Integrating Data on Ethnicity, Geography and Conflict:

The *Ethnic Power Relations* Dataset Family

5978 words

This paper introduces the new Family of Ethnic Power Relations *(EPR) Datasets, version 2014, which is the latest in a series of datasets on ethnicity that have stimulated civil war research in the past decade. The EPR Family provides data on ethnic groups' access to state power, their settlement patterns, links to rebel organizations, trans-border ethnic kin relations, and intra-ethnic cleavages. The new 2014 version does not only extend the dataset's temporal coverage from 2009 to 2013, but it also offers several new features, such as a new measure of regional autonomy that is independent of national-level executive power and a new dataset component coding intra-ethnic identities and cleavages. Moreover, for the first time, detailed documentation of the EPR data is provided through the* EPR Atlas. *The paper presents these novelties in detail and compares the EPR Family 2014 to the most relevant alternative datasets on ethnicity.*

Introduction

The past decade has seen a surge in academic studies of ethnic civil conflicts, supported by significant efforts in data collection. Focusing on one of these new data sources, this article presents the Family of *Ethnic Power Relations* (EPR) Datasets, version 2014, which offers scholars a set of unique instruments to deepen their knowledge about ethnic conflict. The EPR Dataset Family consists of five distinct components built around the EPR Core dataset, which provides annual data on politically relevant ethnic groups and their access to state power from 1946 to 2013. The Core dataset is complemented by geo-spatial information on ethnic groups' settlement patterns, data on ethnic groups' links to rebel organizations, on the trans-border relations of ethnic groups, and on intra-ethnic cleavages.

The EPR Family 2014 offers several new features. In addition to extending the coverage to 2013, it introduces a new, two-level notion of state power that explicitly distinguishes between access to power at the level of the central state and political power at the regional level. Moreover, it presents new structural data on ethnic groups' intra-ethnic cleavages. Finally, for the first time, detailed documentation of the data is provided through the *EPR Atlas*, available from the Geographic Research On War, Unified Platform (GROW[up]).[1]

The purpose of this paper is to present these novelties and to outline the structure and functionality of the entire dataset family. After briefly reviewing existing data on ethnic groups, we introduce the newly added members of the EPR Family 2014 in detail, and explain how researchers can access these data through GROW[up]. We then compare the EPR Family to the most relevant alternative datasets on politicized ethnicity, in terms of both coverage and included variables, before ending with concluding thoughts on future data collection efforts.

Previous Datasets on Ethnicity

The concept of ethnicity first found its way into quantitative research in economics, notably in studies of corruption (e.g. Mauro 1995) and economic growth (e.g. Easterly and Levine 1997). Subsequently, scholars turned their attention to the relationship between ethnic diversity and civil war onset (Collier and Hoeffler 2002, 2004; Fearon and Laitin 2003; Sambanis 2001). These studies all rely on an ethno-linguistic fractionalization (ELF) index derived from the Soviet *Atlas Narodov Mira* (USSR 1964), which focuses on linguistic differences. Hence, they conceptualize ethnicity as a purely demographic concept, measured as the time-invariant degree of linguistic heterogeneity of country populations.

While some fractionalization measures explicitly take into account other sources of ethnic diversity, such as religious or racial differences (e.g. Alesina et al. 2003; Fearon 2003; Vanhanen 1999), they are associated with important limitations. For one thing, these indices disregard the political relevance of ethnic groups (Chandra and Wilkinson 2008; Posner 2004). Moreover, they say nothing about the political power relations between groups as such individualistic conceptualizations of ethnicity ignore the crucial role of the state in shaping groups' access to political power and material resources (Cederman and Girardin 2007).

Adopting an alternative approach, Gurr et al. (1993) collected data on discrimination against ethnic minorities, their grievances, levels of political mobilization, and rebellious activities (see also Gurr 2000). Their Minorities at Risk (MAR) dataset broke new ground as regards the political relevance of ethnic groups and soon became the standard source for studies of ethnic mobilization, protest, and ethnic group rebellion (see, e.g., Birnir 2007; Gurr and Moore 1997; Ishiyama 2009; Olzak 2006; Walter 2006).

However, the MAR dataset's focus on minorities "at risk" limits its applicability. Although featuring some "advantaged" minorities, this dataset remains incomplete since it does not include majority groups, some of which may be at risk of being challenged by disadvantaged minorities or are indeed discriminated against themselves in regimes of ethnic minority rule. Furthermore, the data do not include politically relevant minority groups that are not considered "at risk." For inferential analyses, this particular sample composition is unproblematic as long as the outcome of interest is unrelated to the mechanism of group selection applied in MAR, but may lead to biased results otherwise (Hug 2013).[2]

Inspired by the pioneering approach of the MAR dataset, the *Ethnic Power Relations* (EPR) dataset provides a more comprehensive selection of politically relevant ethnic groups, including minority *and* majority, and discriminated as well as state-controlling groups. EPR version 1.1 was introduced by scholars from ETH Zurich and the University of California in Los Angeles (UCLA) on the basis of an online expert survey (Cederman, Wimmer, and Min 2010). The EPR dataset defines ethnicity as a subjectively experienced sense of commonality based on a belief in common ancestry and shared culture (Weber 1976). Different markers may be used to indicate such shared ancestry and culture, such as a common language, similar phenotypical features, or adherence to the same faith. An ethnic group is considered politically relevant if at least one political organization has claimed to represent its interests at the national level or if its members are subjected to state-led political discrimination (Cederman, Wimmer, and Min 2010, 99). Drawing on the EPR data, several recent studies have generated new empirical evidence on the role of ethnicity in civil wars (see e.g. Asal et al. forthcoming; Cederman, Weidmann, and Gleditsch 2011; Cederman, Wimmer, and Min 2010; Fjelde and Hultman 2013; Roessler 2011).

The EPR Dataset Family 2014

This section outlines the structure of the latest EPR version, the EPR Dataset Family 2014, and presents its main novelties. The term "dataset family" emphasizes the fact that since its original release EPR has grown into a comprehensive system of tightly integrated datasets. By referring to the year of release, it also leaves behind the old versioning system that focused on the individual dataset components, and gives a more precise indication of the data's temporal reach.

The 2014 version builds directly on EPR-ETH version 2.0.[3] It offers several new features:

- The temporal scope is extended by four years from 2009 to 2013.[4]
- Regional autonomy is now coded for all groups regardless of whether they are included

    or excluded at the national level.
- The *EPR Atlas* offers detailed documentation of the data, including graphic presentation

    and explanations of the main coding decisions as well as references to the sources.
- The EPR-Ethnic Dimensions (EPR-ED) component dataset identifies linguistic, religious,

    and racial segments of all ethnic groups included in EPR.
- The "Trans-border Ethnic Kin" (EPR-TEK) data is fully integrated in the dataset

    structure.

Figure 1 shows the organization of the EPR Family resulting from these new features.

[Figure 1]

At the country level, the EPR Family 2014 includes all states with a population, in 1990, of at least 500,000, and in which ethnicity has been politically relevant.[5] For all these states, the EPR Core dataset identifies the politically relevant ethnic groups and codes their access to state

power. GeoEPR provides geo-spatial information on the settlement patterns of these groups.[6] Furthermore, the ACD2EPR component links the EPR groups to the rebel organizations in UCDP's actor database, while the EPR-TEK dataset identifies trans-border ethnic connections between EPR groups in different countries. Finally, at the sub-group level, the EPR-ED dataset codes linguistic, religious, and racial segments of the EPR groups.

The data of all components of the EPR Family 2014 were collected by regional experts and research assistants. Each coding was then evaluated by the EPR Management Committee and in region-specific workshops regarding its consistency with the globally applied coding rules. In some cases of disagreements, we consulted additional country experts.[7] The following sections will present the EPR Core dataset and those individual components that were newly added to the data universe during the recent update process before explaining how the data can be accessed through the online platform GROW[up].

The EPR Core Dataset 2014

The Core dataset of the EPR Family 2014 provides annual data on politically relevant ethnic groups, their relative sizes as a share of the total population, and their access to state power. Power access is measured with an ordinal scale composed of three main categories, depending on whether a group (1) controls power alone, (2) shares power with other ethnic groups, or (3) is excluded from executive state power. Each of these three main categories is divided into several sub-categories:

1. **The group rules alone**: *monopoly* or *dominant*. In contrast to monopoly power, the status of dominant indicates "token" representation of other ethnic groups in the executive.

2. **The group shares power**: *senior partner* or *junior partner,* depending on the group's absolute influence in the executive (i.e. irrespective of group size).
3. **The group is excluded**: *powerless*, *discriminated*, or *self-exclusion*. While powerless means that the group is simply not represented (or does not have influence) in the executive, discrimination indicates an active, intentional, and targeted discrimination by the state against group members in the domain of public politics. The special category of self-exclusion applies to groups that have excluded themselves from central state power, in the sense that they control a particular territory of the state which they have declared independent from the central government.[8]

Groups falling into one of the first two main categories can be regarded as politically *included* in distinction to the *excluded* groups in the third main category. The new version identifies a total of 819 ethnic groups that were politically relevant at one point or the other during the time period from 1946 to 2013.

In addition to the national power variable, the updated EPR Core Dataset measures access to executive power at the regional level with a separate *regional autonomy* variable. In previous versions of EPR, regional autonomy status was coded as a subcategory of exclusion at the national level. However, this coding scheme does not allow identifying those groups that are both included at the national level *and* enjoy regional autonomy. To solve this problem, the current version disentangles the regional level of political power from power access to the state's executive. Groups that only have access to power at the sub-state level are coded as "powerless" at the national level. Included groups in a power-sharing regime (that is, those coded as "senior partner" or "junior partner") may or may not simultaneously enjoy regional autonomy.[9]

For a group to be coded as regionally autonomous, two conditions must be jointly satisfied. First, there must be a meaningful[10] and active regional executive organ that operates below the state level (for example, the departmental, provincial, or district level) but above the local administrative level, *and* group representatives must exert actual influence on the decisions of this entity, acting in line with the group's local interests.

The second condition also implies that a given regional entity must have *de facto* (as opposed to mere *de jure*) political power. Federal states, such as Switzerland or India, are the most typical (but not the only) such systems of regional autonomy. The Kurdistan Regional Government in northern Iraq is another example of meaningful political power at the sub-state level (Katzman 2010). In contrast, the regional administrative subdivisions in many Central and Eastern European countries do not possess any political or fiscal powers and thus cannot be considered meaningful political decision-making bodies. Furthermore, non-territorial forms of autonomy (such as the recently established minority councils in Serbia) do not fall under this definition of regional autonomy.[11]

Table 2 compares the relative frequency of regional autonomy between excluded and power-sharing groups. It shows that autonomy is more frequent among excluded groups. This suggests that ethnic power sharing typically happens along one dimension alone: *either* in the form of a "grand coalition" (Lijphart 1977) at the national level *or* in some sort of territorial power sharing. Nevertheless, the table also shows that a sizable number of autonomous groups were not captured by the former coding scheme, underlining the importance of this change in the data structure of the new EPR version.

[Table 1]

Identifying Trans-border Ethnic Kin Relations: EPR-TEK 2014

While previously separate from the rest of the EPR data, the Trans-border Ethnic Kin (EPR-TEK) dataset has been fully integrated in the new EPR Family 2014 in the context of the update process. EPR-TEK identifies all EPR groups with settlements in at least two countries through nominal matching. Thus, groups in different countries are coded as trans-border kin if they share the same ethnographic name, including synonyms. The Kurds who live in Turkey, Iran, Iraq, and Syria are a typical example of such a transnational ethnic group (see Figure 2).

[Figure 2]

Overall, EPR-TEK identifies 149 unique TEK groups. Out of the 819 EPR groups, 418 have at least one TEK connection to another country. Many TEK groups are spread across a multitude of states, such as the Arabs or Russians. Moreover, only in fifteen countries worldwide, EPR groups lack any cross-border links to ethnic groups in other countries. This attests to the relevance of trans-border ethnic links and of the EPR-TEK dataset itself, which allows researchers to account for this transnational dimension in their analyses of ethnic conflict.

Coding Intra-ethnic Identities and Cleavages: EPR-ED 2014

The most recent addition to the EPR Dataset Family is the Ethnic Dimensions (EPR-ED) data, which identify the linguistic, religious, and racial segments of all EPR groups.[12] This is the first dataset that codes both multiple cleavage dimensions and several segments within a cleavage

9

dimension for ethnic groups. For each group a maximum of three linguistic and religious segments are reported, along with their relative sizes (as a share of the total group population). In the case of race, the segments indicate miscegenation by denoting up to three different racial origins of a given ethnic group.

Figure 3 illustrates the basic setup of the EPR-ED data with two ethnic groups from Nigeria: the Hausa-Fulani from the northern part of the country and the Yoruba from the southwest. The Hausa are a religiously homogeneous but linguistically divided group, while the Yoruba are united by their language, yet religiously divided between Christians and Muslims. The largest religious segment of the Yoruba is made up by Sunni Muslims from the Maliki background, the same religion practiced by the Hausa group. Thus, on the religious dimension there is a substantial minority among the Yoruba that shares a trait with the Hausa. On the racial dimension, EPR-ED does not differentiate the two groups as they both originate from Sub-Saharan Africa.

[Figure 3]

Overall, the dataset identifies 635 unique language segments in the Ethnologue database (Lewis 2009) that are spoken by the 819 EPR groups. Relying on the Joshua Project (2011), EPR-ED codes 71 distinct religious creeds that ethnic groups adhere to. Finally, the data distinguish between eight different regional origins that have become relevant as social categories through European colonization of the world.

Accessing the EPR Dataset Family

All individual components of the EPR Dataset Family 2014 can be downloaded from the

GROW[up] web portal at http://growup.ethz.ch. GROW[up] provides a user-friendly Research Front-

End (RFE), which allows users to assemble and download customized panel datasets composed

of EPR-related variables. The data offered via the RFE are pre-aggregated to the level of group-

years and country-years. Formatted in order to facilitate statistical analysis, the selection of

variables includes conflict onset and incidence dummies, as well as various peace-years variables

and other temporally defined conflict indicators.[13]

Comparison with Other Datasets on Politicized Ethnicity

This section compares the EPR Family 2014 to three alternative datasets on ethnicity and

discusses the potential limitations of the EPR data in empirical applications. Table 1 lists the

main characteristics of the EPR Family, the MAR Project, two samples of the A-MAR dataset,

and Fearon's (2003) list of ethnic groups.[14]

Table 1: The EPR Family 2014 and Its Alternatives. Comparison of Coverage and Variables

	EPR Family 2014	Minorities at Risk (MAR)	A-MAR group list	A-MAR sub-sample	Fearon (2003)
Sample					
Inclusion criteria	Discrimination or political representation	Discrimination or mobilization	Social relevance	Stratified, representative sample of new groups included in A-MAR	Countries' main ethnic groups
N groups	819	342[a)]	1157	74	822
N countries	141 (165)[b)]	123[a)]	169	55	160
Time horizon	1946-2013	1940-2006[c)]	-	???	-

11

Variables					
Group size	Yes	Yes	No	Yes	Yes
Spatial extension	Yes	No[d]	No	No	No
Violent conflict	Yes	Yes	-	Yes	-
Political power	Yes	Yes	-	Yes	-
Economic inequality	Yes[e]	Yes	-	Yes	-
Cultural rights	No	Yes	-	Yes	-
Non-violent mobilization	No	Yes (group organization and protest)	-	Yes (group organization and protest)	-
Trans-border ethnic links	Yes	No[f]	-	No[f]	-
Intra-ethnic cleavages	Yes	No[g]	-	No[g]	-
Documentation of codings	Yes	Yes	-	???	-

Notes: a) In the 2004-2006 data of MAR, these numbers dropped to 284 groups in 117 countries.

b) 141 groups in countries where EPR considers ethnicity to be politically relevant. Where this is not the case, EPR lists one national group as placeholder, which increases the total number of included groups to 165. These national groups are relevant for the coding of trans-border ethnic links in the EPR-TEK data.

c) Until 1985, the MAR variables are only coded in 5-year intervals, and some of the variables are not coded over the full time period.

d) The MAR dataset contains categorical and ordinal variables on groups' settlement patterns, such as their spatial distribution.

e) Cederman, Weidmann, and Gleditsch (2011) have derived estimates of horizontal economic inequality, by combining ethnic groups' settlement patterns from the GeoEPR dataset and geo-referenced income data by Nordhaus (2006).

f) While the MAR data code whether ethnic groups have transnational links (and whether kin groups hold political power in their countries), they do not identify these kin groups.

g) The MAR data provide measures of group distinctiveness to the country's majority group, but they do not identify the different linguistic, religious, and racial segments of each group.

As a result of their different inclusion criteria, these datasets cover quite different and differently sized samples of ethnic groups. Yet, all of these data sources rely on ethnic groups as their units of analysis, which makes them subject to the criticism of reifying and attributing power of agency to mere social categories (see Brubaker 2004).[15] Although ethnic identities may change

over time, there is usually sufficient group cohesion for ethnic groups to be considered stable identity categories within the time spans of conflict processes (Cederman, Gleditsch, and Buhaug 2013, 23). Nevertheless, because ethnic groups are treated as monolithic, none of these datasets is able to reveal potentially competing agendas and claims of different organizational representatives of the same group.[16]

In terms of temporal coverage, the EPR Family 2014 offers the most up-to-date codings of ethnic groups and their social and political characteristics, ranging up to 2013. The original MAR dataset has a similar temporal coverage although until 1985, the variables are only coded in 5-year intervals, and some of the variables are not coded over the full time period. In contrast, Fearon's data and the new A-MAR group are time-invariant.

There are also considerable differences between these datasets with regard to the variables included, as Table 1 reveals. Compared to the original MAR dataset, the EPR Family 2014 offers less information on group characteristics, such as their level and strategies of political mobilization, but includes a much more extensive sample of ethnic groups. Moreover, the GeoEPR dataset provides geocoded maps of the settlement areas of all EPR groups, while the EPR-TEK and EPR-ED extensions offer additional information on groups' trans-border links and intra-ethnic divisions that is not available in MAR. Although many classes of variables are found in both datasets, there are important differences in the type of measures employed to capture similar theoretical concepts. For example, in terms of groups' political status MAR relies on a political discrimination index, whereas EPR measures groups' access to state power over time using an ordinal classification of power status. With respect to the ethnic conflict coding EPR offers onset and incidence variables whereas MAR uses an ordinal index of the level of group rebellion.[17] On balance, the two datasets have distinct strengths and weaknesses as regards the

13

variables they include. Yet, the EPR Family 2014 offers a more extensive geographic and temporal coverage than MAR. In addition, it has the advantage of providing researchers with geocoded data.

It should be noted that although including a much larger sample of ethnic groups than the original MAR dataset, the EPR Family 2014 still only codes a subset of all ethnic identities. The selection criterion of political relevance ensures that groups included in EPR can reasonably be assumed to possess a minimal level of political agency. However, the pre-selection of relevant groups is associated with two potential shortcomings for empirical applications. First, the data are not well suited for the analysis of ethnic mobilization processes, since the sample does not include ethnic identities that are not (yet) politically relevant. Second, empirical analyses employing EPR-coded ethnic groups as units of analysis may be subject to selection bias (Heckman 1979). This is the case if unmeasured variables exist affecting both the studied group-level outcome *and* the probability of a given ethnic group of being coded as politically relevant. However, the issue of selection bias depends on the particular research question at hand and cannot be determined *a priori* (Hug 2013). In particular, group-level analyses that focus on outcomes unrelated to the employed selection criteria are unlikely to be biased (Hug 2013, 199-201). Moreover, selection bias can also be addressed by replicating group-level studies using different units of analysis (for example, countries or geographically defined sub-national units) onto which the EPR data can be mapped, but which themselves are not subject to selection.

The new A-MAR data were collected in response to these concerns. As Table 1 shows, A-MAR comprises an even larger sample of ethnic groups than the EPR Dataset Family. Yet, at the current state it consists only of a list of socially relevant ethnic groups of each country, lacking any information on groups' social and political characteristics. So far the A-MAR authors have

coded the full range of MAR variables for a stratified but representative sample of 74 new ethnic groups that were not previously included in MAR and, thus, not selected according to the political relevance criterion. These groups can be matched with the original MAR groups to arrive at a new, theoretically unbiased sample of ethnic groups (Birnir et al. 2012, 15-7). Although still limited to the group list, the new A-MAR data constitute an important instrument to examine potential selection biases in quantitative studies of ethnic conflict. Nevertheless, at the current stage, the variables relevant to conflict researchers are only coded for a relatively small sample of groups, which for many empirical applications may not be sufficient. In contrast, the EPR Dataset Family offers full (including geo-spatial) information for all groups included.[18]

Finally, while Fearon's data encompass a slightly higher number of countries and ethnic groups, they possess two crucial disadvantages vis-à-vis the EPR Dataset Family besides their time-invariance. First, they do not provide any information on groups' social and political characteristics beyond demographic size and, second, they are limited no non-spatial representations.

Conclusion and Outlook

The EPR Dataset Family 2014 provides conflict researchers with new instruments to improve their understanding of ethnic conflict processes. Thanks to its extensive scope, this system of datasets facilitates the study of multiple causes of ethnic conflict. However, much work still remains to be done with regard to the quality of quantitative data on ethnic politics. In this regard, there are three promising directions of future data collection. First, it would be useful to identify the actual political actors that are at the roots of the collective action undertaken by ethnic groups. For this purpose, we plan to launch a new data collection project, the EPR-

15

Organizations Dataset, which will offer information on ethnically based political organizations in all countries of the world. The goal is to break up the monolithic concept of ethnic groups by taking into account the diverse agendas and claims of different organizational representatives, and to study ethnic mobilization processes.

Second, broadening the spectrum of political violence covered by EPR beyond civil war, upcoming extensions of the ACD2EPR dataset will link EPR groups to all UCDP actors, including those involved in one-sided violence, non-state conflict, and interstate war. Finally, going beyond the nominal coding of EPR-TEK, future versions of the dataset will adhere to a claim-based coding that allows for changes in transnational ethnic identifications.

While these and similar projects promise significant improvements in the quality of data on politicized ethnicity, for the time being the EPR Datasets Family enables researchers to study complex research problems related to politicized ethnicity and political violence within a consistent data structure. Future updates will ensure that the data remain relevant to conflict analysis for years to come.

Notes

References

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8 (2):155-194.

Asal, Victor, Michael Findley, James A. Piazza, and James I. Walsh. forthcoming. "Political Exclusion, Oil, and Ethnic Armed Conflict." *Journal of Conflict Resolution*.

Asal, Victor, Amy Pate, and Jonathan Wilkenfeld. 2008. Minorities at Risk Organizational Behavior Data and Codebook Version 9/2008. Available from http://www.cidcm.umd.edu/mar/data.asp (accessed March 1, 2015).

Birnir, Jóhanna Kristín. 2007. *Ethnicity and Electoral Politics*. New York: Cambridge University Press.

Birnir, Jóhanna Kristín, Jonathan Wilkenfeld, James D. Fearon, David D. Laitin, Ted Robert Gurr, Dawn Brancati, Stephen M. Saideman, Amy Pate, and Agatha S. Hultquist. 2015. "Socially Relevant Ethnic Groups, Ethnic Structure, and AMAR." *Journal of Peace Research* 52 (1):110–115.

Birnir, Jóhanna, Jonathan Wilkenfeld, James Fearon, David Laitin, Ted Robert Gurr, Dawn Brancati, Stephen Saideman, and Amy Pate. 2012. The Reason We Don't Yet Know "Why Groups Rebel": Introducing A-MAR (All-Minorities at Risk). Paper prepared for the Comparative Politics Workshop, Stanford University, February 23.

Brubaker, Rogers. 2004. *Ethnicity Without Groups*. Cambridge, MA: Harvard University Press.

Cederman, Lars-Erik, and Luc Girardin. 2007. "Beyond Fractionalization: Mapping Ethnicity onto Nationalist Insurgencies." *American Political Science Review* 101 (1):173-185.

Cederman, Lars-Erik, Luc Girardin, and Julian Wucherpfennig. 2014. "Exploring Inequality and Ethnic Conflict: EPR-ETH and GROW[up]." In *Peace and Conflict 2014*, edited by D. A. Backer, J. Wilkenfeld and P. K. Huth. Boulder, CO: Paradigm and University of Maryland. 74-87.

Cederman, Lars-Erik, Kristian Skrede Gleditsch, and Halvard Buhaug. 2013. *Inequality, Grievances and Civil War*. New York: Cambridge Univerisity Press.

Cederman, Lars-Erik, Simon Hug, Andreas Schädel, and Julian Wucherpfennig. forthcoming. "Territorial Autonomy in the Shadow of Conflict: Too Little, Too late?" *American Political Science Review*.

Cederman, Lars-Erik, Nils B. Weidmann, and Kristian Skrede Gleditsch. 2011. "Horizontal Inequalities and Ethnonationalist Civil War: A Global Comparison." *American Political Science Review* 105 (3):478-495.

Cederman, Lars-Erik, Andreas Wimmer, and Brian Min. 2010. "Why Do Ethnic Groups Rebel? New Data and Analysis." *World Politics* 62 (1):87-119.

Chandra, Kanchan, and Steven Wilkinson. 2008. "Measuring the Effect of "Ethnicity"." *Comparative Political Studies* 41 (4/5):515-563.

Collier, Paul, and Anke Hoeffler. 2002. "On the Incidence of Civil War in Africa." *Journal of Conflict Resolution* 46 (1):13-28.

———. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56 (4):563-595.

Easterly, William, and Ross Levine. 1997. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *The Quarterly Journal of Economics* 112 (4):1203-1250.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth* 8 (2):195-222.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *The American Political Science Review* 97 (1):75-90.

Fjelde, Hanne, and Lisa Hultman. 2013. "Weakening the Enemy A Disaggregated Study of Violence against Civilians in Africa." *Journal of Conflict Resolution* 58 (7):1230-1257.

Gleditsch, Kristian Skrede, and Michael D. Ward. 1999. "Interstate System Membership: A Revised List of the Independent States since 1816." *International Interactions* 25 (4):393-413.

Gurr, Ted Robert. 2000. *Peoples versus States: Minorities at Risk in the New Century*. Washington, D.C.: United States Institute of Peace.

Gurr, Ted Robert, Barbara Harff, Monty G. Marshall, and James R. Scarritt. 1993. *Minorities at Risk: A Global View of Ethnopolitical Conflicts*. Washington, D.C.: United States Institute of Peace Press.

Gurr, Ted Robert, and Will H. Moore. 1997. "Ethnopolitical Rebellion: A Cross-Sectional Analysis of the 1980s with Risk Assessments for the 1990s." *American Journal of Political Science* 41 (4):1079-1103.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the Econometric Society*:153-161.

Hug, Simon. 2013. "The Use and Misuse of the "Minorities at Risk" Project." *Annual Review of Political Science* 16:191–208.

Ishiyama, John. 2009. "Do Ethnic Parties Promote Minority Ethnic Conflict?" *Nationalism and Ethnic Politics* 15 (1):56-83.

Joshua Project. *Unreached Peoples of the World.* 2011. Available from http://www.joshuaproject.net/ (accessed December 4, 2014).

Katzman, Kenneth. 2010. The Kurds in Post-Saddam Iraq. *CRS Report for Congress* (RS22079), http://www.fas.org/sgp/crs/mideast/RS22079.pdf (accessed June 2, 2014).

Lewis, M. Paul, ed. 2009. *Ethnologue: Languages of the World*. Vol. 16. Dallas, TX: SIL International.

Lijphart, Arend. 1977. *Democracy in Plural Societies: A Comparative Exploration*. New Haven: Yale University Press.

Mauro, Paolo. 1995. "Corruption and Growth." *The Quarterly Journal of Economics* 110 (3):681-712.

Nordhaus, William D. 2006. "Geography and Macroeconomics: New Data and New Findings." *Proceedings of the National Academy of Sciences* 103 (10):3510-3517.

Olzak, Susan. 2006. *The Global Dynamics of Racial and Ethnic Mobilization* Stanford, CA: Stanford University Press.

Posner, Daniel N. 2004. "Measuring Ethnic Fractionalization in Africa." *American Journal of Political Science* 48 (4):849–863.

Roessler, Philip G. 2011. "The Enemy Within: Personal Rule, Coups, and Civil War in Africa." *World Politics* 63 (2):300-346.

Sambanis, Nicholas. 2001. "Do Ethnic and Nonethnic Civil Wars Have the Same Causes? A Theoretical and Empirical Inquiry (Part 1)." *Journal of Conflict Resolution* 45 (3):259-282.

USSR. 1964. *Atlas Narodov Mira*. Moscow: Glavnoe upravlenie geodezii i kartografii.

Vanhanen, Tatu. 1999. "Domestic Ethnic Conflict and Ethnic Nepotism: A Comparative Analysis." *Journal of Peace Research* 36 (1):55-73.

Wade, Peter. 2010. *Race and Ethnicity in Latin America*. 2nd ed. London: Pluto Press.
Walter, Barbara F. 2006. "Information, Uncertainty, and the Decision to Secede."
    International Organization 60 (1):105-135.
Weber, Max. 1976. *Wirtschaft und Gesellschaft: Grundriss der verstehenden
    Soziologie*. 5th ed. Tübingen: J.C.B. Mohr.
Wucherpfennig, Julian, Nils Metternich, Lars-Erik Cederman, and Kristian S.
    Gleditsch. 2012. "Ethnicity, the State, and the Duration of Civil Wars." *World
    Politics* 64 (1):79-115.

Table 2: Frequency of regional autonomy among included and excluded groups, 1946-2013

	No autonomy	Regional autonomy	Total
Excluded groups	16,169	5,812	21,981
	(74%)	(26%)	(100%)
Included groups	8,420	2,180	10,600
	(79%)	(21%)	(100%)
Total	24,589	7,992	32,581
	(75%)	(25%)	(100%)

Notes: The figures in each cell indicate the absolute number of observations in the respective category as well as the row

percentages. The unit of analysis is the group year.

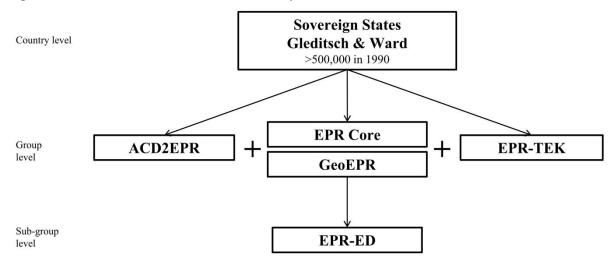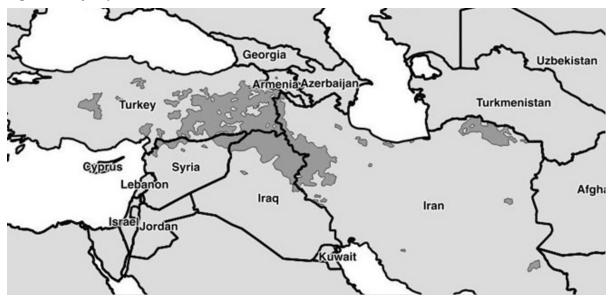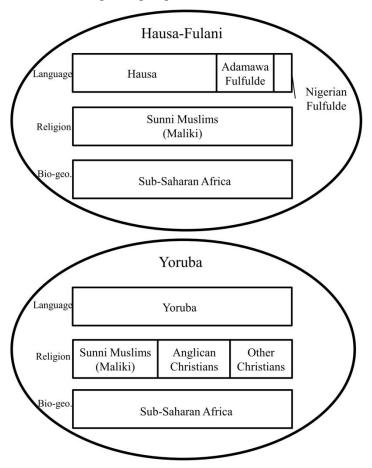**Figure 1: The structure of the EPR Dataset Family 2014**

Country level

┌─────────────────────────┐
│   **Sovereign States**   │
│   **Gleditsch & Ward**   │
│   >500,000 in 1990   │
└─────────────────────────┘

Group level

| **ACD2EPR** | + | **EPR Core** / **GeoEPR** | + | **EPR-TEK** |

Sub-group level

EPR-ED

Figure 2: Majority settlement areas of the Kurds

Note: Based on maps from GeoEPR 2014.

Figure 3: Ethnic identities of two Nigerian groups

1 See http://growup.ethz.ch/. For a detailed overview of GROW^up, see Cederman, Girardin, and Wucherpfennig (2014).

2 In response to these concerns, the MAR research team recently presented a new dataset of ethnic groups, labeled A-MAR ("All Minorities at Risk"), that introduces a comprehensive list of "socially relevant" ethnic identities by relying on a mainly "cultural" definition of (potentially) relevant ethnic groups (Birnir et al. 2015). We will discuss this dataset in more detail below.

3 The EPR-ETH version 2.0 was compiled by a team at ETH Zurich, extending the coverage of the original data from 2005 to 2009 and thoroughly revising the codings for various countries. Subsequently, a team of researchers at UCLA recently assembled their own version of the dataset (labeled EPR version 3) that builds directly on version 1.1 and extends it to 2010. However, because it is currently limited to non-spatial data and incompatible with the other EPR-related datasets, this version was not taken into account in the composition of the new EPR Family 2014.

4 Thanks to the lower population threshold for case selection introduced in EPR-ETH 2.0, the EPR Family 2014 also extends the geographic coverage of the original EPR data. The following countries were added to the dataset: Bahrain, Bhutan, Cyprus, Djibouti, Fiji, Guyana, Mauritius, and Singapore.

5 Note that the EPR Family 2014 draws on Gleditsch and Ward (1999) for its list of sovereign states.

6 In addition to extending the coverage to 2013, the new GeoEPR version now distinguishes between "statewide" groups that constitute a majority throughout the entire territory of a state and "dispersed" groups: minority groups whose members are scattered throughout different regions of a state, such as the Jews in Poland. Previously, these groups were both combined in the same category of "dispersed" groups.

7 In the case of the ACD2EPR data, we are also grateful for excellent assistance from the UCDP team at Uppsala University.

8 Note that this category was labeled "separatist autonomy" in previous EPR versions.

9 Note that the autonomy dimension is not coded for "monopoly" and "dominant" groups since their political interests are assumed to be sufficiently represented at the level of the central state.

10 The term "meaningful" here refers to executive organs that carry out core competencies of the state, involving, for example, cultural rights (language and education) and/or significant economic autonomy (e.g. the right to levy taxes, or very substantial spending autonomy).

11 See Cederman et al. (forthcoming) for more information on the regional autonomy coding.

12 Instead of resorting to assumptions about differences in the physical appearance of individual ethnic group members, the EPR-ED dataset defines race as ethnic groups' origins from particular world regions, such as Europe, Sub-Saharan Africa, Oceania, etc. These regional origins – expressed at the individual level through certain phenotypical markers (particularly skin color) – have become relevant as social categories in the context of European colonization of the world and the related process of racial classification (see e.g. Wade 2010, 5-19).

13 Since EPR takes into account shifts in the relevant ethnic cleavages, ethnic groups may split into different, politically relevant sub-groups or, reversely, lower-level ethnic categories may become politically relevant as parts of an overarching umbrella category. This means that the construction of "historical" variables for EPR groups, such as the number of ethnic conflicts a group has experienced, is not trivial. Yet, the data available to researchers through GROW[up] provide variables that keep track of a group's history even across changes in group hierarchy. For instance, they include a variable indicating the number of conflict onsets a group *or* any of its (potential) ancestors has ever experienced.

14 Since the sub-sample of groups, for which A-MAR provides variable codings, is currently not yet publicly available, the information in Table 1 is based on Birnir et al. (2012).

15 Note that the Minorities at Risk Organizational Behavior (MAROB) dataset, a subsidiary of the MAR Project, uses ethno-political organizations as units of analysis (Asal, Pate, and Wilkenfeld 2008). However, at the current stage it is limited to 26 countries of the Middle East and North Africa.

16 Note, however, that this is not true for conflict periods since the ACD2EPR dataset identifies multiple rebel organizations fighting for the same ethnic group (Cederman, Gleditsch, and Buhaug 2013, ch. 8; Wucherpfennig et al. 2012).

17 In addition, the MAR dataset includes variables on intercommunal conflict and government repression.

18 Note also that using the new A-MAR data, Birnir et al. (2012) replicated the main results of Cederman, Wimmer and Min (2010), providing evidence against selection bias in the EPR data.