. The techniques that we explore here usually start from a graph or list L of relations between connected observations. For many purposes, it is practical to use a matrix to represent the connectivities between observations. For example, we can define a binary matrix $\mathbf{C}$ that specifies connectivities between individual observations. We have an entry $c_{ij} = 1$ if two observations i and j are considered connected, $c_{ij} = 0$ if not.

The basic ideas of measuring spatial associations and correlations can be thought of as cross-product statistics, following Hubert et al. (1981), which cross-multiply a measure of spatial proximity with a measure of the similarity of values on some particular attribute.[3] Let S_{ij} be some

[3] In the context of spatial point processes, these are sometimes known as join-count statistics, since they count the number of neighboring points with similar attributes.

measure of the spatial proximity of two observations i and j and let U_{ij} be the similarity on some underlying variable of concern. Cross-product statistics will have the *general* form

$$\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}U_{ij} \qquad \forall \quad i \neq j.$$

If U_{ij} defines similarity as a mean normalized cross-product on the underlying variable, say $[(y_i - \bar{y})(y_j - \bar{y})]$, then with appropriate scaling, summing this product over all observations yields a measure of spatial correlation known as the Moran's $\mathcal{I}$ statistic. If U_{ij} is defined as a squared difference, such as $(y_i - y_j)^2$, the resulting statistic is known as Geary's $\mathcal{C}$. We primarily focus on Moran's $\mathcal{I}$, which we define in more detail subsequently.

For example, spatial association in the case of measures of democracy would join a measure of how close countries were to one another in terms of some spatial measurement, such as whether their outer boundaries are within 200 kilometers of one another, with a measure of the similarity of democracy scores for each pair of countries examined. These statistics are useful as heuristics for identifying spatial patterns. Perhaps they are most useful as a diagnostic heuristic for examining the residuals from modeling exercises in which it is believed that there is no (remaining) spatial patterning not accounted for by the model used.

The first task in formally assessing such correlations is to specify the interdependencies among data. This requires developing a list of which observations are connected to one another.[4] This is an important step, but one that we will only illustrate here. Linkages might be established by physical distance, say the distance between capital cities, as in our example. However, other transmission mechanisms such as the density of transportation networks via roadways, trains, waterways, and air carriers may be a better indicator of connection in particular circumstances. Similarly, instead of capital city distances, scholars have used the length of the border between neighboring countries, for example, as a measure of interaction opportunities among adjacent countries.

[4] There are different types of spatially organized data. The data we are exploring are often called areal data, or (irregular) lattice data—deriving from a lattice of field experiments. There are problems with each of these terms, and throughout this monograph these are denoted as *regional* data. We do not examine any approaches that explicitly deal with individual points.

Weidmann et al. (2010) developed a database and an $\mathcal{R}$ package (cshapes) that has shapefiles for all the countries in the world, back to 1946. This is a very flexible $\mathcal{R}$ package that allows one to select a set of countries and a date, and the appropriate shapefiles are provided (Weidmann and Gleditsch, 2015). At the same time, it allows the specification of a distance matrix for the chosen set of countries. This package computes a distance matrix for the given data. It can compute different types of distance matrices: capital distances, centroid distances, and minimum distances between polygons.

A subset of these data are portrayed in Tables 3.3 and 3.4 in two ways: as a list and as a matrix. Many computer programs organize large

Table 3.3. A List Representation of Connections for Eight European Countries

List Format	
Country	Connections
Denmark	Germany, Norway, Sweden
Finland	Norway, Sweden
France	Germany, Italy, UK
Germany	Denmark, France, Italy, Sweden
Italy	France, Germany
Norway	Denmark, Finland, Sweden
Sweden	Denmark, Finland, Germany, Norway
UK	France

Table 3.4. (Adjacency) Matrix Representation of Connections for Eight European Countries

Connectivity Matrix Format								
	Denmark	Finland	France	Germany	Italy	Norway	Sweden	UK
Denmark	0	0	0	1	0	1	1	0
Finland	0	0	0	0	0	1	1	0
France	0	0	0	1	1	0	0	1
Germany	1	0	1	0	1	0	1	0
Italy	0	0	1	1	0	0	0	0
Norway	1	1	0	0	0	0	1	0
Sweden	1	1	0	1	0	1	0	0
UK	0	0	1	0	0	0	0	0

matrices as lists, since it allows a more efficient storage of information, allowing only the nonzero elements to be included in memory. Indeed, for small subsets, it is easier, perhaps, to derive spatial characteristics and record them as lists of connections. However, each list can be converted easily into a square matrix that portrays the observations along the rows and columns and the linkages in the interior of the matrix. A matrix representation is also helpful for defining certain variables or measures reflecting spatial structures and variation. Table 3.3 presents a set of connectivity data as a list; Table 3.4 illustrates the corresponding binary matrix **C** of these connections.

These data can also be presented as a simple network graph (see Figure 3.4). Such graphs are illuminating, but they quickly become convoluted, crowded, and difficult to read when the number of nodes is high.

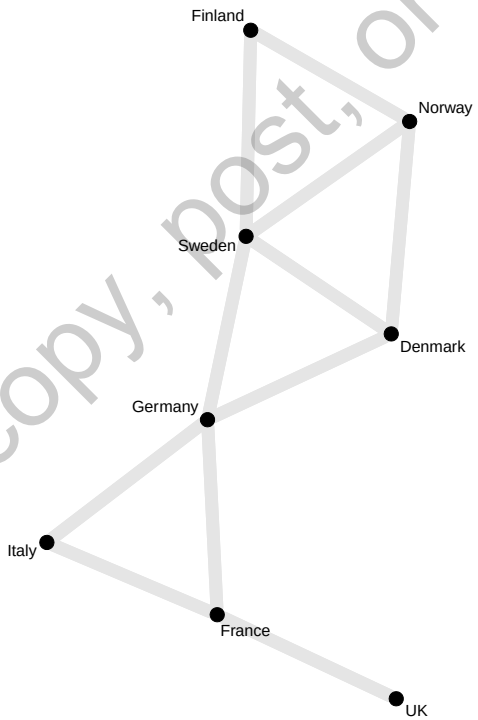

Figure 3.4. A Simple Network Visualization of the Linkages Among Eight European Countries

Source: Created by Ward and Gleditsch.

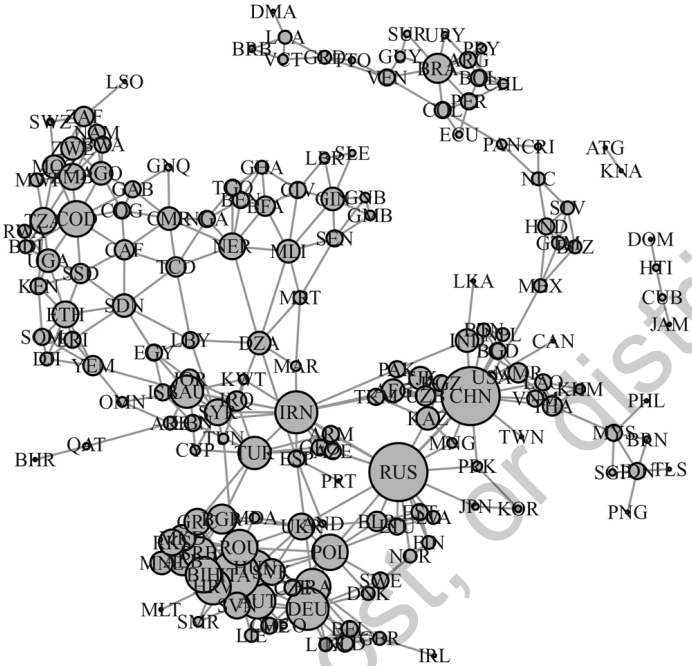

Figure 3.5. A Simple Network Visualization of the Linkages Among 195 Countries in cshapes in 2015

Note. Adjacency set to true for distances less than 200 kilometers. The size of vertices is proportional to outdegree.

Source: Created by Ward and Gleditsch.

The network map of the links between all 195 countries shows the crowding, as many countries have a large number of connections (Figure 3.5). Both Russia and China have 19 other countries to which they are connected (by a minimum distance of 200 kilometers). Such visual network representations may be a useful way to examine some data sets, especially those that are smaller or very much bigger.

Once we have a potential network of connections between observations specified by a list L or a connectivity matrix $\mathbf{C}$, we can explore whether the values on a variable of concern, which we denote here as y, are similar across connected or neighboring observations. One way to do this would be to look at whether two connected observations i and j tend to be similar to each other, for example by determining whether high or low values for i tend to go together with high or low values for

j. But *i* is usually connected to many observations, and we do not have spatial clustering unless it is similar to many of its neighbors. To combine information about the connected observations, we usually assume that all neighbors carry equal weight and that the weight of each is proportional to 1 over the total number of connectivities. The main goal of getting a *spatial lag* is to derive an average value that exists in a neighboring region. What is the average value of democracy in the neighbors of the United States? What is the average value of GDP per capita of Ghana's neighbors? Are these average values of neighboring observations correlated with each country's own score on democracy or GDP per capita? We present a heuristic statistic for gauging this, a statistic that measures the spatial correlation. In much the same way that a researcher might generate the correlation matrix among independent variables, this spatial correlation might also provide heuristic information about the observed data.

Let y_i^s denote the mean or average of y across all connected observations, or the "lag" of y over space. Matrix representation makes it easier to see the construction of the spatial lag y_i^s from y and the connectivity matrix **C**. We can create a row-normalized connectivity weight matrix **W**, where each row sums up to 1 by dividing each row vector $c_{i.}$ of the binary connectivity matrix **C** by the total number of links $\sum c_{i.}$. An example is given in Table 3.5.

In this context, the scalar $y_i^s = c_{i.}y$ calculates (by summing) the average or mean across all neighboring observations of one unit *i*. This is often referred to as the *spatial lag*. The relationship $y^s = \mathbf{W}y$ reminds us how each y_i^s is related to values of y for other states and the connectivity weights $w_{i.}$. Table 3.6 presents the 10 largest positive and negative spatial

Table 3.5. Row-Standardized Connectivity Matrix for a Subset of Eight European Countries

	Denmark	Finland	France	Germany	Italy	Norway	Sweden	UK
Denmark	0	0	0	1/3	0	1/3	1/3	0
Finland	0	0	0	0	0	1/2	1/2	0
France	0	0	0	1/3	1/3	0	0	1/3
Germany	1/4	0	1/4	0	1/4	0	1/4	0
Italy	0	0	1/2	1/2	0	0	0	0
Norway	1/3	1/3	0	0	0	0	1/3	0
Sweden	1/4	1/4	0	1/4	0	1/4	0	0
UK	0	0	1	0	0	0	0	0

lags for the democracy variable. Bahrain has a democracy score of −8, for example, but is surrounded by neighboring countries that all have the maximum negative democracy score, −10. Ireland and Portugal, on the other hand, have the highest possible democracy score as do all their neighbors.

Table 3.6. Democracy Data (PITF: 2015)

Country	Polity	Polity, Spatially Lagged
Canada	10.00	10.00
Ireland	10.00	10.00
Belgium	8.00	10.00
France	9.00	10.00
Switzerland	10.00	10.00
Portugal	10.00	10.00
Germany	10.00	10.00
Czech Republic	9.00	10.00
Sweden	10.00	10.00
Denmark	10.00	10.00
Saudi Arabia	−10.00	−5.00
Iraq	6.00	−6.00
Cambodia	2.00	−6.00
Sudan	−4.00	−7.00
Jamaica	9.00	−8.00
Bahrain	−10.00	−9.00
Qatar	−10.00	−9.00
United Arab Emirates	−8.00	−9.00
Oman	−8.00	−9.00
Dominican Republic	8.00	−10.00

Note. Top and bottom 10 countries in terms of spatially lagged Polity scores.

Measuring Proximity

For many social scientists, developing a measure of the proximity of units being studied is perhaps the most important step in spatial analysis. What is distance, in a social context? While many physical scientists will be able to use a strict measure of geographical or Euclidean distance to gauge how close trees are to one another, for example, this issue is considerably

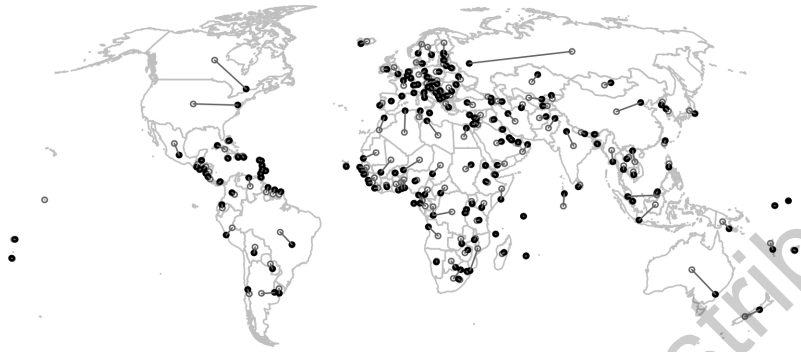

Figure 3.6. Map of Centroids (Open Circles) and Capital Cities (Filled Circles)

Note. The United States, Canada, Russia, China, and India all have large distances between these two locations. Norway has a centroid outside its boundaries.

Source: Created by Ward and Gleditsch.

more complicated for many social science analyses. How close are, for example, the United States and Mexico? If we use a strict contiguity measure, they are contiguous neighbors since they share a land border. But Canada also shares a land border with the United States. Does this imply that it is equally close to the United States? The straight-line distance from Washington, D.C., to Mexico City is approximately 3,000 kilometers, while the distance from Washington, D.C., to Ottawa is about 700 kilometers. We might use the length of borders between countries, or the distances between the average of the 10 largest population centers in each country. Figure 3.6 illustrates the difference between these two specifications. In some countries, the centroid is quite distant from the actual capital city, but in small countries this cannot be the case. China, Canada, Russia, Australia, and the United States are examples that illustrate the distance between these two locations. By contrast, in North and South Korea there is little distance between the centroids and the capital cities.

Another important issue in applied work is how to deal with missing spatial data. Imputation may be one approach, though other alternatives exist (Griffith, 2003). A real problem is that social science data are frequently missing, but rarely randomly missing. In nonspatial applications, this may be handled in the standard fashion—by imputation or, more frequently, by deletion of observations with missing information.

However, in the spatial framework, such missing data may create *holes* in the spatial representation and undermine establishing a salient and complete representation of the spatial proximities. Another problem that can occur in some kinds of spatial setups is that some observations will not be linked to other observations. For example, New Zealand is not within 200 kilometers of any other independent Polity. Two strategies are widely employed to circumvent these situations. Islands isolates are often deleted from the analysis, since at a substantive level they are not *connected*, and thereby will not affect other observations via the spatial process being studied. More prosaically, deleting them will purge the resulting spatial weights matrix of certain singularities (rows and columns composed entirely of zeros). A second strategy is simply to choose the *nearest* or most plausible neighbors for the islands, linking Australia and New Zealand as neighbors, for example, even if all other linkages are set for 200 kilometers. More generally, one can use nearest k neighbor distances for all units.

Above we have suggested two basic metrics for measuring distance, but this just scratches the surface. This metric of distance could be thought of in terms of average travel times, the number of mobile phone conversations between each pair of points, the amount of tourism from each point to every other location, or any variety of different measures of distance and interactions. Countries that have a large amount of commerce with each other, for example, can be thought of as economically "close" (Lofdahl, 2002). Griffith (1996) offers some ideas about how such measures can and should be developed.

It would seem natural to estimate the similarity between states' own level of democracy and the levels of their neighbors by the correlation between y and y^s. The linear association between a value and a weighted average of its neighbors is known as the Moran's $\mathcal{I}$ statistic (Moran, 1950a,b), a global correlation of the values of an observation with those of its neighbors. The generalized Moran's $\mathcal{I}$ is given by a weighted, scaled cross-product:

$$\mathcal{I} = \frac{n \sum_i \sum_{j \neq i} w_{ij} (y_i - \bar{y}) (y_j - \bar{y})}{\left(\sum_i \sum_{j \neq i} w_{ij} \right) \sum_i (y_i - \bar{y})^2},$$

where w denotes the elements of the row standardized weights matrix $\mathbf{W}$ and y is the variable of concern.

If the observations of y are *iid*, then $\mathcal{I}$ can be considered normal (asymptotically) with a mean that is $\frac{-1}{n-1}$. The variance of Moran's $\mathcal{I}$ is

then given by

$$\text{var}(\mathcal{I}) = \frac{n^2(n-1)\frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})^2 - n(n-1)\sum_k(\sum_j w_{kj}+\sum_i w_{ik})^2 - 2(\sum_{i\neq j} w_{ij})^2}{(n+1)(n-1)^2(\sum_{i\neq j} w_{ij})^2}.$$

If the variable of concern is standardized as z_i, Moran's $\mathcal{I}$ is simply

$$\mathcal{I} = \frac{1}{2}\sum_{ij} c_{ij} z_i z_j, \quad \forall \quad i \neq j.$$

The Moran's $\mathcal{I}$ statistic is often used as a test of spatial correlation by constructing a Z-score with the mean and variance components.

Moran's $\mathcal{I}$ does not really have a fixed metric, and its expected value is $-1/(n-1)$ rather than 0. However, the Moran's $\mathcal{I}$ statistic can be given a graphical interpretation that helps convey how spatial association among individual cases will give rise to different values of the statistic. Consider a scatterplot of $\tilde{y}$ against its average among neighbors' $\tilde{y}^s$ (we use a standardized $\tilde{y} = [y - \bar{y}]/\text{sd}[y]$ so that the value has a mean of 0 and a standard deviation of 1). In this plot, the distribution of observations in the four quadrants around the mean of $\tilde{y}$ and $\tilde{y}^s$ captures a picture of the spatial association of the variable y. If there is no spatial clustering or association in y, the individual values of y^s should not vary systematically with y. However, if there is a positive spatial association, individual observations that have values above or below the mean on y should also be low and high, respectively, on y^s, or among proximate countries. The bulk of the cases should fall in the South-West and North-East quadrants where units are similar to their neighbors, and we should have few observations in the North-West or South-East quadrants. If we fit a regression line to this scatterplot, its slope is the Moran's $\mathcal{I}$ correlation given the original variable y and the connectivity list L or matrix $\mathbf{C}$.

Figure 3.7 provides a stylized plot illustrating the Moran's $\mathcal{I}$ statistic and the interpretation of a scatterplot of a variable and the first-order spatial lag. The slope of the regression line is the average spatial correlation in the data; it is the Moran's $\mathcal{I}$ statistic. To illustrate this concept, we present a plot of the residuals from the OLS regression against the spatial lag of those residuals, where the weights were created by a 400-kilometer distance band. This kind of plot is known as an Anselin-Moran plot.

The computed Moran's $\mathcal{I}$ statistic for these OLS residuals is 0.85, with a variance of 0.003. This has an associated standard score of 7.803 that

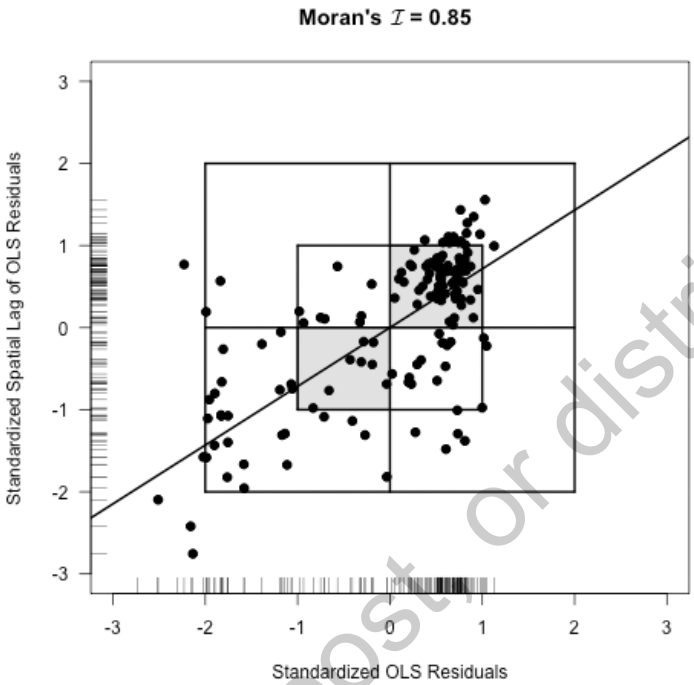

Figure 3.7. Anselin-Moran Plot of OLS Residuals

Note. Polity scores are modeled as a function of logged GDP per capita. Data are from 2015 for the Polity scores. GDP is from 2014. Spatial weights are dichotomized at 400 kilometers.

Source: Created by Ward and Gleditsch.

is much larger than $\frac{-1}{180}$ and has an associated p value that is ≈ 0. This tells us that the OLS results, which assume independent observations, are strongly affected by the spatial clustering in the dependent and independent variables. As a result, they are likely to be misleading for both the statistical and substantive inferences that we may wish to draw about the relationship between democracy and its social requisite of wealth, as captured in GDP per capita. What is pretty clear is that the residuals are skewed to the left, having more negative than positive values (when standardized). There is a strong cluster of observations between 0 and 1, but very few large positive residuals. Thus, the model tends to underpredict the Polity score by a substantial amount. Some of this underprediction may be because of the spatial correlation that characterizes these data, and indeed the residuals of the OLS estimates.

Estimating Spatial Models

What might constitute a simple set of steps for spatial analysis?

1. Map the data, especially the dependent variable. This can be done in a variety of contexts, ranging from spreadsheet plug-ins, map mash-ups, and GIS packages, but we find it best to undertake this in the context of a platform that will permit statistical analysis of these data. We illustrate the use of $\mathcal{R}$ libraries, especially maptools and spdep for constructing simple maps of the distribution of variables.

2. Determine if there is some discernible spatial correlation in the dependent variable. For most applications—that is, not point processes—that we consider, this means calculating the Moran's $\mathcal{I}$ statistic, to gauge the magnitude of spatial correlation. Analysts may in some cases wish to proceed to examine and plot/map each observation's contribution to spatial correlation, through a local indicator of spatial association (aka LISA). We do not pursue this in any detail. See Gleditsch and Ward (2000), Anselin (1995), and Ord and Getis (1995) for further discussion and examples.

3. Precisely incorporate these spatially lagged variables into a statistical framework and examine the resultant residuals for remaining spatial association. Subsequent chapters detail several common specifications of spatial models.

4. In addition to employing the OLS model heuristics to gauge the fit of the model and the degree of uncertainty in the estimated parameters, the equilibrium impact should be computed and examined. This means teasing out the equilibrium, feedback implications of the estimated spatial model for the dependent variable, and gauging their plausibility.

We now turn to an illustration of these steps in the context of our running example.

Mapping the Data and Constructing the Spatial Weights Matrices

We have illustrated mapping of data with the democracy scores for 195 countries in 2015. In this subsection, we illustrate the use of mapping with the residuals from the OLS model. We have also shown that the residuals from the regression of democracy on income display spatial association. We calculated the Moran's $\mathcal{I}$, using a 400-kilometer distance

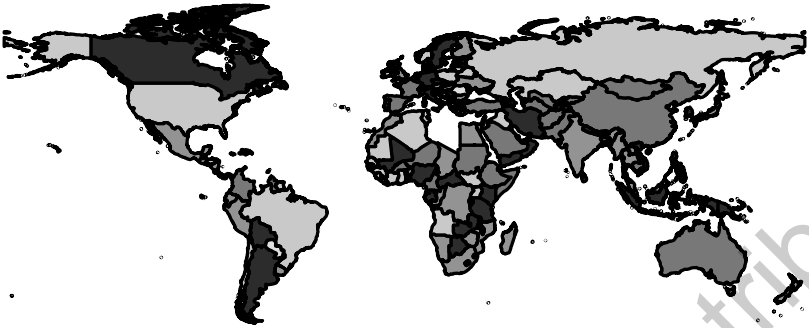

Figure 3.8. Choropleth of Local Indicators of Spatial Autocorrelation—the Localized Moran's $\mathcal{I}$

Note. Spatial weights dichotomized at 400 kilometers.

Source: Created by Ward and Gleditsch.

band from the outer boundaries of countries to determine each country's "neighbors." As previously reported, the Moran's $\mathcal{I}$ in this case had a value of 0.85. This is significant in a classical sense, and allows us to be confident that the spatial patterns are actually influencing the regression results in a substantial fashion, that is, introducing bias into estimates and standard errors.

Looking for Spatial Patterns

We also illustrate the construction of the so-called Anselin-Moran plot, adapted from the works of Anselin (1996) and Shin (2001). This plots the standardized value of each input variable—in this case the residuals—against its spatial lag or average value for its connected observations. The shaded boxes indicate concordant observations where a value above the mean of the residual is accompanied by a positive value for its neighbors. The axes contain a "rug plot," indicating the distribution of the variables.

In addition to mapping the first-order spatial lag of democracy, it is also useful to map the contributions of each observation to the global Moran's $\mathcal{I}$ statistic. This quantity is known as the LISA statistic. Herein we standardize these, and provide a mapping that is displayed in Figure 3.8. The local Moran is developed in Ord and Getis (1995), Anselin (1995, 1996), and Getis and Ord (1996).

This map illustrates which countries have the most unusual situations in terms of their neighbor's level of democracy. Southern and

Western Africa fall into this category, as does India. Jamaica, Bosnia-Herzegovina, Haiti, Swaziland, and Mongolia have the largest negative values of the localized Moran's $\mathcal{I}$. Bahrain and Qatar have the highest LISA values. These have been colored based on the LISA values, which have been cut at the first and third quartiles, along with the median.

Summary

Having first carefully examined our data and visual displays of these data, we explored the results of an OLS regression positing that the level of democracy is a linear function of wealth, measured as logged GDP per capita. We inspected the residuals from this regression and found convincing evidence that the residuals appear to display spatial clustering, violating the regression assumption that the error terms of individual observations can be considered independent of one another. As such, OLS assuming independent observations will not be a compelling method for analyzing the relationship between income and democracy. More fundamentally, a model assuming independent observation where only income matters for democracy ignores important features of obvious geographical clustering. We have also shown how maps and simple statistics can be used as informative heuristics to assess the extent and nature of spatial clustering.

Even if one is not interested in regression analysis, there is room for examining spatial patterns in social science data. We show that whether one is going to simply do a test of means or use a regression approach to examining data that are spatially organized, failure to take the spatial correlation into account will lead to incorrect inferences that are generally biased away from rejecting the stated hypotheses.

Cartographic displays of correlational data provide an exploratory heuristic for determining the presence of spatial patterns, patterns that can complicate statistical inference. We turn next to estimation of regression models with spatially lagged dependent variables, an approach that can take spatial dependencies explicitly into account.