International Interactions: Empirical and Theoretical Research in International Relations
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gini20

The Geography of the International System: The CShapes Dataset

Nils B. Weidmann [a], Doreen Kuse [b] & Kristian Skrede Gleditsch [c]
[a] Princeton University
[b] Dresden University of Technology
[c] University of Essex
Published online: 09 Mar 2010.

International Interactions, 36:86–106, 2010
Published with license by Taylor & Francis
ISSN: 0305-0629 print/1547-7444 online
DOI: 10.1080/03050620903554614

Routledge
Taylor & Francis Group

The Geography of the International System: The CShapes Dataset

NILS B. WEIDMANN
Princeton University

DOREEN KUSE
Dresden University of Technology

KRISTIAN SKREDE GLEDITSCH
University of Essex

We describe CShapes, a new dataset that provides historical maps of state boundaries and capitals in the post-World War II period. The dataset is coded according to both the Correlates of War and the Gleditsch and Ward (1999) state lists, and is therefore compatible with a great number of existing databases in the discipline. Provided in a geographic data format, CShapes can be used directly with standard GIS software, allowing a wide range of spatial computations. In addition, we supply a CShapes package for the R statistical toolkit. This package enables researchers without GIS skills to perform various useful operations on the GIS maps. The paper introduces the CShapes dataset and structure and gives three examples of how to use CShapes in political science research. First, we show how results from quantitative analysis can be depicted intuitively as a map. The second application gives an example of computing indicators on the CShapes maps, which can then be used in

We are grateful to Luc Girardin, Håvard Hegre, and Jan Ketil Rød for their help and feedback during this project, and to Elin Arbin for advice on writing and presentation. Nils Weidmann acknowledges financial support from ETH Zurich (Research Grant TH -4/05-3). Kristian Skrede Gleditsch would like to acknowledge the support of the Economic and Social Research Council (RES-062-23-0259) and the Research Council of Norway (180441/V10).
Address correspondence to Nils B. Weidmann, Woodrow Wilson School of Public and International Affairs, Robertson Hall, Princeton University, Princeton, NJ 08544, USA. E-mail: nils.weidmann@gmail.com

statistical tests. Third, we illustrate the use of CShapes for generating different weights matrices in spatial statistical applications. All the examples can be replicated using the freely available R package and do not require specialized GIS skills. The dataset is available for download from the CShapes website (http://nils.weidmann.ws/projects/cshapes).

KEYWORDS distance metrics, GIS, historical boundaries, maps, spatial weights

The field of international relations centers around the state as a key unit for research (see Lake 2002; Kahler 2002), and comparative research usually conducts cross-country comparisons of state characteristics (see Doggan and Pelassy 1984; Bollen, Entwisle, and Alderson 1993). The last decades have seen a tremendous growth in the availability of empirical data at the state level, which in turn have been used in many empirical studies. However, most existing data on states and their characteristics often lack a clear spatial referent, and the specific units and boundaries that observations in a data set refer to are often unclear. Many common data sets contain inaccuracies or anachronisms that, we surmise, are not at all obvious to users and only become apparent upon close inspection. The Penn World Tables data, for example, impose current state/border configurations on all the historical data reported (Heston, Summers, and Aten 2006). This in turn means that historical observations for "Germany" prior to unification encompass population estimates for both the Federal and the Democratic Republics together, although the two were clearly administered as separate states. Moreover, data for "Russia" prior to the end of the Soviet Union are estimates of the Russian Federation only, although the former Soviet Republics clearly were part of an integrated state at the time.

Likewise, although the use of Geographic Information Systems (GIS) and maps have become increasingly common in international relations and comparative research, most standard software packages tend to have default global maps that conform to current state boundaries, without any additional information to reflect changes over time. As such, these maps do not incorporate changes in the configuration and shape of states (for example the secession of Bangladesh/East Pakistan from Pakistan in 1971). Not surprisingly, using current maps to represent historical time periods can introduce many inaccuracies in applied work.

In this paper we present the CShapes dataset, which contains a set of historically accurate country boundaries starting in 1945, as well as routines to perform useful operations on these in the R statistical package. Beyond displaying historically accurate maps, our package allows users to take advantage of geographical information for different purposes. This paper starts by discussing how existing research has taken into account the geography of the international system. We show that the development of multiple databases for this purpose leaves much to be desired with regards to data consistency and compatibility. We argue that a direct representation of state boundaries as electronic maps can improve the already existing numeric indicators. The paper proceeds with an introduction of our CShapes dataset that follows this approach. We discuss the coding scheme and the dataset structure. The remainder of the paper illustrates the use of CShapes with three applications: First, the creation of maps to visualize quantities measured at the state level; second, the computation of geographic indicators on CShapes maps; and third, the creation of weights matrices for spatial regression modeling.

THE GEOGRAPHY OF THE INTERNATIONAL SYSTEM

For the field of international relations, the geography of the international system is important for at least three reasons. First, it can tell us something about states. Country size, for example, is held to influence the likelihood that a state will see conflict (see for example, Fearon and Laitin 2003), or a phenomenon important to explain in its own right (for example, Lake and O'Mahony 2004). In order to compute geographic variables such as country size, we clearly need to know the spatial extent of states and their configurations. Second, geography provides the topology in which the states interact. This is reflected in research on territorial conflict, the diffusion of war, and the role of distance as a determinant for cooperative and conflictual interactions (for prominent examples, see Hensel and Diehl 1994; Siverson and Starr 1991; Gleditsch 2002a; Vasquez 1995; Gochman 1991). In order to understand the influences a state is exposed to, we need to know where states are located relative to each other, so that we can compute characteristics such as adjacency and distance between boundaries. Third, there are many important phenomena that do not necessarily follow state boundaries such as environmental influences, for which many data sources are now provided in geographically disaggregated form. Prominent examples include depositions of important environmental pollutants, such as sulphur and nitrous oxide (see for example, Sandler 1997). This also holds for rainfall, which some researchers have suggested is a key influence on economic growth in rain-fed agricultural societies, with the advantage that it can be assumed to be exogenous to conflict and hence be helpful as an instrument in assessing the effects of growth shocks on conflict (Miguel, Satyanath, and Sergenti 2004). Computing state-level estimates of these phenomena and assessing their variation within states required us to overlay accurate boundary maps.

A number of data sources have been compiled to reflect various geographical characteristics deemed important in theoretical studies on conflictual interactions. These include data on contiguity (Gochman 1991), minimum distances between states (Gleditsch and Ward 2001), shared boundary length (Furlong and Gleditsch 2003) and territorial change (Tir, Schafer, Diehl, and

Goertz 1998). However, previous data projects have emphasized a single aspect of geography, and have been collected independently of one another. Although these individual databases measure aspects of the same system, they lack any comparable underlying data. This is obviously a less than optimal strategy from an information representation point of view. First, it makes it difficult to compute comparable measures—how can we be sure that geographic variables in two different datasets refer to the same geopolitical entity? Second, these compatibility issues across datasets can make error checking and cross validation particularly difficult. Third, and most importantly, updating the datasets is inefficient and time-consuming. For example, with a newly released system membership list, one would have to update multiple databases at the same time. We believe that a more consistent approach to generating data on underlying geographic characteristics may be helpful.

All the above-described geographic variables depend on the same information, which is the underlying maps of countries and their international boundaries. Once these maps are available, very little is required for measures such as minimum or capital distance to be computed from this geographic information. In fact, the only step that needs human intervention is the creation of the electronic maps; the computation of the dependent information—distance, area, adjacency—can be left to a computer. This approach ensures that the variables are comparable, since they refer to the same geopolitical entities. Also, modifications in such a dataset can be done simply by changing the underlying cartographic data and rerunning the variable computation. The CShapes dataset and statistical package seeks to implement this approach and to develop a geographic database for international boundaries.

THE CSHAPES DATASET

The aim of the CShapes project is to provide a direct representation of state boundaries in a GIS dataset. However, there are a number of practical problems to be considered when developing historical maps of the international system. The most obvious decision that has to be made concerns membership in the international system, or the issue of determining and identifying what qualifies as an independent state. For the representation of state boundaries, we then need to identify the spatial extent of states that we want to include in our maps. Lastly, since we are aiming for a representation of state boundaries over time, we need to identify what constitutes a change in these boundaries. This section presents our coding decisions and introduces the CShapes data representation as a GIS dataset.

Defining States

Many different definitions have been proposed for the coding of geopolitical entities. The *International Organization for Standardization* (ISO) publishes the frequently used ISO 3166 country codes standard that defines numeric or alphanumeric codes for a global list of countries (International Organization for Standardization 2008). Another coding system for geopolitical units often used in geographic applications is the “Federal Information Processing Standard” (FIPS), number 10-4 (Information Technology Laboratory and National Institute of Standards and Technology 2008). The FIPS standard is issued by the US Federal Government and partly relies on ISO 3166. However, there are two problems associated with these systems, which makes them difficult to apply in the social sciences. First, there is a lack of clear rules regarding the inclusion of cases.[1] Second, none of the mentioned country coding systems includes a time dimension. ISO was first published in 1974 and is regularly updated, but there is no consistent list for dates earlier than this.

Research in political science has produced alternative lists of states or members in the international system, which address some of these problems. The first widely used state list was originally proposed by Russett, Singer, and Small (1968), and has subsequently become known as the *Correlates of War* (COW) system membership list (Correlates of War Project 2008). COW uses as its main criteria the recognition by the UK and France, membership in the League of Nations and the United Nations, and various ad hoc decisions. The COW coding scheme has been employed—in varying degrees of consistency—by a number of other data projects, such as the widely used Polity dataset on regime characteristics (Marshall and Jaggers 2008), and the Minorities at Risk database on discriminated ethnic groups (Minorities at Risk Project 2005). However, the COW list has been criticized for lacking face validity and problematic coding decisions. Alternative lists have been proposed, the most important of which, for our purposes, is the state list by Gleditsch and Ward (1999). Gleditsch and Ward (GW) consider a minimum population threshold of 250,000, whether self-declared states have territorial control and recognition by some states, even if they do not seek an active international role or membership in international organizations. Gleditsch and Ward also provide a supplementary list of microstates with populations less than 250,000 for researchers interested in including small formally independent states. A variety of data projects uses the GW list, for example the Uppsala/PRIO Armed Conflict Dataset (Gleditsch, Wallensteen, Eriksson, Sollenberg, and Strand 2002), or the “Scalar Index of Polities” by Gates, Hegre, Jones, and Strand (2006).

We decided to make CShapes compatible with the COW and GW lists as both of these sources are widely used in the research community. The differences between the two datasets mainly arise in the pre-World War II period, where the two often have dramatically different dates of independence

[1]For example, ISO 3166 includes the Svalbard islands as a separate unit. However, although Svalbard has a special status under the 1920 treaty, it falls under Norwegian sovereignty and is administrated by a governor directly appointed by the Norwegian government.

for many states, such as Canada or Iran/Persia. The differences are much smaller in the period after 1945, which is currently covered by CShapes. In this period, most of the discrepancies between the two data sources arise from minor differences in the start and end dates of states, as well as some major differences arising for microstates such as Andorra that often have a substantial period of formal independence prior to becoming system members in the COW list due to membership in the United Nations.

The Spatial Extent of a State

Once we use either the COW or the GW list to identify the relevant state entities to be coded in CShapes, we need to define their spatial extent explicitly. According to the classic Weberian notion of the state as having a monopoly on violence, we could define states boundaries by the territorial limits of control of the state. This criterion is difficult to apply in practice, however, as there is a lot of variation in the level of control that states actually exert over their territorial units. As of the time of writing, for example, the central government of Afghanistan does not actually exercise control over much of the territory for which it claims sovereignty.

Another possible criterion, conceptually similar to that of the COW list, would be to consider the internationally accepted boundaries of states. In practice, boundaries are often disputed, and many states often claim the same territory. Venezuela, for example, makes a claim to all of the territory west of the Essequibo River, which constitutes more than half of the territory claimed by Guyana, although they have so far not used force to seize the territory. In other cases such as Kashmir, China and Pakistan occupy territory seized by force from India, but without India recognizing their claim to the territory. Finally, many states formally recognize states that do not control the territory claimed, such as the Sahrawi Arab Democratic Republic in Moroccan-controlled Western Sahara.

For the CShapes project, we take a pragmatic approach and consider the conventional international boundaries recognized by most states and where another state is not clearly exercising control over the territory. Our coding approach is stepwise in that we first identify the relevant states in the system and then find spatial representations of their extent. We include only what we call the "core" territory of a state, and leave out the dependent territories such as colonies or militarily-occupied territories outside the recognized boundaries of states. As such, not every point on Earth will be included in a CShapes polygon at any given point in time, as there may be significant territories that do not fall within the core of a state. For example, sub-Saharan Africa in 1947 consists of only three polygons (Ethiopia, Liberia, and South Africa), since the rest of the continent was under colonial rule at that time. More recently, territorial units such as Antarctica are not coded as part of any state.

An alternative to our approach of proceeding from states to territories would be to determine for each individual territorial unit at different points in time whether it belongs to an independent state, and what its status is. Such information would certainly be desirable, but this approach is much more complex and would require a great deal more work and resources than we have at our disposal. However, our approach of starting with a positive list of states can certainly be extended to include territories other than the core if need be, and it would be possible to code the remaining parts of the globe outside the core territories into residual polygons based on the information in CShapes.

Territorial Changes

The shape and configuration of a state's core territory can change over time. We distinguish between two important cases: First, territorial changes may occur when states merge or dissolve, and second, state configurations may change in the absence of emergence or disappearance of states. One example of the first kind of territorial change is the dissolution of Czechoslovakia in 1992, when the two constituent members of the federation (that is, the Czech and Slovak Republics) emerge as new independent states. In CShapes, we see two new polygons replacing the former polygon for Czechoslovakia, but other polygons bordering Czechoslovakia are not affected. Conversely, German unification, where the former German Democratic Republic joined the German Federal Republic in 1990, implies a merger of two polygons, but also without any changes in other polygons. Examples of territorial adjustments unrelated to changes in the number of independent states include the entry of Newfoundland to Canada in 1948, or the incorporation of the Azouzou strip, previously held by Chad, into Libya in 1972.

Information on territorial changes related to system membership can be obtained from the state lists that CShapes relies on (COW or GW). Both lists provide entry and exit dates at the level of days for all independent states they contain. We aligned all the dates of territorial changes where states enter and leave the international system with the date given in the two state lists. For example, in line with both state lists, the boundaries of East and West Germany are coded as active until October 2, 1990. The new boundaries of the unified Germany replace the old territorial division beginning October 3, 1990. As we show below, CShapes contains separate entry/exit dates for the two state lists, since in a number of cases the two state lists have different dates of independence.

Changes of the second type—boundary adjustments without changes in system membership—are taken from the *Territorial Changes* dataset by Tir et al. (1998).[2] The dataset lists all territorial changes that involved at least

[2]Jaroslav Tir kindly provided a supplementary list of entity codes and names used in the original dataset.

one member of the international system in the period 1816–1996, including the exact dates. Each change is coded in a directional fashion, by listing the gainer and the loser of a particular piece of land. In order to make our effort of representing changes over space and time feasible, we had to limit the number of cases from the Territorial Changes dataset included in CShapes. In particular, we apply three restrictions. First, we select all changes that occurred during and after 1946, because CShapes covers only the post-World War II period. Second, CShapes aims to represent only the core territory of states, so we disregard all changes not affecting the homeland territory of a state. This selection was done using the "gaintype" and "losetype" variables in the Territorial Changes dataset. Third, we limit ourselves to major territorial changes in area and only code transfers affecting an area equivalent to 100 km × 100 km. Also, this selection was performed using the information in the Territorial Changes dataset by setting a minimum threshold of 10,000 in the "area" variable. This leaves us with a final set of 35 territorial changes relevant for CShapes. Note, however, that the Territorial Changes dataset has not been updated beyond the year 2000. Hence, for the 2001–2008 period CShapes includes only changes related to the emergence/disappearance of new states, but no boundary adjustments between existing states that may have occurred.

The Representation of State Boundaries in a GIS Dataset

Software and standards for the processing of spatial data have developed at an amazing speed over the last decade. These so-called geographic information systems (GIS) are also finding their way into the social sciences. Since the aim of CShapes is to give an explicit spatial representation of state boundaries, it is a straightforward decision to use a geographic data format for CShapes. GIS data come in two fundamentally different types. First, raster datasets divide the geographic space into equal-sized cells, and a particular value is stored for each of them. For example, territorial elevation is typically represented as a raster dataset with an elevation level for each cell. Second, vector datasets represent the geographic features of interest directly, either as points, lines or polygons. For example, in a dataset on rivers, we can store each river as a line represented by a number of connected points in the geographic space. We refer readers interested in more details about GIS tools and data to other specialized references such as Longley, Goodchild, Maguire, and Rhind (2005).

The GIS vector data approach is particularly suitable for CShapes, since we are dealing with a finite number of geographic entities (states in our case). In addition, vector data typically allow for additional, nonspatial information to be stored along with each geographic feature. This is done by providing a so-called "attribute table" in addition to the spatial features, for example, rivers or roads. Each feature is linked to a record in the

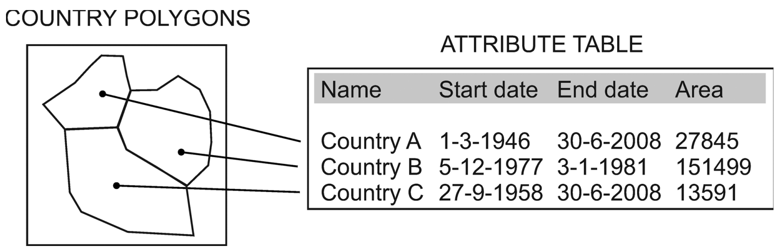

FIGURE 1 CShapes data representation as a GIS shapefile. Countries are represented by polygons, each of which is linked to a record in the attribute table where additional information about the country is stored.

attribute table, which contains supplementary information. In a river dataset, for example, the attribute table may contain the name of the river. In CShapes, we represent states as polygons, where each polygon is linked to the corresponding record in the dataset's attribute table. Figure 1 illustrates the dataset structure graphically.[3] Note, however, that with a simple polygon representation we would have to represent noncontiguous states with a set of polygons and their corresponding rows in the attribute table. This would lead to a very complex and error-prone dataset. Instead, we rely on an extension of the polygon vector format, the *multi-polygon* representation. Instead of linking each polygon to a single row, a *set of polygons* is represented by a record in the attribute table. This allows for a convenient data format where, for example, a country like Indonesia corresponds to exactly one record in the attribute table, even though it consists of 121 disconnected polygons.

For each territorial change, CShapes contains polygons that reflect the status quo before and after the change. For example, the breakup of a country A into countries B and C would be coded with three polygons: one representing the extent of country A before the dissolution, and two polygons giving the extent of countries B and C after the breakup. How do we code the temporal dimension of these changes? Whereas GIS vector data formats are perfectly suited for the representation of spatially explicit information, there is no support for linking this information to a particular point in time. We solve this problem by following the common approach of coding a "lifetime" for each state polygon. More precisely, each polygon has a *start date* and an *end date* stored in its attribute table, which indicates the period when it is active. For each territorial change, we assign a lifetime that *ends the day before the change occurs* to all polygons that reflect the status

[3]CShapes uses the shapefile format, a frequently used vector data format, originally developed by the Environmental Systems Research Institute (ESRI). A technical description is available online (http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf).

before the change. All polygons representing the status *after the change* are assigned a lifetime that starts on the day the change happened. In the above example, if we assume the breakup of country A to occur on a given day *d*, the lifetime of the polygon representing country A ends at time *d-1*, and the lifetime of the two successor polygons starts at *d*. If there is no information on the particular day a change occurred, we code the middle of the respective month.

Coding Procedure and Result

CShapes was created by starting with a 2006 GIS dataset and backdating it step by step, according to the changes given in the system membership and territorial changes datasets. The original dataset is provided by ESRI (2006) and contains the boundaries of states and dependent territories. We started by matching the entities in the original dataset to the state lists and deleted all polygons that did not match an independent state as coded by the state lists. We then worked our way backwards in time, creating new polygons for each relevant territorial change as introduced above. We consulted additional sources in all cases where the extent of the respective boundary change was not readily available. This was almost always the case for boundary changes of the second type (that is, adjustments without changes in system membership). The final dataset contains 241 polygons, where the vast majority of countries (177) are represented by one polygon. 15 countries have two polygons in the dataset, and 9 countries have three or more.

Figure 2 illustrates the result of the coding process for the breakup of Yugoslavia. We present the example as series of snapshots, taken at the end of each month when a territorial change occurred. Each panel shows the currently active polygons as solid lines. Note that because the breakup of the Federation occurred in a stepwise fashion with the former Republics seceding at different times, we need to code multiple polygons for what constitutes the core territory of the Federation. The lifetimes of these polygons are shown in the plot.

USING CSHAPES: APPLICATIONS

We believe that CShapes can be useful for a variety of applications in quantitative research, and in this section we present some examples to illustrate this. The first example shows how to use CShapes for creating map visualizations of variables of interest. Our second application deals with the computation of state-level geographic variables, which require information about the extent of states, as contained in our dataset. Lastly, we demonstrate the use of CShapes in the context of spatial regression models.

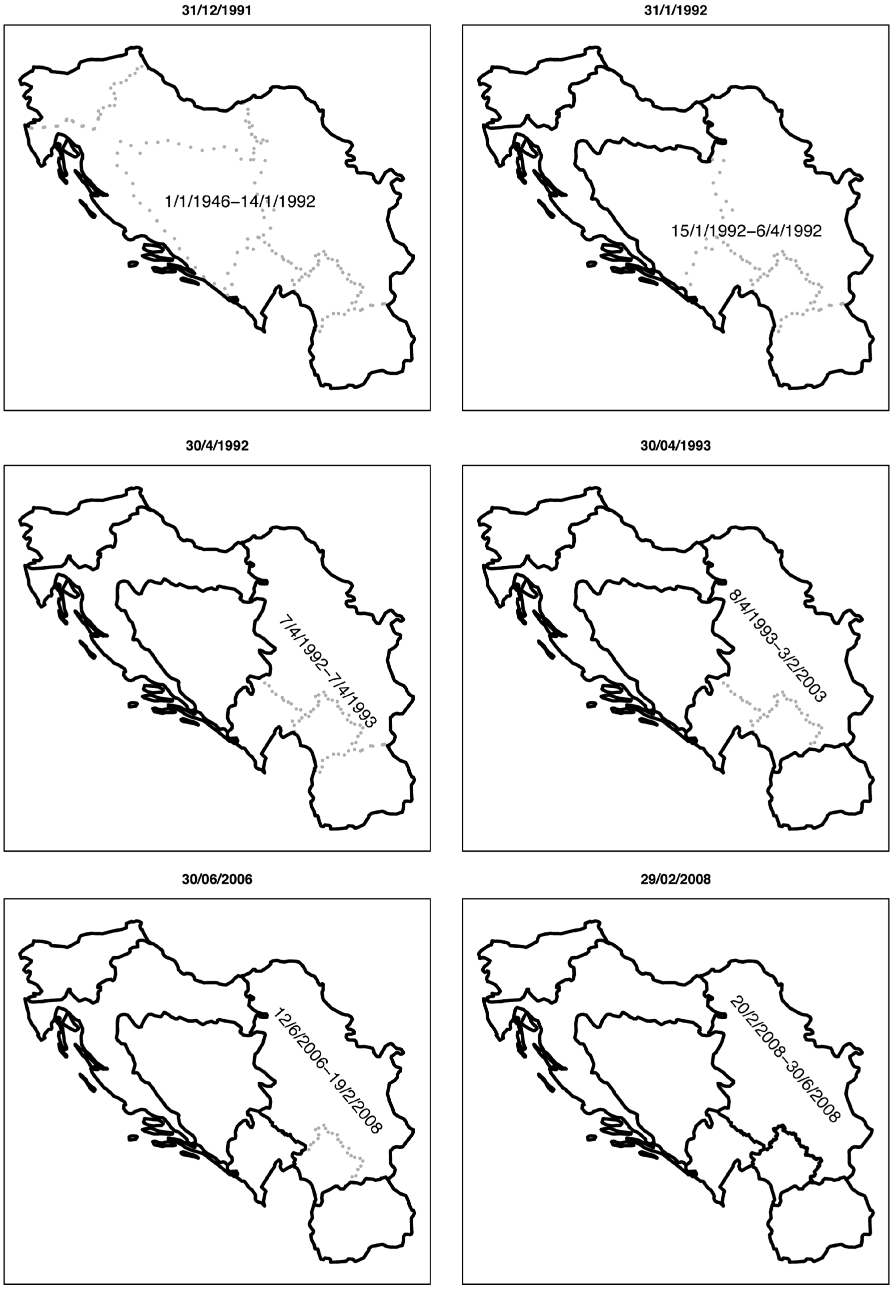

FIGURE 2 Example of the CShapes coding of Yugoslavia. The figure shows the break up of the Federation by a series of snapshots, aligned with the COW dates for when new states were established. Solid lines indicate polygons that are active at the given date. For illustration purposes, the date range given in the plot shows the lifespan of the currently active polygon that represents Yugoslavia (and later Serbia, COW code 345).

As described above, CShapes relies on a GIS vector data format. Although this data format by itself is perfectly adequate for the storage of spatial information, lack of GIS skills among researchers might limit its spread in the research community. We have therefore decided to distribute CShapes in two ways: the raw dataset itself in a vector data format, accompanied by a software package that connects our dataset to the R statistical toolkit, the latter providing functionality for processing spatial data (Weidmann and Gleditsch 2010).[4] This dual dissemination strategy builds on recent efforts to introduce new, open-source tools for spatial analysis in the social sciences (Rey and Anselin 2006), among them the R spatial extensions (Bivand 2006). The CShapes package for R allows scholars without a GIS background to get access to the information in CShapes and to perform a variety of tasks required to use the dataset for their own research. To illustrate this approach, the applications presented in the remainder of this section rely only on the CShapes R package and do not require any additional proprietary software.

Spatio-Temporal Mapping

Quantitative studies often rely exclusively on numbers and tables to communicate their results. Recent attempts have been made to find graphical ways to present the outcome of quantitative analysis (Kastellec and Leoni 2007). However, for many applications in international relations where states are the units of analysis, it might be beneficial to opt for a geographic way of showing results pertaining to these. This is a straightforward idea, but might be more difficult to implement in practice. What is required is a set of geographic entities that correspond to the commonly used units of analysis, and an easy way to link other variables of interest to these entities. CShapes makes it possible to do this, as it provides state polygons that are compatible with the commonly used state lists in political science.

We demonstrate the creation of spatial plots using the Fearon and Laitin (2003) model for the onset of civil war. Using Model 1 from their study, we compute the predicted risk of civil war onset and show its geographical variation.[5] Figure 3 shows the model predictions for 1985 and 1995, with darker colors corresponding to a higher predicted risk. We limit our presentation here to two years, but it could obviously be repeated for other points in time. These maps show how there are many countries with a low predicted risk of conflict, located mostly in Europe and the Americas, and that certain subregions contain the countries with higher risks of conflict. One difference between the maps is that the 1995 predicted conflict

[4]The R statistical package can be downloaded free of charge (http://www.r-project.org).

[5]In the original Fearon and Laitin model, missing values for GDP lead to many cases being dropped from the original analysis, including cases of civil war, such as Bosnia. Therefore, we replace the original GDP variable with a more complete version provided by Gleditsch (2002b).

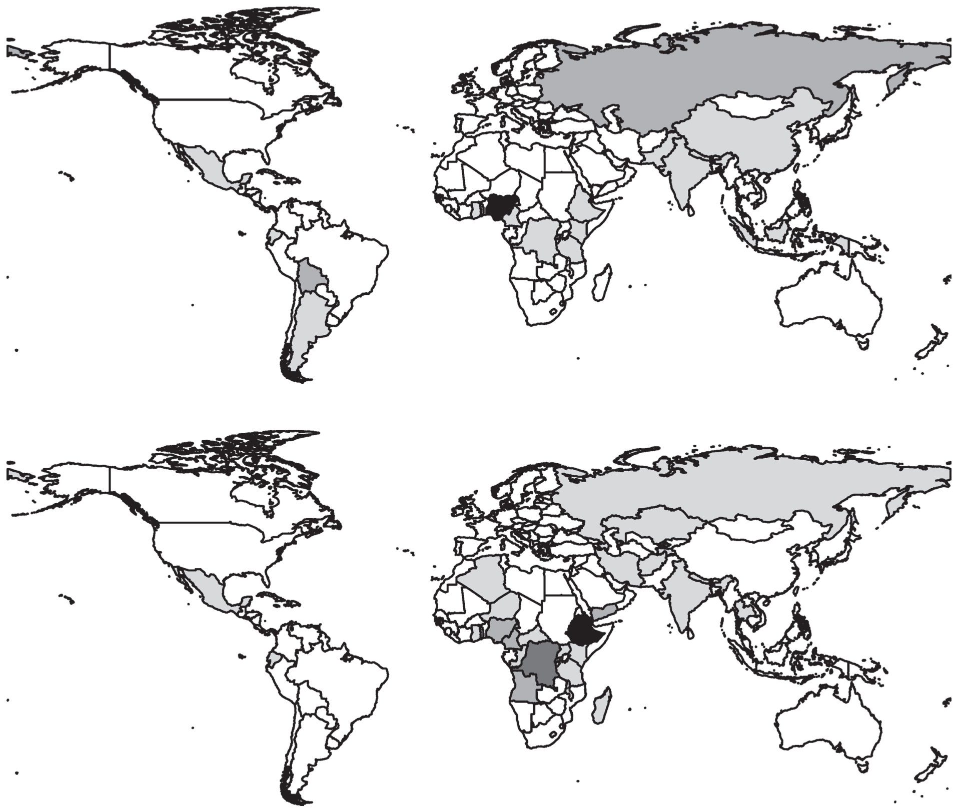

FIGURE 3 Predicted risk of civil war onset for 1985 (top) and 1995 (bottom), computed using. Model 1 in Fearon and Laitin (2003). Darker colors correspond to a higher predicted risk of conflict (4 shadings, based on a equal-sized class intervals over the range of predicted values).

propensity in Africa reaches higher values than in the 1985 example, as indicated by the higher number of dark shaded countries in Africa according to the 1995 map (Figure 3, right).

Geographic Variables for Countries

Quantitative analysis on states often requires indicators of certain aspects of a state's geography. For example, the study by Fearon and Laitin (2003) mentioned in the previous paragraph includes a variable for "roughness of terrain," measured at the state level. If geographic variables cannot readily be obtained from existing sources, they need to be computed from scratch. For simpler applications, this can be done using the R package, but as indicators get more complex, advanced GIS software is required. However, the computation of spatial indicators at the state level—whether it is done in R or a more advanced package—requires information about state boundaries over time, as provided in CShapes.

In this section, we present an example of computing a new spatial indicator using the CShapes R package. As in the previous section, we study civil war and how it relates to a country's geographic makeup. More precisely, we depart from the assumption that a state's risk of experiencing civil war should be related to its capability of repressing dissent in peripheral areas. For example, Herbst (2000) argues that the difficult geographies of African states can partly account for the many failed attempts at state-building in this region. Here, a difficult geography relates to problems of projecting power across the entire territory of a state. Unrest in peripheral regions is more likely to develop into full-fledged civil war if the state fails to provide a sufficient level of control. Existing research has found that distance from the capital is related to a higher risk of conflict. For example, the analysis by Buhaug, Cederman, and Rød (2008) shows that ethnic groups which are located far away from the capital face a higher risk of conflict. If the hinterland is indeed characterized by lower state control and thus a higher risk of conflict, states that have more extensive hinterlands should generally see more conflict.

In the following example we consider a simple test of this relationship. We posit that states in which the capital is located at a strategically advantageous position should be better able to contain unrest. We further posit that the optimal location for a capital is at the center of the state's boundaries, so that a state can maximize its ability to project power across the state's territory. A territorially unbalanced configuration, in which the capital is located at a corner or edge of a state, may create a fair amount of distant hinterland territory, where state control is relatively weak and the risk of conflict is high. As a simple measure of the territorial balance of a state, we consider the deviation between the centroid (that is, the geographic center point of its boundaries) and the actual location of the capital. Two cases help to illustrate this measure. Figure 4 shows the borders of two African countries, the Democratic Republic of the Congo and Nigeria, as well as their capitals and their polygon centroids. In the Democratic Republic of the Congo (left), the capital Kinshasa is located in the far West of the country, which creates large hinterlands in the East. Nigeria (right) has a more balanced configuration, with the capital Abuja located close to the centroid.

We compute two indicators for the deviation from the capital to the polygon centroid, the first one being the absolute distance between the two points. However, since this deviation is likely to be correlated with the country's absolute size, we also compute a relative deviation measure by dividing the absolute value by the diameter of the country.[6] We test the

[6]The diameter of a polygon is defined as the maximum distance of any pair of points on the polygon boundary. However, for the sake of computational simplicity, here we approximate the polygon diameter by the diameter of the polygon's bounding box, i.e. the smallest rectangle with straight North–South and East–West lines that includes the polygon.

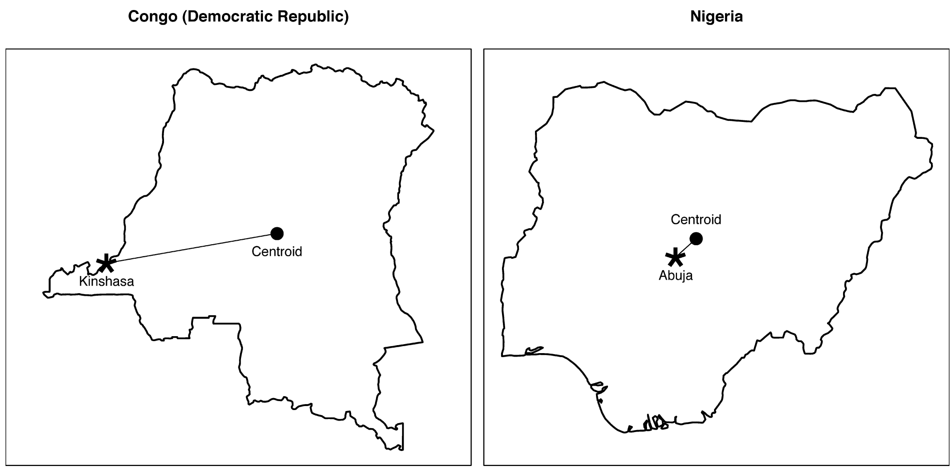

FIGURE 4 Illustration of our capital deviation measure, which corresponds to the distance between the geographic centroid of a polygon and the true location of the capital.

TABLE 1 Logit Regressions of Civil War Onset, 1950–1999

	Model 1	Model 2	Model 3	Model 4
(Intercept)	−5.21* (0.51)	−4.18* (0.20)	−5.78* (1.10)	−5.84* (1.07)
log (capital deviation)	0.48* (0.21)		−0.01 (0.25)	
Capital deviation (relative)		0.39 (0.86)		0.60 (0.85)
Prior war			−0.69* (0.31)	−0.68* (0.31)
log (GDP per capital)			−0.66* (0.23)	−0.68* (0.23)
log (population)			0.26* (0.08)	0.26* (0.08)
log (% mountainous)			0.24* (0.09)	0.25* (0.09)
Noncontiguous state			0.10 (0.30)	0.05 (0.31)
Oil exporter			0.72* (0.28)	0.71* (0.28)
New state			2.19* (0.37)	2.18* (0.37)
Instability			0.84* (0.24)	0.84* (0.24)
Democracy			0.00 (0.02)	0.00 (0.02)
Ethnic fractionalization			0.75 (0.40)	0.71 (0.40)
Religious fractionalization			0.26 (0.52)	0.25 (0.52)
N	6137	6137	6137	6137
AIC	1020.36	1025.55	955.68	955.20
BIC	1074.14	1079.32	1305.23	1304.75
log *L*	−502.18	−504.77	−425.84	−425.60

Note: Standard errors in parentheses.

explanatory impact of both variables on civil war using the Fearon and Laitin (2003) dataset.[7] As in the original paper, we use logit regression models with conflict onset as the dependent variable. Table 1 reports the results. Models 1 and 2 include only our new indicators as independent variables in

[7]The original GDP variable was again replaced by a more complete one provided by Gleditsch (2002b), see above.

the regression. In Models 3 and 4, we add these variables to the original Fearon and Laitin Model 1.

Absolute distance between capital and centroid seems to be related to civil war onset when included as a single explanatory variable in Model 1. As we hypothesized, we find a significant and positive coefficient. This is not the case for relative capital deviation (Model 3), where we also see a positive coefficient, but the estimate is far from statistically significant. Model 3 reveals that the positive effect of absolute capital distance disappears once we control for other civil war determinants included in the original model. In conclusion, our analysis provides little evidence that supports the proposed relationship between the optimal capital location and the risk of civil war. However, the above measure is very crude and does not take other factors into account, that is, the population distribution in the country. Rather than the deviation from the purely geographic center of gravity, a better measure would be the deviation from the *population* center of gravity, since ultimately it is people that need to be controlled by the state to suppress rebellion. Even though this is beyond the scope of this illustrative example, such alternative indicators can be computed using the CShapes data in combination with other geographic datasets.

Distances Between States

Regression models with spatial dependence are an important application for data on distances between states. Social scientists are becoming increasingly sensitive to how observations pooled over time require special attention to possible problems of serial dependence. However, cross-sections—samples of observations at a given point in time—are usually taken to be independent of one another, and there is less attention to the possibility that observations may be dependent across space. In an influential early comment, Galton (1889) argued that an analysis comparing marital institutions across societies could face difficulties in making valid inferences if the observed institutions were the result of diffusion processes (hence the name "Galton's problem"). Likewise, if policies enacted in one country are the result of emulation of other states or competitive processes (Simmons and Elkins 2004), the policies of individual states will not be independent of those adopted in other states.

In many instances, it is likely that such dependence between observations will be a function of spatial distance, since states tend to interact more with proximate countries and closer countries are more likely to serve as role models or reference points. As an illustrative example, consider the extent to which states have democratic institutions. Many have argued that democracy is likely to be a function of economic or social characteristics, but there are many reasons to expect that countries will be more/less likely to have democratic institutions if their neighbors have more/less democratic institutions (see Gleditsch 2002a for a more comprehensive discussion).

One way to determine whether a country's degree of democracy depends on that of its neighbors is to examine the similarity between a country's own institutional makeup and a weighted average of the level of democracy in its neighboring countries. The latter can be seen as a "spatial lag," analogous to the temporal lag of variable. More specifically, we can represent connections between states by a connectivity matrix C where individual entities c_{ij} acquire non-zero values if two units i and j are connected to one another. The spatial lag of a variable y is then computed by Wy, where W is a row-normalized version of the connectivity matrix C where the individual rows $w_{i\cdot}$ add up to 1. A variable displays spatial dependence or correlation to the extent that individual observations of y_i tend to be similar to the values of the spatial lag $w_{i\cdot}y$. CShapes allows us to create a connectivity matrix C based on different connectivity criteria; for example direct contiguity, distance between capital cities, or minimum distance between boundaries.

In our example, the outcome of interest—democracy—is also likely to be influenced by other characteristics such as income, which may also be spatially correlated. Hence, it is useful to consider spatial dependence in democracy jointly with what may be attributed to income. One way to do this is to consider democracy (y) as a conditional function of logged per capita income (x_1) and the spatial lag of democracy, i.e.,

$$E(y|x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\rho} w_{i\cdot} y$$

This conditional expected value can be estimated via linear regression. Such a regression model is often called a "spatially lagged y" model, and can be seen as a spatial analogy to a model with a temporal lag y_{t-1} on the right hand side. The fact that y appears on both sides of the equation implies a simultaneity problem that makes estimation problematic, but it is possible to fit such a model via maximum likelihood. We refer to Ward and Gleditsch (2008) for further details on spatial dependence, constructing spatial variables, and model estimation, and focus here only on an example of how measures of spatial connectivities can be derived from CShapes.

Table 2 shows the results of an OLS regression of democracy (measured by the 21-point POLITY scale) and logged per capita income from Gleditsch (2002b) for data on states at December 31, 2004, as well as a second expanded model including a spatial lag. The lag is based on an adjacency matrix where states are considered connected if within 500 km of one another, which is computed using the distmatrix() R function provided by CShapes. For simplicity, we drop all islands or observations with no neighbors within 500 kms. As can be seen, we find a positive and highly significant estimate for the parameter ρ, indicating that a country's level of democracy is

TABLE 2 Regression of Democracy on Logged Per Capita Income

	OLS			Spatial lag		
	$\hat{\beta}$	$SE(\hat{\beta})$	t-value	$\hat{\beta}$	$SE(\hat{\beta})$	z-value
Intercept	−18.378	3.667	−5.012	−12.727	3.415	−3.727
Ln(GDP per capita)	2.478	0.427	5.801	1.652	0.408	4.049
$\hat{\rho}$				0.519	0.087	5.938
N		166			166	
DF		2			3	
log L		−549.398			−532.795	

strongly associated with that of its neighbors. The coefficient estimate for the log of GDP per capita, while still significant, is considerably smaller for the spatial lag model. However, the two parameters have different interpretations, since assessing the total impact of a change in x_i in the spatial lag model would need to take into account the implied feedback between observations.[8]

CONCLUSION

The geography of the international system plays a major role in quantitative research, but is often not considered in a systematic manner. In this paper, we have introduced the CShapes project, a geographic dataset of historical state boundaries that covers the period 1946–2008. CShapes is a GIS dataset that represents state boundaries with polygons. In order to capture changes over time, these polygons are assigned a "lifespan" during which they are active. The direct representation of state boundaries has a number of advantages over the creation of separate datasets, e.g., for minimum or capital distance, since it avoids problems of consistency and updating. In addition to the GIS dataset, CShapes comes with an accompanying R package that enables users who are unfamiliar with GIS to take full advantage of the data. We have illustrated the use of the package with three examples. First, the package allows for the creation of plots to map particular quantities of interest, providing users a quick and easy way to communicate results from their research. Second, using CShapes, it is possible to compute geographic variables at the state level as for example a measure for the strategic quality of the capital location as we have shown. Third, CShapes is a valuable basis for the computation of weights matrices for spatial statistics. The corresponding

[8]More specifically, with a spatially lagged y, a change in a right-hand-side x_{ki} for country i will first change y_i directly, and then indirectly, through the effects of y_i on the neighbors of i, which in turn feed back onto i, and reverberate through the system until reaching a new equilibrium. Hence, the full "equilibrium impact" needs to be computed using the spatial multiplier $(I - \rho W)^{-1}$ (see Ward and Gleditsch 2008).

functions provided in the R package developed for CShapes enable the computation of different kinds of matrices and make the readily available for model estimation.

REFERENCES

Bivand, Roger S. (2006) Implementing Spatial Data Analysis Software Tools in R. *Geographical Analysis* 38:23–40.

Bollen, Kenneth A., Barbara Entwisle, and Arthur S. Alderson. (1993) Macrocomparative Research Methods. *Annual Review of Sociology* 19:321–351.

Buhaug, Halvard, Lars-Erik Cederman, and Jan Ketil Rød. (2008) Disaggregating Ethnic Conflict: A Dyadic Model of Exclusion Theory. *International Organization* 62(3):531–551.

Correlates of War Project. (2008) State System Membership List, v2008.1. Available online at http://correlatesofwar.org

Doggan, Mattei, and Dominique Pelassy. (1984) *How to Compare Nations: Strategies in Comparative Politics*. Chatham, NJ: Chatham House.

Environmental Systems Research Institute. (2006) Data and Maps collection for ArcGIS. Available at http://www.esri.com/data/data-maps/overview.html

Fearon, James D., and David D. Laitin. (2003) Ethnicity, Insurgency, and Civil War. *American Political Science Review* 91(1):75–90.

Furlong, Kathryn, and Nils Petter Gleditsch. (2003) The Boundary Dataset. *Conflict Management and Peace Science* 29(1):93–117.

Galton, Francis. (1889) Comment on E.B. Tylor's "On a Method of Investigating the Development of Institutions: Applied to Laws of Marriage and Descent." *Journal of the Royal Anthropological Institute* 18:268–269.

Gates, Scott, Håvard Hegre, Mark P. Jones, and Håvard Strand. (2006) Institutional Inconsistency and Political Instability: Polity Duration, 1800–2000. *American Journal of Political Science* 50(4):893–908.

Gleditsch, Kristian Skrede. (2002a) *All Politics is Local: The Diffusion of Conflict, Integration, and Democratization*. Ann Arbor, MI: University of Michigan Press.

Gleditsch, Kristian Skrede. (2002b) Expanded Trade and GDP Data, 1946–99. *Journal of Conflict Resolution* 46(5):712–724.

Gleditsch, Kristian Skrede, and Michael D. Ward. (1999) Interstate System Membership: A Revised List of the Independent States since 1816. *International Interactions* 25:393–413.

Gleditsch, Kristian Skrede, and Michael D. Ward. (2001) Measuring Space: A Minimum Distance Database and Applications to International Studies. *Journal of Peace Research* 38(6):749–768.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. (2002) Armed Conflict 1946–2001: A New Dataset. *Journal of Peace Research* 39(5):615–637.

Gochman, Charles S. (1991) Interstate Metrics: Conceptualizing, Operationalizing, and Measuring the Geographic Proximity of States since the Congress of Vienna. *International Interactions* 17(1):93–112.

Hensel, Paul R., and Paul F. Diehl. (1994) Testing Empirical Propositions about Shatterbelts. *Political Geography* 13(1):33–51.

Herbst, Jeffrey. (2000) *States and Power in Africa: Comparative Lessons in Authority and Control*. Princeton, NJ: Princeton University Press.

Heston, Alan, Robert Summers, and Bettina Aten. (2006) Penn World Table Version 6.2. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Information Technology Laboratory and National Institute of Standards and Technology. (2008) Countries, Dependencies, Areas of Special Sovereignty, and their Principal Administrative Divisions. Available at http://www.itl.nist.gov/fipspubs/fip10-4.htm

International Organization for Standardization. (2008) ISO 3166 Maintenance agency. Available at http://www.iso.org/iso/country_codes.htm

Kahler, Miles. (2002) The State of the State in World Politics. In *Political Science: State of the Discipline*, edited by Ira Katznelson and Helen V. Milner. New York: Norton.

Kastellec, Jonathan P., and Eduardo L. Leoni. (2007) Using Graphs Instead of Tables in Political Science. *Perspectives on Politics* 5(4):755–771.

Lake, David A. (2002) The State and International Relations. In *The Oxford Handbook of International Relations*, edited by Christian Reus-Smit and Duncan Snidal. New York: Oxford University Press, pp. 41–61.

Lake, David A., and Angela O'Mahony. (2004) The Incredible Shrinking State: Explaining the Territorial Size of Countries. *Journal of Conflict Resolution* 48(5):699–722.

Longley, Paul A., Michael F. Goodchild, David J. Maguire, and David W. Rhind. (2005) *Geographic Information Systems and Science*. 2nd edition. Chichester: Wiley.

Marshall, Monty G., and Keith Jaggers. (2008) Polity IV Project: Political Regime Characteristics and Transitions, 1800–2007. Available at http://www.systemicpeace.org/polity/polity4.htm

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. (2004) Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy* 112(4):725–753.

Minorities at Risk Project. (2005) Minorities at Risk Dataset. Available at http://www.cidcm.umd.edu/mar/

Rey, Sergio J., and Luc Anselin. (2006) Recent Advances in Software for Spatial Analysis in the Social Sciences. *Geographical Analysis* 38(2006):1–4.

Russett, Bruce M., J. David Singer, and Melvin Small. (1968) National Political Units in the Twentieth Century: A Standardized List. *American Political Science Review* 62(3):932–951.

Sandler, Todd. (1997) *Global Challenges: An Approach to Environmental, Political, and Economic Problems*. Cambridge: Cambridge University Press.

Simmons, Beth, and Zachary Elkins. (2004) The Globalization of Liberalization: Policy Diffusion in the International Political Economy. *American Political Science Review* 98(1):171–189.

Siverson, Randolph M., and Harvey Starr. (1991) *The Diffusion of War: A Study in Opportunity and Willingness*. Ann Arbor, MI: University of Michigan Press.

Tir, Jaroslav, Philip Schafer, Paul F. Diehl, and Gary Goertz. (1998) Territorial Changes, 1816–1996. *Conflict Management and Peace Science* 16:89–97.

Vasquez, John. (1995) Why Do Neighbors Fight? Proximity, Interaction, and Territory. *Journal of Peace Research* 32(3):277–293.

Ward, Michael D., and Kristian Skrede Gleditsch. (2008) *Spatial Regression Models*. Thousand Oaks, CA: Sage.

Weidmann, Nils B., and Kristian Skrede Gleditsch. (2010) Mapping and Measuring Country Shapes: The CShapes Package. *R Journal*, Forthcoming.