Erschienen in: Journal of Peace Research. Sage. 2010, 47(4), pp. 491-499
https://doi.org/10.1177/0022343310368352

Representing ethnic groups in space: A new dataset

Nils B. Weidmann

Center for Comparative and International Studies, ETH Zurich

Jan Ketil Rød

Department of Geography, Norwegian University of Science and Technology & Centre for the Study of Civil War, PRIO

Lars-Erik Cederman

Center for Comparative and International Studies, ETH Zurich

Abstract

Whether qualitative or quantitative, contemporary civil war studies have a tendency to over aggregate empirical evidence. In order to open the black box of the state, it is necessary to pinpoint the location of key conflict parties. As a contribution to this task, this article describes a data project that geo references ethnic groups around the world. Relying on maps and data drawn from the classical Soviet *Atlas Narodov Mira* (ANM), the 'Geo referencing of ethnic groups' (GREG) dataset employs geographic information systems (GIS) to represent group territories as polygons. This article introduces the structure of the GREG dataset and gives an example for its application by examining the impact of group concentration on conflict. In line with previous find ings, the authors show that groups with a single territorial cluster according to GREG have a significantly higher risk of conflict. This example demonstrates how the GREG dataset can be processed in the *R* statistical package without specific skills in GIS. The authors also provide a detailed discussion of the shortcomings of the GREG dataset, resulting from the datedness of the ANM and its unclear coding conventions. In comparing GREG to other datasets on ethnicity, the article makes an attempt to illustrate the strengths and weaknesses associated with the GREG database.

Keywords

ethnic conflict, ethnic groups, geographic information systems, group settlement patterns

Introduction

The role of ethnicity in conflict processes remains as controversial as ever. The ethnic wars following the end of the Cold War triggered a surge of interest in ethnic conflict (e.g. Posen, 1993; Kaplan, 1993). In this wave of scholarship, inter national relations specialists attempted to apply their tradi tional tools that had been developed to study primarily superpower relations and other interstate exchanges (Ceder man, 2002). In contrast, more recent research has turned attention more specifically to civil wars. Inspired by a promi nent project funded by the World Bank, a number of political scientists and economists have expressed doubts as to whether ethnic grievances really drive such conflicts (Fearon & Laitin, 2003; Collier & Hoeffler, 2004).

Since the current literature on civil wars relies extensively on cross national statistics, the question of what mechanisms drive observed macro patterns can only be answered indirectly (Sambanis, 2004). This problem pertains to the effect of pro minent indicators, including GDP per capita, democracy and geography. Most importantly, however, the seemingly absent effect of ethnic grievances is also based on aggregate indicators. In order to get closer to the micro mechanisms that drive ethnic civil wars, it is necessary to focus more explicitly on the actors in these conflicts. One way to do this is to use ethnic groups as the unit of analysis. Even though 'groups' do not themselves commit acts of violence in conflict (Brubaker, 2004), ethnic divisions are the relevant societal cleavages along which many internal conflicts are fought. Ethnic groups thus constitute a meso level of analysis in the study of civil war a first step down from the state as an over aggregated analytical unit, but still above the micro level of political organizations,

Corresponding author:
Nils B. Weidmann: nils.weidmann@gmail.com

rebel movements, and individuals, where limited data availability makes large N comparisons at a global scale impossible.[1]

Why do some groups experience violence, whereas others do not? More careful scrutiny of ethnic groups requires better indicators at the group level. In this article, we present the GREG project ('Geo referencing of Ethnic Groups'), which attempts to attain this goal by disaggregating ethnicity spatially. More specifically, our goal is to place ethnic groups on the map as a way to locate key actors of conflict processes. Information about the settlement regions of ethnic groups make it possible to describe group characteristics that potentially make them more susceptible to conflict. The GREG project is only one in a series of recent attempts to geo code conflict and its determinants, such as for example location and scope of civil wars (Buhaug & Gates, 2002). None of these projects, however, provides a systematic treatment of ethnic groups. This is the gap the GREG project aims to fill.

In the following, we proceed in four steps. First, we survey possible data sources that could potentially support spatial disaggregation of ethnicity. Second, we introduce our GREG dataset and its structure. Third, we describe possible applications of the dataset, starting with an example that replicates Toft's (2003) analysis on group concentration and conflict. Fourth, we analyze potential problems associated with the use of the GREG data.

Spatial data on ethnic groups

In the literature, ethnicity has typically entered analysis of conflict processes either as qualitative, historical entities or as quantitative indices, such as the ethno linguistic fractionalization index (ELF, Fearon, 2003; Posner, 2004). In neither case, however, has space played a prominent role in the development of causal arguments. Historical and other qualitative accounts of ethnic conflict occasionally provide maps that show the spatial distribution of groups, but such information is hardly ever supplied for a larger sample of states (Horowitz, 1985; Herbst, 2000). While the quantitative literature does offer some references to group geography, such as settlement concentration, such information is usually narrowed down to a few variables. The *Minorities at Risk* dataset (MAR) includes indicators for group concentration or urban rural settlement of groups (MAR, 2005). However, since MAR does not provide the spatial distribution of groups directly, one quickly reaches the dataset's limits if more complex indicators are required. To our knowledge, there are no data sources that systematically pin down the location of ethnic groups in a large number of comparable cases. This raises the question of where such information could be found. Several candidates exist.

Linguists have developed detailed maps of language diffusion; see, for example, *Ethnologue* (Gordon, 2005). However, this data resource is problematic because it is narrowly focused on linguistic traits. Its linguistic charts are typically either too detailed to serve as a guide to ethnic group delimitation or too sketchy as they often represent a linguistic group with a point, thus making delimitation or inference on spatial dissemination fuzzy. Moreover, *Ethnologue* does not include spatial information for all countries, making it unsuitable for the development of a global dataset.

Another possibility would be to infer the location of ethnic groups from census or survey data. Yet, such an approach is only viable where such data contains references to ethnicity, which is often not the case. Furthermore, it also hinges on the presence of a reasonably fine grained provincial structure. Where federal subunits are large, the necessary degree of spatial disaggregation may never be attained. A few cases exist for which this approach is possible: for example, the 1991 census for Bosnia provides detailed information about the ethnic composition of municipalities (Petrovic, 1992). Nevertheless, providing spatially referenced census data for a larger set of cases is not possible.

For these reasons, we have chosen to rely on data and maps from the well known *Atlas Narodov Mira* (ANM, Bruk & Apenchenko, 1964), which stems from a major project of charting ethnic groups worldwide, undertaken by Soviet ethnographers in the 1960s. The ANM has several strengths: it is complete and carefully researched, it relies on a uniform group list that is valid across state borders, and it provides high quality maps. At the time of publication, the ANM received excellent reviews (Hewes, 1966). Among its advantages, Harris (1965) explicitly mentions the recognition of minority groups in states dominated by other groups, which makes it particularly suitable for the study of ethnic conflict. Until now, the ANM has been widely used in contemporary research, mainly as a basis for calculating the ELF index (Taylor & Hudson, 1972). However, it should be stressed that the ANM also has its weaknesses, which we discuss in detail below.

Creating a GIS dataset on ethnic groups

In this section, we describe our efforts to create a GIS dataset on ethnic groups on the basis of the ANM. We start with a detailed introduction of the ANM data and then show how this information was converted into GIS format.

The ANM consists of 57 ethnographic maps, covering all regions of the world at various scales.[2] Each map shows the geographic distribution of the relevant groups, indicated by colored areas. In addition to the color coding, the areas are marked with numbers which refer to the respective group's

[1] Efforts have been made to code non-state actors in civil wars, but in such cases the focus is on rebel organizations rather than on nonviolent political organizations (Cunningham, Gleditsch & Salehyan, 2009).

[2] The ANM includes maps on population density. However, better data on population exist, so we did not make use of these maps.

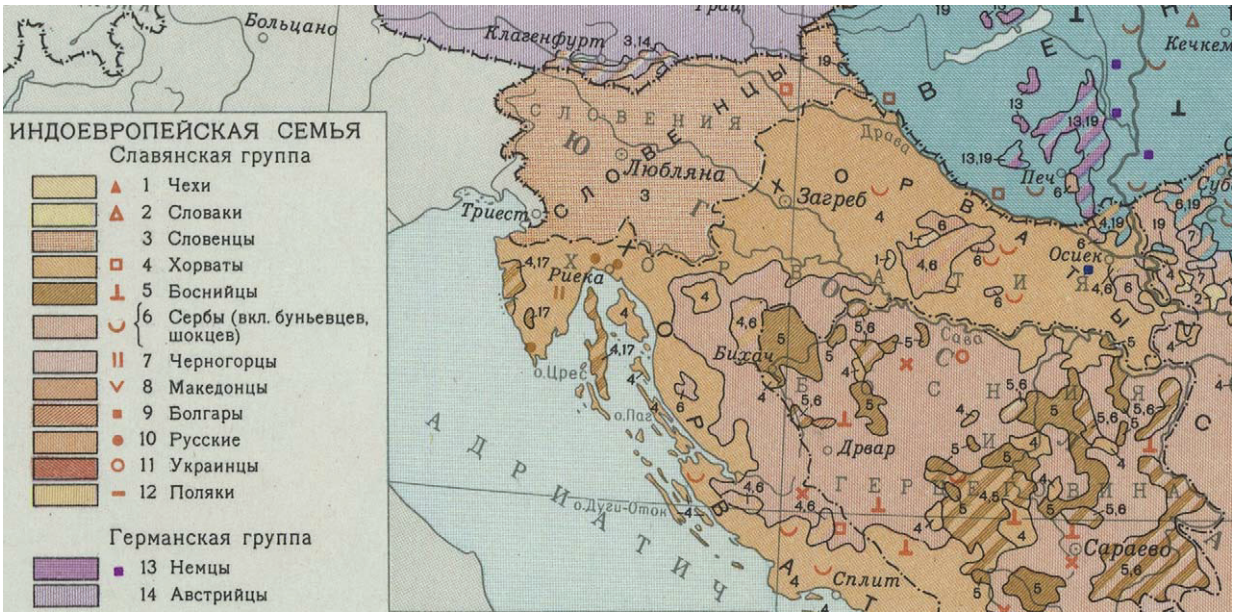

Figure 1. Map of former Yugoslavia from the *Atlas Narodov Mira*

Figure reproduced from Bruk and Apenchenko (1964: 40). Translation of the legend: Indo-European family. Slavic group: (1) Czechs, (2) Slovaks, (3) Slovenes, (4) Croats, (5) Bosniaks, (6) Serbs, (7) Montenegrins, (8) Macedonians, (9) Bulgarians, (10) Russians, (11) Ukrainians, (12) Poles. Germanic group: (13) Germans, (14) Austrians.

number in the legend. Most areas are coded as pertaining to one group only, but in some cases there can be up to three groups sharing a certain territory (although the latter case is quite rare, see below). This is indicated on the map by a striped fill of the respective areas. Figure 1 shows a part of the map covering former Yugoslavia. For each area, one or two num bers indicate the respective group.

The ANM also provides information about groups without a clear territorial base. The presence of these groups is indi cated by symbols rather than areas. To give an example in the map in Figure 1, minor occurrences of group 4 (Croats, square symbol) can also be found in Northeast Slovenia. Sparsely populated areas can be distinguished from others by their grey raster fill and the missing group color fill. However, for these regions group presence is still indicated by symbols and num bers as explained above. Unpopulated regions are left white.

The source of the information contained in the ANM remains somewhat obscure. A short text at the beginning of the volume list three different types of sources: (1) ethno graphic and geographic maps assembled by the Institute of Ethnography at the USSR Academy of Sciences, (2) popula tion census data, and (3) ethnographic publications of govern ment agencies. Still, it remains unclear what kind of information was used for which maps, and how groups were selected in the first place. This is an issue to which we return later. Apart from the map collection, the ANM features a sta tistical appendix complementing the geographic information. The appendix contains two major lists. The first one gives the full set of groups mentioned in the ANM along with their relative population sizes within each country, and the second contains all countries together with their groups. It is the latter list that has served as a basis for the computation of ELF scores in the literature (Taylor & Hudson, 1972).

The aim of the GREG project is to make the ethnic maps usable for spatial analysis by converting them into a GIS dataset. Groups without a territorial basis (those marked with symbols on the map) were not included. For the dataset creation, three steps were carried out. First, the maps from the ANM were scanned to obtain an image file for each map. Second, these images were spatially referenced using a geographic information system (see the online appendix for details on this procedure). Third, since all the maps in the ANM are annotated in Russian, the English group names were assigned to the polygons by a native Russian speaker, using the translations of the Russian group names provided in the ANM's appendix.[3] The final result is a master list of ethnic groups, each with a unique numeric identifier, and a set of polygons in ESRI's shapefile format (ESRI, 1998), each of which contains the identifiers of the corresponding group(s) (more details given in the online appendix).

The full GREG dataset has global coverage and consists of 929 groups represented with 8,969 geo referenced polygons. In the ANM, there are 1248 groups in total but as 319 of these do not have any territorial basis, they are not contained in the

[3] An English translation of the ANM map legends exists (Telberg, 1965), but only became available to us after completion of the project. However, its group name translations largely correspond to those given in the ANM appendix used for GREG. A PDF copy of Telberg (1965) is provided as part of our replication file.

GREG GIS dataset. If a given research design requires also the inclusion of groups without a territorial basis, these groups can be found in the ANM's demographic appendix. Of the 8,969 polygons, the vast majority (7,383) contain one group. 1,552 polygons contain two groups and for only 34 polygons, there are three groups listed in the dataset. The size of the polygons varies considerably: the smallest polygon occupies an area of 0.59 km^2, and the largest polygon extends over 6,954,564 km^2.[4]

Applications

Since the GREG dataset is provided independently of state boundaries, it is possible to analyze ethnic groups at different levels: transnationally, at the state level, or even down to the level of subnational units and groups. Our examples focus on the latter.

Group concentration and conflict

Recent research found that geographically concentrated groups face a higher risk of conflict (Toft, 2003). In this example, we replicate Toft's analysis by computing a simple measure of group concentration based on GREG (see Weidmann, 2009, for an extended example). Typically, processing spatial data requires substantial skills in GIS techniques. However, our con centration measure does not involve any complex GIS compu tations and can therefore be implemented in *R*, a freely available statistical software package.[5] Therefore, this example also serves as a hands on exercise on how to use GREG outside the context of GIS, which we hope can lower the entry costs for quantitative researchers lacking a GIS background. The computation is dis cussed in detail in the appendix.

Geographically concentrated groups occupy a single, contiguous region in a country. In GREG, we measure group concentration by the number of territorial clusters occupied by the group. For each group, we retrieve all its polygons and assign them to clusters, such that all polygons in a cluster are contiguous to each other. The number of these clusters is then used as an independent variable in a regression analysis. In the online appendix, we give a detailed description of how this variable is computed in *R*.

For testing the impact of group concentration on conflict, we rely on the data coded by Buhaug, Cederman & Rød (2008). They analyze ethnic conflict in a dyadic framework, pitting ethnic groups in power against peripheral groups. The dataset includes 67 countries from Europe, Asia, and North Africa. The dependent variable, onset of ethnic conflict, is coded as 1 for dyad years when conflict erupted between a peripheral group and the government. We follow the approach in the original paper and estimate logit regression models with standard errors clustered by country.[6] Besides our independent variable 'number of clusters', our models include the three main independent variables of the original study: (i) the dyadic *power balance* between the government groups and the respective peripheral group. Higher values of this variable indicate that the peripheral group is strong compared to the government. We include (ii) a measure of the group's distance to the capital, and (iii) an indicator of whether the group lives in a mountainous area. At the country level, we control for population and economic performance, two variables that have been found to have a strong impact on internal conflict. Since observations of conflict onset are likely to have strong time dependence, we include a variable measuring the number of peace years, as well as its square and cubic transformation (Carter & Signorino, 2006). Model 1 replicates the original model. Model 2 includes our independent variable, number of clusters. Model 3 tests the impact of group concentration using a dummy variable for a group with one territorial cluster. All results are reported in Table I.[7]

In line with the original findings, Model 1 shows the strong effect of power balance on conflict onset. Strong groups compared to the government are significantly more likely to experience conflict. The same holds for groups settling in regions distant from the capital, but unlike in the original paper, the effect is not significant. Also, groups in mountai nous areas face a higher risk of conflict. At the country level, we see that more populous and poorer countries receive a positive sign, but the effects are not significant. When we add the number of territorial clusters as an additional variable in Model 2, these effects largely persist. We see that more dispersed groups measured by the number of territorial clusters face a lower risk of conflict, a finding which is in line with previous research (Toft, 2003). The effect of the number of clusters is considerable: increasing the number of clusters from 1 to the sample mean of 6.86, while keeping all other variables in the model at their mean, decreases the predicted probability by about 20%. In Model 3, we find that this effect is largely driven by the difference between groups with a single territorial cluster and those with more than one: comparing groups with one territorial cluster to those with more than one cluster is linked to a decrease of about 50% in the predicted conflict probability. This finding is in agreement with Toft's conclusion that the absence of perfect territorial concentration seems to be almost a sufficient cause for peace.

Combining GREG with GIS raster data

Besides variables computed directly on GREG, researchers can take advantage of advanced GIS techniques and combine

[4] All area computations were performed with ArcGIS 9.2 using an Eckert VI equal area projection.

[5] See http://www.r-project.org/.

[6] Software: *R* version 2.7.1 with *Design* package, version 2.1-1.

[7] We restrict the sample period from 1946–89 to simplify the analysis. After the Cold War many new states emerged. To include this period would require us to introduce a time dimension to the computation of group concentration, since polygons might belong to one state at a given time, but to another state later.

Table I. Logit models of ethnic conflict onset at the group level, 1946 89

	Model 1 *Coefficient (Std. error)*	*P*	*Model 2* *Coefficient* (Std. error)	*P*	*Model 3* *Coefficient* (Std. error)	*P*
Number of clusters			0.03 (0.01)	0.00		
One territorial cluster					0.76 (0.30)	0.01
Power balance	0.55 (0.12)	0.00	0.66 (0.14)	0.00	0.60 (0.13)	0.00
Distance to capital	0.55 (0.36)	0.12	0.70 (0.31)	0.02	0.58 (0.34)	0.09
Mountains	0.66 (0.39)	0.09	0.69 (0.48)	0.15	0.55 (0.45)	0.22
Country population	0.20 (0.26)	0.26	0.32 (0.17)	0.06	0.26 (0.16)	0.10
Per capita GDP	0.03 (0.16)	0.87	0.03 (0.13)	0.84	0.01 (0.14)	0.95
Intercept	8.50 (2.35)	0.00	10.03 (2.30)	0.00	9.23 (2.43)	0.00
N	25,674		25,674		25,674	
Likelihood ratio	58.87		66.58		64.52	
Model significance	0.00		0.00		0.00	

Peace years control (and square and cubic transformations) not shown.

GREG with other GIS databases, as for example raster datasets. In contrast to the polygon format used by GREG, a raster dataset divides the globe into small quadratic cells and stores one value for each cell. For example, population counts are available as a raster dataset, making it possible to compute the population for group polygons. GREG can also be com bined with other raster datasets, for example on territorial ele vation, land use, etc.

We illustrate the general procedure with an example of how to compute population estimates for groups. The 'Gridded Population of the World' dataset (GPW, CIESIN, 2005) pro vides population figures for small raster cells.[8] The group poly gons are then superimposed on the GPW raster such that the cells covered by a polygon can be selected and their population values added, resulting in a population estimate for the group polygon. Figure 2 illustrates this. The color of the raster cells indicates the population in the respective cell, with darker shading corresponding to a higher population. The GREG polygons (solid lines) represent the zones for which the popu lation values are aggregated. An obvious simplification in this estimation procedure is that the entire population within a zone is assumed to belong to one group. However, unless we have better information about the distribution of groups at particular locations, this is the best estimation one can make.

Ethnic variables for sub-national units

The GREG data can also be used to create indicators of ethni city for geopolitical units other than the state. For example, Cunningham & Weidmann (in press) study the effect of eth nic polarization on conflict at the level of administrative dis tricts. Data about ethnicity is almost impossible to obtain for a global sample of units, so they employ a geographic esti mation of group sizes, illustrated in Figure 3 (example: Sino county in Liberia). The shaded polygons indicate the group polygons from GREG, two of which intersect with the Sino administrative boundaries (the Babinga and Gere). Using GIS software, the GREG polygons are clipped along the district boundaries such that we get two polygons for the Sino district, one for each group (Figure 3, right). In order to obtain popu lation estimates for groups, these clipped polygons can now be combined with population raster data as described above in order to obtain ethnic population figures at the district level.

Further applications

The information in GREG can also be used in conjunction with geographic datasets on conflict to examine the relation ship between the ethnic distribution and the occurrence of vio lence. For example, in their study on the diffusion of civil war, Buhaug and Gleditsch (2008) use GREG to determine whether a country has ethnic ties to groups within the conflict zone of a neighboring state. Other applications are presented in Cederman, Buhaug & Rød (2009) and Cederman, Girardin & Gleditsch (2009). Using conflict event data at an even more

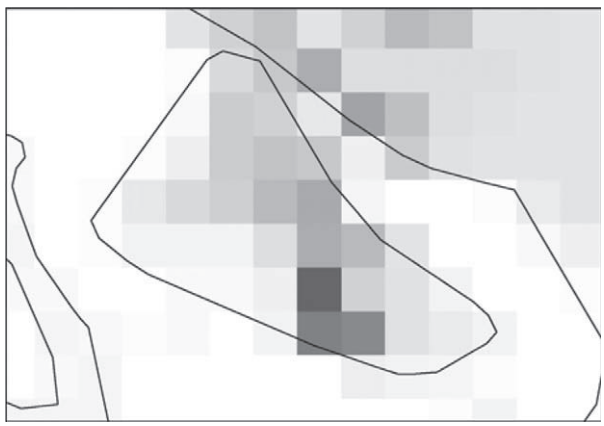

Figure 2. Obtaining population estimates for geographic areas

Population estimates for group areas in GREG (solid lines) are obtained by overlaying a raster dataset (shaded cells) with the GREG polygons and sum-ming up the population of the cells covered by each polygon.

[8] 2.5′ x 2.5′, which corresponds to approximately 5 km x 5 km at the equator.

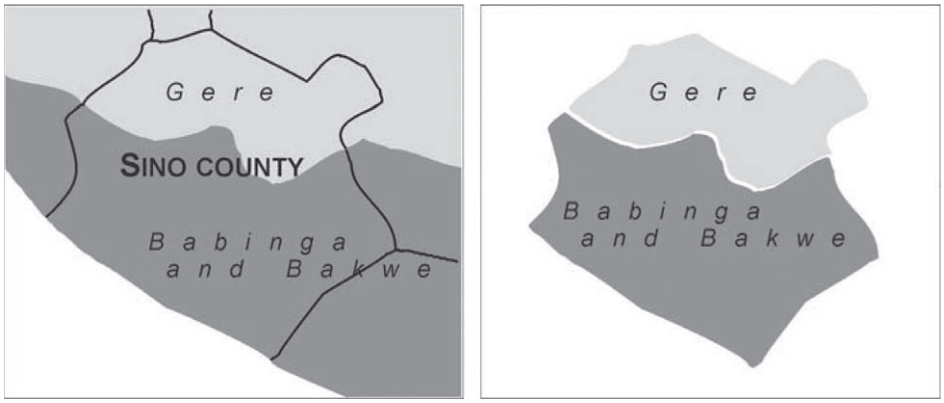

Figure 3. Combining GREG data with sub state administrative boundaries

Intersecting group areas from GREG (shaded polygons) with sub-national units (solid lines) results in a set of polygons that correspond to the settlement area of a group in the unit. These polygons can then be used for computing population estimates as described above.

detailed level (Raleigh & Hegre, 2010), GREG even makes it possible to study the extent to which the geographic ethnic dis tribution determines the location of violence. In general, owing to the increasing availability of geographic data, the number of possible applications of GREG is likely to grow. For an in depth introduction to GIS and spatial analysis, see Longley et al. (2005) for ArcGIS, or Bivand, Pebesma & Gómez Rubio (2008) for the *R* package.

Problems

Outdated

An obvious problem with the GREG data is that the data in the ANM were collected in the early 1960s and might not reflect the current ethnic configuration. Still, ethnic settlement pat terns exhibit a lot of inertia, so that it is plausible to also use the GREG data as the basis for measuring ethnic geography in recent times. Especially during conflicts, however, ethnic configurations might change significantly, for example by conflict induced population movements, the systematic expul sion of people, or ethnic cleansing. In other words, the accuracy of the dataset could suffer precisely in cases with a recent history of conflict. Whereas this can obviously constitute a limiting fac tor in the study of ethnic geography and conflict, the upside is that the ethnic configuration as captured by GREG is causally prior to the majority of ethnic conflicts in the post World War II period. Still, a static conception of the ethnic configuration in space leaves much to be desired. In order to capture the dynamics of settlement pattern changes more accurately, one would have to introduce a time dimension to GREG, for exam ple by coding snapshots of the settlement areas at different time points. However, this is a complex undertaking, which we plan to address in a follow up project.

Group categories problematic

What is an ethnic group? Ethnic groups can be distinguished along linguistic or religious lines, with or without a territorial basis, or with respect to their political relevance. The coding conventions of the ANM are not documented anywhere in the volume. As a result, we can only infer the coding criteria by comparison with existing data sources on ethnic groups. Bridgman (2008) does this at the state level and compares inter marriage rates as predicted by the ANM to other data sources. In general, he finds that the ANM categorization largely corre sponds to societal cleavages as defined by a lack of intermarriage, but a few cases exist where the ANM underestimates ethnic divisions. However, in order to find out more about the ANM coding conventions, a more thorough assessment at the group level is required. We do so by comparing GREG group cate gories to other frequently used datasets in the field: first, the list of ethnic groups compiled by Fearon (2003), and second, the above mentioned MAR dataset (MAR, 2005).

The comparison of group lists is a difficult effort because in many cases, group names need to be matched manually. For that reason, we limit our comparison to four countries: Belgium, Iraq, Georgia, and Cambodia. Table II summarizes the comparison. Shaded lines list the matching groups in the three datasets. Empty cells indicate that no match was found in the respective dataset.

Belgium For Belgium, GREG and the Fearon dataset list the major groups of the country (Flemings and Walloons). However, there is disagreement on what minority groups should be included. GREG includes the small German speaking population in the east of the country. It is unclear why the Fearon dataset excludes this group, given that it fits most of his inclusion criteria (Fearon, 2003: 201). On the other hand, Fearon includes Italians and Moroccans, who are most likely foreign workers in Belgium with limited political relevance. These groups, however, do not have a territorial base in Belgium (and were probably not present at the time the ANM was created), so they are not included in GREG. MAR does not list any groups for Belgium, since none of the groups is seen as discriminated.

Iraq The ANM's focus on linguistic boundaries poses a problem for cases where alternative group distinctions are

Table II. Comparing the GREG, Fearon (2003), and MAR group lists

Country	*GREG*	*Fearon (2003)*	*MAR*
Belgium	Flemings	Flemings	
	Germans		
	Walloons	Walloons	(none)
		Italians	
		Moroccans	
Iraq	Assyrians		
	Circassians		
	Iraq Arabs	Shi'is	Shi'is
		Sunni Arabs	Sunnis
	Kurds	Kurds	Kurds
	Lur		
	Persians		
	Turkmens	Turkomans	
Georgia	Abkhaz	Abkhazians	Abkhazians
	Armenians	Armenians	
	Azerbaijanians	Azeris	
	Bats		
	Estonians		
	Georgians	Georgians	
	Ingushes		
	Moldavians		
	Ossetes	Ossetians (South)	Ossetians (South)
	Russians	Russians	Russians
		Adzhars	Adzhars
Cambodia	Boloven		
	Cham	Chams	Chams
	Jarai		
	Khmers	Khmers	
	Kui		
	Lao		
	Ma		
	Malays of Malaya		
	Muong and Brao		
	Siamese		
	Stieng		
	Vietnamese	Vietnamese	Vietnamese
		Chinese	

relevant. An example is Iraq, where divisions between different Muslim denominations Shi'ites and Sunnis have dominated the country for a long time. This distinction is made both in the Fearon dataset and in MAR. However, the GREG dataset summarizes these groups under a single identity, 'Iraq Arabs', since both speak Arabic. Sunnis and Shi'ites differ significantly in terms of their spatial distribution, with the Shi'ites dominating the south of the country and the Sunnis mostly present in the north of Baghdad and in the west.[9] Relying on alternative data sources, it would therefore be possible to pin down the relevant groups geographically. Again, MAR's selection of discriminated minorities leads to the exclusion of the governing ethnic groups. GREG includes smaller groups that speak a different language, for example the Assyrians, Circassians, and Persians. The Kurds are present in all three datasets. The Turkmens of Iraq, however, are not listed in the MAR dataset.

Georgia As in the case of Belgium, GREG and the Fearon dataset agree with respect to the most important groups in Georgia (Georgians, Azeris, Armenians, Russians, Abkhazians, and Ossetians). However, the Fearon group list distinguishes the Adzhars (or Ajars) from the rest of the Georgians because of their different religion (Islam), whereas the ANM does not make this distinction. This might be due to the radical

[9] See e.g. http://www.lib.utexas.edu/maps/middle east and asia/iraq ethno 2003.jpg.

suppression of Islam during the Soviet era (Hughes & Sasse, 2002), where the ANM criteria for inclusion might have been guided by political aspirations. The ANM, however, lists other minority groups without political relevance. Four groups are regarded as discriminated and are therefore included in the MAR dataset, among them the Adzhars that the ANM fails to list and the secessionist Abkhaz and South Ossetians. By definition of MAR's coding rules, the titular majority the Georgians fails to show up in the MAR group list.

Cambodia Among the four examples described here, Cambodia is the one where the high level of detail of the ANM is most obvious. The GREG dataset contains detailed information about the highland tribes (Jarai, Kui, Muong, Stieng) which is not present in any of the other datasets. GREG and the Fearon dataset agree on the major groups (Khmer, Vietnamese, and Cham), but GREG does not list the Chinese, which already constituted a significant minority in the 1960s (Ross, 1987). This is due to the fact that the Chinese were engaged in commerce all across Cambodia, so they do not have a delimited territorial basis in the country.

In sum, our comparison highlights major differences between datasets on ethnic groups. By definition, MAR shows only parts of the ethnic landscape by focusing on discriminated groups only. In the case of Iraq, the ANM's focus on linguistic boundaries causes GREG to omit the Sunni Shi'ite division, one of the most important cultural cleavages in the country. Since the coding criteria for the ANM are not spelled out in detail, it is in some cases difficult to reproduce the group categorizations used. Most of the distinctions are clearly based on linguistic differences. Additionally, however, national boundaries were introduced. For example, the ANM distinguishes between Germans and Austrians, even though they belong to the same language family. Similarly, the ANM separates 'English Irish' from 'Irishmen'. These coding decisions might be problematic in some cases and require additional inspection of the group list. Furthermore, it is important that as a spatial dataset, GREG only includes territorial groups, that is, those that have one or more settlement regions in a country. Migrant groups and foreign workers residing only in urban areas are not represented in GREG.

Conclusion

The quantitative literature on civil war has usually relied on highly aggregated determinants. With the GREG data project, we aim to disaggregate these indicators by providing a comprehensive and complete geographic dataset on ethnic groups. As we have demonstrated above, GREG can be used to derive various measures for the ethnic groups. Alternatively, using GREG it is possible to compute ethnic indicators for political units other than the state. The use of GREG is not unproblematic, as the focus on linguistic boundaries could limit the application of the dataset to political science questions. However, we believe that the GREG dataset constitutes a significant step forward in the quantitative study of ethnicity and conflict that will allow researchers to derive new insights about the role of ethnic groups.

Acknowledgments

We are grateful to the staff of the Zurich Central Library Map Collection, especially Hans Peter Hoehener, and to Helena Kusch, Doreen Kuse, and Olga Nikolayeva for their excellent research assistance.

Funding

Nils Weidmann is supported by ETH (Research Grant TH 4/05 3).

References

Bivand, Roger S; Edzer J Pebesma & Virgilio Gómez Rubio (2008) *Applied Spatial Data Analysis with R*. New York: Springer.

Bridgman, Benjamin (2008) What does the Atlas Narodov Mira measure? *Economics Bulletin* 10(6): 1 8.

Brubaker, Rogers (2004) *Ethnicity without Groups*. Cambridge, MA: Harvard University Press.

Bruk, Solomon I & V S Apenchenko, eds (1964) *Atlas narodov mira* [Atlas of the Peoples of the World]. Moscow: Glavnoe Upravlenie Geodezii i Kartografi.

Buhaug, Halvard & Scott Gates (2002) The geography of civil war. *Journal of Peace Research* 39(4): 417 433.

Buhaug, Halvard & Kristian Skrede Gleditsch (2008) Contagion or confusion? Why conflicts cluster in space. *International Studies Quarterly* 52(2): 215 233.

Buhaug, Halvard; Lars Erik Cederman & Jan Ketil Rød (2008) Disaggregating ethnic conflict: A dyadic model of exclusion theory. *International Organization* 62(3): 531 551.

Carter, David B & Curtis S Signorino (2006) Back to the future: Modeling time dependence in binary data. Working paper, University of Rochester, NY (http://polmeth.wustl.edu/retrieve.php?id=710).

Cederman, Lars Erik (2002) Nationalism and ethnicity. In: Walter Carlsnaes, Thomas Risse & Beth A Simmons (eds) *Handbook of International Relations*. London: Sage, 409 428.

Cederman, Lars Erik; Halvard Buhaug & Jan Ketil Rød (2009) Ethno nationalist dyads and civil war: A GIS based analysis. *Journal of Conflict Resolution* 53(4): 496 525.

Cederman, Lars Erik; Luc Girardin & Kristian Skrede Gleditsch (2009) Ethno nationalist triads: Assessing the influence of kin groups on civil wars. *World Politics* 61: 403 437.

CIESIN (2005) Gridded Population of the World v3 (http://sedac.ciesin.columbia.edu/gpw).

Collier, Paul & Anke Hoeffler (2004) Greed and grievance in civil war. *Oxford Economic Papers* 56(4): 563 595.

Cunningham, David E; Kristian Skrede Gleditsch & Idean Salehyan (2009) It takes two: A dyadic analysis of civil war duration and outcome. *Journal of Conflict Resolution* 53(4): 570 597.

Cunningham, Kathleen & Nils B Weidmann (in press) Shared space: Ethnic groups, state accommodation and localized conflict. *International Studies Quarterly*.

ESRI (1998) *ESRI Shapefile Technical Description*. ESRI whitepapers (http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf).

Fearon, James D (2003) Ethnic and cultural diversity by country. *Journal of Economic Growth* 8(2): 195 222.

Fearon, James D & David D Laitin (2003) Ethnicity, insurgency and civil war. *American Political Science Review* 97(1): 75 90.

Gordon, Raymond G, Jr, ed. (2005) *Ethnologue: Languages of the World*, 15th edn. Dallas, TX: SIL International.

Harris, Chauncy D (1965) Review of the Atlas Narodov Mira. *Geographical Review* 55(4): 608 610.

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

Hewes, Gordon W (1966) Review of the Atlas Narodov Mira. *American Anthropologist* 68: 532 534.

Horowitz, Donald (1985) *Ethnic Groups in Conflict*. Berkeley, CA: University of California Press.

Hughes, James & Gwendolyn Sasse (2002) *Ethnicity and Territory in the Former Soviet Union: Regions in Conflict*. London: Routledge.

Kaplan, Robert D (1993) *Balkan Ghosts: A Journey through History*. New York: St. Martin's.

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

MAR (2005) Minorities at Risk Dataset. College Park, MD: Center for International Development and Conflict Management (http://www.cidcm.umd.edu/mar/).

Petrovic, Ruza (1992) The national composition of Yugoslavia's population, 1991. *Yugoslav Survey* 33(1): 3 24.

Posen, Barry R (1993) The security dilemma and ethnic conflict. In: M E Brown (ed.) *Ethnic Conflict and International Security*. Princeton, NJ: Princeton University Press, 103 124.

Posner, Daniel N (2004) Measuring ethnic fractionalization in Africa. *American Journal of Political Science* 48(4): 849 863.

Raleigh, Clionadh, et al. (2010) Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research* 47(5): In press.

Ross, Russell R, ed. (1987) *Cambodia: A Country Study*. Washington, DC: GPO for the Library of Congress (http://countrystudies.us/cambodia/).

Sambanis, Nicholas (2004) Using case studies to expand economic models of civil war. *Perspectives on Politics* 2(2): 259 279.

Taylor, Charles & Michael C Hudson (1972) *World Handbook of Political and Social Indicators*. New Haven, CT: Yale University Press.

Telberg, Vladimir G (1965) *Telberg Translation to Atlas Narodov Mira*. New York, NY: Telberg.

Toft, Monica Duffy (2003) *The Geography of Ethnic Violence: Identity, Interests, and the Indivisibility of Territory*. Princeton, NJ: Princeton University Press.

Weidmann, Nils B (2009) Geography as motivation and opportunity: Group concentration and ethnic conflict. *Journal of Conflict Resolution* 53(4): 526 543.

NILS B. WEIDMANN, b. 1976, MSc (University of Freiburg, 2003), MA (Swiss Federal Institute of Technology, ETH, Zurich, 2008); PhD candidate in Political Science, ETH Zurich.

JAN KETIL RØD, b. 1966, PhD (NTNU Trondheim, 2002); Associate Professor, Department of Geography, NTNU Trondheim.

LARS ERIK CEDERMAN, b. 1963, PhD (University of Michigan, 1994); Professor of International Conflict Research, ETH Zurich.