

Addressing group-based inequalities

Ethnic conflict, tensions, and protests: taking stock of available cross-country data

Ines A. Ferreira*

December 2021

Abstract: This note provides an overview of available cross-country data on ethnic conflict, tensions, and protests. First, it documents the steps taken in the selection of 16 different relevant data sources, before they are briefly described, and their distinct aspects are highlighted. Second, it includes some descriptive analysis of selected indicators, as well as comparisons between them. Overall, it suggests that data on ethnic armed conflict is easily accessible, especially through the GROW^{mp} – Geographical Research on War, Unified Platform. Moreover, the updated version of the Minorities at Risk (MAR) and accompanying datasets will also provide useful data. However, data on ethnic tensions, protests, and other types of violent and non-violent events (including the actors involved) is currently scattered across different sources, with differences in coverage and in the information provided.

Key words: data, ethnic conflict, ethnic tensions, ethnic protests

JEL classification: D74

Acknowledgements: I am grateful to Rachel M. Gisselquist for comments on an early draft. The usual caveats apply.

* University of Copenhagen, Copenhagen, Denmark; if@econ.ku.dk

This study has been prepared within the UNU-WIDER project [Addressing group-based inequalities](#).

Copyright © UNU-WIDER 2021

UNU-WIDER employs a fair use policy for reasonable reproduction of UNU-WIDER copyrighted content—such as the reproduction of a table or a figure, and/or text not exceeding 400 words—with due acknowledgement of the original source, without requiring explicit permission from the copyright holder.

Information and requests: publications@wider.unu.edu

<https://doi.org/10.35188/UNU-WIDER/WTN/2021-22>

**United Nations University World Institute for Development
Economics Research**



Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

The views expressed in this paper are those of the author(s), and do not necessarily reflect the views of the Institute or the United Nations University, nor the programme/project donors.

1 Introduction

This note provides a stock-taking of available cross-country data on ethnic conflict, tension, and protest. Before giving more details about the selected datasets (Section 2) and showing some descriptive analysis and comparison between the variables (section 3), the following paragraphs make a brief note on definitions and give a succinct overview of the steps taken to select the datasets.

1.1 Definitions

Given the purpose of taking stock of existing data, a broad approach was taken to the concepts of ethnic conflict, tension, and protest. Ethnicity is understood here to consider a sense of shared identity on the basis of different aspects, particularly race, language, and religion.¹ In the process of selecting data during the initial search on ethnic conflict, different types of violent and non-violent events were considered.

Among the datasets selected, several variables focus on armed conflict.² A first insight from this exercise is that there is significant variation in the definitions and coding of ethnic conflict across datasets (see Table A1 in the appendix).³ Most of them base their definition on conflict events in which the actors/organizations involved have been linked to ethnic groups. In some cases, this link is conditional on ethnic recruitment and claims (e.g. GROW^{up} – Geographical Research on War, Unified Platform [GROW^{up}]), while in others the link is simply based on the ethnic composition of conflicting parties (e.g. Categorically Disaggregated Conflict [CDC]). Others focus on conflict or violence of ethnic or religious nature (e.g. Institutional Profiles Database [IPD]).⁴

Other variables cover different types of events, including protests.⁵ Ethnic tensions were also understood broadly to consider divergence between different ethnic groups, but it proved more difficult to find data on this dimension. While one of the variables selected covers ethnic and religious tensions, it has also been used, for example, as a measure of ethnic division (Keefer and Knack 2002). Indicators of ethnic fractionalization or concentration are not covered in this note.

1.2 Steps for data selection

The first step consisted in compiling an overview of useful resources, achieved through three types of purposive search:

- 1) Scoping reading based on different Google and Google Scholar searches on ‘ethnic conflict’, as well as snowballing from different articles.

¹ While most of the selected datasets cover different elements of ethnicity, one focuses only on religious groups.

² Some follow the Uppsala Conflict Data Program’s (UCDP) definition of armed conflict.

³ In fact, different elements of the concept and even its importance have been contested (see Gilley 2004; Gurr 2017; Horowitz 2008).

⁴ Note also that some definitions are based on the criteria of politically significant ethnic division, namely variables from UCDP datasets.

⁵ Additional types of events include rebellion or repression, terrorism, and considering a broader scope for ethnic violence, genocides/politicides.

- 2) Consultation of existing related compilations (e.g. Forsberg et al. 2012).
- 3) Consultation of well-known datasets on institutions and conflict (e.g. Uppsala Conflict Data Program [UCDP]).
- 4) Manual browsing of article titles (2017–2021, inclusive) and lists of special issues in relevant journals, such as *Journal of Conflict Resolution*.

The next steps consisted in consolidating the results from the different searches into a list of datasets and papers, after eliminating the dataset duplication and excluding some sources. Sources excluded did not meet one or several of the following criteria: (i) they could not be accessed at the time of writing (e.g. Country Indicators for Foreign Policy [CIFP]); (ii) they were too broad (e.g. indicator of social, ethnic, and religious conflict from the BTI Transformation Index by Bertelsmann Stiftung); (iii) the last data point included was from before 2001 (e.g. datasets in Fearon and Laitin 2003 and Sambanis 2001). Additionally, datasets including only geocodes were also not included (e.g. Spatially Interpolated Data on Ethnicity [SIDE]). Finally, while not selected for this note, reference is made to two datasets that may be relevant to the analysis of ethnic conflict focusing on organizational behaviour: the Minorities at Risk Organizational Behavior (MAROB) and the EPR-Organizations dataset.

Despite efforts to cover a broad range of sources, the resulting list of datasets makes no claims of completeness. The selected sixteen sources (Table 1) are described in the next section.

Table 1: Some characteristics of the selected data sources

Name of data source	Disaggregation level			Information sources			Data coding			
	Group	Country	Event	Perpe- trator	Qual. sources ¹	Existing dataset(s)	Primary data	1 Expert ²	>1 Experts	Software
Armed Conflict Location & Event Data Project (ACLED)			X		X				X	
Categorically Disaggregated Conflict (CDC)			X		X	X	X	X		
Ethnic conflict (Pischedda 2020)			X		X	X		X		
Ethnic One-Sided Violence dataset (EOSV)				X	X	X			X	
Fragile States Index (FSI)		X			X	X			X	X
GROW ^{up} – Geographical Research on War, Unified Platform	X	X				X			X	
GTD2EPR	X				X	X		X		
Institutional Profiles Database (IPD)		X					X		X	
Minorities at Risk (MAR)	X				X				X	
Major Episodes of Political Violence (MEPV)		X			X			X		
Political Instability Task Force (PITF)			X		X				X	
Religion and Armed Conflict (RELAC)			X		X				X	
Social Conflict Analysis Database (SCAD)			X		X				X	
Targeted Mass Killing (TMK)		X	X		X	X			X	
The Global State of Democracy Indices (GSoDI)		X				X			X	
UCDP Non-State Conflict Dataset			X		X	X			X	

Note: ¹ this refers to several types of qualitative sources, including media reporting and reports from international organizations. In certain cases (e.g. EOSV) quantitative datasets are used when they provide details on the character of violence. ² Expert(s) also include(s) the researcher(s).

Source: author's elaboration.

2 Describing available datasets

In this section, I delve more deeply into the existing datasets, describing their content and highlighting their specificities. The search process described in the previous section suggested the Minorities at Risk (MAR) and GROW^{up} datasets as the two most widely used/referred sources. I start by describing them, before turning to the remaining datasets.⁶

2.1 MAR and GROWup

The MAR dataset (CIDCM 2009, 2016) is used in many studies on ethnic conflict, including those focusing on protests. According to the project website, the focus is on the status and conflicts of politically-active communal groups in all countries with a current population of at least 500,000 and includes both qualitative and quantitative data. The criteria to define a ‘minority at risk’ was revised in the 2004–2006 update, and among different criteria, includes ‘members share some distinguishing cultural features, such as common language, religion and customs’ and a threshold level for the number of members.

In addition to different group characteristics and status, the quantitative dataset includes indicators on inter- and intra-communal conflict. Moreover, it provides data on other types of conflict, namely protest, rebellion, and repression, which are less common in other datasets. Unfortunately, while the previous release included data prior to 2004, it has been discontinued, and the current data release contains data only for the period 2004–2006. Moreover, the All Minorities at Risk (AMAR) Project—created to address the problem of selection bias identified in the MAR dataset (resulting from the selection criteria used for discrimination and/or mobilization)⁷—is also under construction, and while the codebook can be consulted, the Phase I data is not available at the time of writing.⁸

In addition to portals for data visualization and detailed documentation by country, the GROW^{up} platform (Girardin et al. 2015) offers datasets at the country-year and group-year level which compile different variables from other UCDP datasets. Focusing on conflict-related variables (compiled from UCDP Actor Dataset, Uppsala/PRIO Armed Conflict Database, and ACD2EPR), they cover ethnic onset and incidence, further disaggregated into territorial and governmental based on the incompatibility they originate from. The complete dataset matches the information to other variables characterizing the different groups, which might also be relevant in conflict analysis. Of all the datasets examined, this is the most comprehensive and enables an analysis of dynamic settlement patterns.

2.2 Datasets on conflict and institutions that include relevant variables

Among the other different datasets offered by UCDP, two are relevant for this note. The Non-State Conflict Dataset provides data on the use of armed force between two organized armed groups, neither of which is the government of a state (Pettersson et al. 2021; Sundberg et al. 2012). Focusing on ethnic conflict, the selection can be restricted to events with warring sides that are informally organized groups that share a common identification along ethnic, clan, religious,

⁶ Table A1 in the appendix summarizes the main details and includes a description of the definitions used.

⁷ See also Cederman et al. (2010) for further details on the limitations.

⁸ See Hug (2013) for a detailed discussion of the advantages and pitfalls of the MAR dataset and Gurr (2017) for the potential for using it in future research.

national, or tribal lines ('organizational level 3'), thus capturing conflict related to communal identity. While UCDP data is widely used and regularly updated, this dataset is perhaps better suited for the analysis of ethnic dimensions within the study of non-state conflict.

The second is the Ethnic One-Sided Violent Conflict (EOSV) dataset. This perpetrator-year dataset provides information on 'the ethnic identity of the victims of deliberate lethal violence against non-combatants by state and non-state actors' (Fjelde et al. 2021) using the identifiers in the EPR Core Dataset (EPR). According to the source, this is the first dataset of this type with global coverage, and it enables the examination of the links between armed actors and their targeted civilian constituencies.

The next two datasets are provided by the Center for Systemic Peace. The Major Episodes of Political Violence (MEPV) dataset covers interstate, societal, and communal warfare for the period between 1946 and 2018 (Marshall et al. 2021). Ethnic violence and ethnic warfare were selected among the different indicators included. While the temporal coverage is wide, some have argued that they apply restrictive inclusion criteria and that there is no explanation of the definition of 'ethnic' used (Bartusevicius 2006).

The second is the PITF (Political Instability Task Force) – State Failure Problem Set, which includes annual data on different episodes of state failure between 1955 and 2018, in particular on cases of ethnic war and genocide/politicide (Marshall et al. 2021). Ethnic wars are distinct from other episodes of violent conflict in that challengers seek changes in their status, e.g. independence or regional autonomy. While there is a text description for each event, more detailed analysis of the actors requires further coding by the user.

With a regional focus, the Social Conflict Analysis Database (SCAD) covers different social disturbances, including protests, riots, and strikes in Africa, Latin America, and the Caribbean, and was last updated in 2017 (Salehyan and Hendrix 2017; Salehyan et al. 2012). It includes information on the different actors and targets involved, the type of event, number of deaths, and the source of tension/disorder. Data can be restricted to events in which the first issue mentioned was ethnic or religious discrimination or issues (with a distinction between ethnic and religious events).

The next three datasets provide indicators at the country-year level. A word of caution is included to highlight issues related to the clarity of the meaning of the indicators and the sources of data used. The Global State of Democracy Indices (GSoDI) capture democratic trends over the period 1975–2019 for 163 independent countries, based on the compilation of 116 indicators from different types of data (International IDEA 2021). One of the indicators used measures religious and ethnic tensions, and results from taking the highest score of these two composing elements, originally obtained from the International Country Risk Guide (ICRG) dataset. While this offers access to these elements, which are beyond a paywall in the original dataset, the original variables are not available, and the resulting indicator does not allow for a distinction between religious and ethnic tensions.

The Fund for Peace provides a measure of state fragility, the Fragile States Index, which results from the aggregation of twelve different indicators obtained from a conflict assessment framework (Fund for Peace 2021). These indicators provide a snapshot of a state in a given moment. Among them is group grievance, which captures divisions and schisms between different groups in society.

This variable was included here for completeness, but we warn against using it in the study of ethnic conflict as it is a fuzzy measure that encompasses different elements.⁹

The IPD (CEPII et al. 2021) captures different institutional characteristics of countries based on perceptions data, one of which is an indicator of conflicts of ethnic, religious, and regional nature. It is noted that scope of the indicator is broader, and the coverage is more limited, with data available only for five years and the last update from 2016.

Finally, I refer to data on different types of conflict from the Armed Conflict Location and Event Data Project (ACLED) dataset (Raleigh et al. 2010), which also distinguishes between the types of actors that participate. It has a global coverage and provides data based on a range of sources for the period 1997–2021. Some insights can be gained based on events involving political militias. While access to this data is free, it requires registration.

2.3 Datasets introduced in journal articles and data features

I now turn to recently proposed new datasets.¹⁰ The CDC provides data on armed conflict, distinguishing between ethnic and non-ethnic, and between governmental and territorial (Bartusevicius 2016). It is an alternative to the ACD2EPR dataset (included in GROW^{mp}), and the novelty comes from the coding of ethnic conflicts. Instead of focusing on the aims declared by the conflicting parties, it relies on the ethnic composition and the patterns of the confrontation, considering the features of contenders at the outset of conflict. Thus, it is more appropriate for the study of conflict onset.

The Targeted Mass Killing (TMK) dataset (Butcher et al. 2020) offers data on a type of violence that is less covered in the remaining selected datasets. It spans the period 1946–2017 and includes data on episodes of genocide, politicide, and other mass atrocities that deliberately target ethnic, religious, or political groups. The distinction between episodes targeting ethnic or religious groups is possible when using the data at the event-year level, which also allows for matching with the EPR database.

Given the specific focus of the remaining three datasets, they are more indirectly relevant to the analysis of ethnic conflict. Pishedda (2020) provides an indicator of non-violent campaigns with ethnic character. The new dataset of ethnic and ethnoreligious terrorism (GTD2EPR) proposed by Polo (2020) identifies the specific subnational ethnic communities from which terrorist organizations have originated and enables the study of mechanisms of terrorism diffusion. Finally, Svensson and Nilsson's (2018a, b) Religion and Armed Conflict (RELAC) dataset contains information on the religious dimensions of conflicts, namely whether parties involved belong to different religions or whether there is a religious dimension in the incompatibility. It focuses on armed conflict between the government and the rebels.¹¹

⁹ For a more detailed account of datasets on grievances, see Marquardt and Herrera (2015).

¹⁰ A word of caution that these datasets are recent and have not yet been widely used or reviewed in subsequent publications.

¹¹ See Svensson and Nilsson (2018a) for a reference to other publicly available datasets on religious conflicts and a description of the advantages of RELAC.

3 Descriptive analysis and comparisons

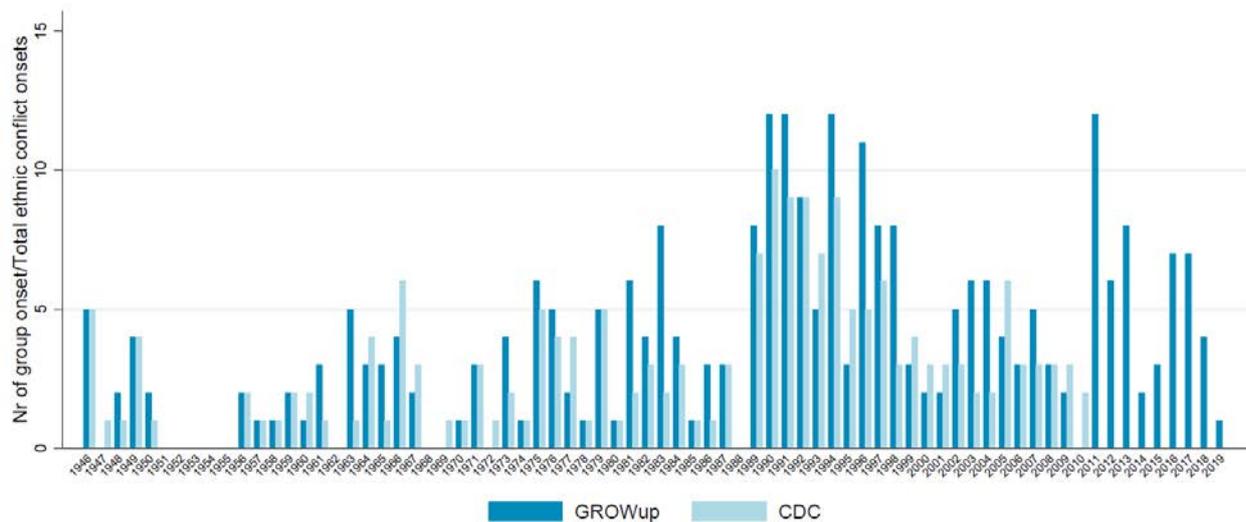
This section includes some descriptive analysis using different indicators from the selected datasets, including comparisons of the insights obtained using similar measures. Basic descriptive statistics of the variables used are included in Table A2 in the appendix.

3.1 Ethnic armed conflict

Figure 1 represents the total number of ethnic conflicts onsets by year, according to the indicators from the GROW^{up} (based on the ‘keep ongoing’ criteria) and CDC datasets. While the former is measured at the group level for the full period represented, the latter is at the event level and offers data until 2010. Overall, the figure suggests high levels of new ethnic conflicts during the 1990s and a decline since then, except a peak in 2011, with the highest number of conflict onsets during this year in the territory of Sudan.

With the exceptions of a few years (e.g., 1961, 1964, 1994, and 2005), the two indicators show similar trends. As described in detail in Bartusevicius (2016), the CDC dataset results from a different coding of ethnic. While conflict is coded as ethnic based on both ethnic claim and recruitment in the GROW^{up}, it is coded as ethnic based only on group composition (at the outset of conflict) in the CDC. In terms of the greatest discrepancies in the numbers of conflicts between these two datasets, these correspond to the years 1983, 1996, and 1998, followed by 1963, 1981, and 2003–2004.¹²

Figure 1: Total number of ethnic conflict onsets by year, 1946–2019



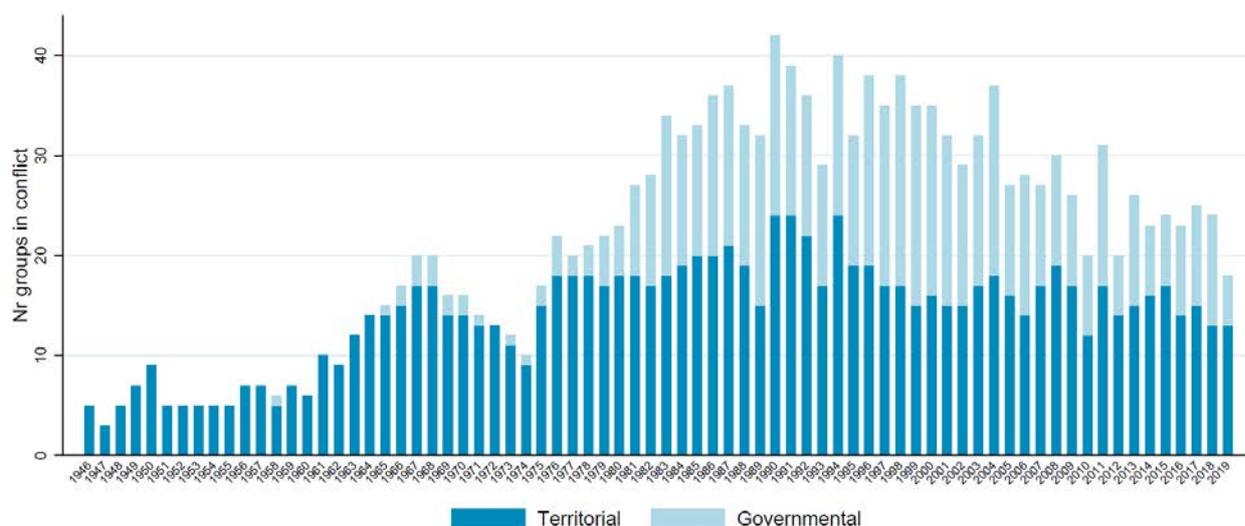
Source: author's elaboration based on data from Girardin et al. (2015) and Bartusevicius (2016).

Turning to conflict incidence (represented by the overall height of the bars in Figure 2), data from GROW^{up} suggests an increasing trend until the mid-1990s, when it reached the highest number of ongoing conflicts in the period considered. There has been a decline since then, with the total number below 20 in 2019. When looking at the split by source of incompatibility (the height of

¹² Given that data from GROW^{up} is at the event-year level, we do not have information on the differences in terms of group. Still, in terms of country-year differences, the biggest discrepancies are for Sudan in the years 1963, 1983, and 2003, Iraq in 2004, and Democratic Republic of the Congo in 1998.

each coloured bar), one notices that there are more ongoing ethnic conflicts resulting from territorial than governmental incompatibilities.

Figure 2: Total number of ongoing ethnic conflicts, divided by incompatibility (stacked bars), 1946–2019



Source: author's elaboration based on data from Girardin et al. (2015).

Using the indicators from the MEPV and IPD datasets for a rough comparison, Table 2 gives a snapshot of the magnitude of ethnic violence episodes in five different years. The MEPV ethnic war and violence variables are measured from 1 to 10, with lower values meaning lower magnitudes. The level of intensity of ethnic, religious, or regional internal conflicts from IPD ranges from 0 to 4, but in this case lower numbers represent more serious conflicts.

The first three columns represent the averages of the indicator for all episodes in all countries in that year, while the last three indicate the number of countries with high magnitude scores. The average values of the MEPV variables indicate low and similar magnitudes for both ethnic violence and ethnic war throughout the period, while the average values from IPD suggest medium levels of intensity. Turning to the number of episodes of high intensity, there is some discrepancy in the absolute values obtained from the two sources, but both suggest more conflicts of high magnitude in 2012 and 2016.

Table 2: Magnitude/intensity of episodes of ethnic conflict

Year	Average values			High intensity		
	Magnitude of ethnic violence (MEPV)	Magnitude of ethnic war (MEPV)	Intensity of conflict (IPD)	Magnitude of ethnic violence (MEPV)	Magnitude of ethnic war (MEPV)	Intensity of conflict (IPD)
2001	2.17	0.83	2.80	9	10	8
2006	2.00	0.50	2.80	10	10	13
2009	2.10	0.44	2.93	15	16	13
2012	1.88	0.50	2.17	16	16	65
2016	1.78	1.00	2.73	16	16	43

Note: high intensity means MEPV > 5 and IPD < 2.

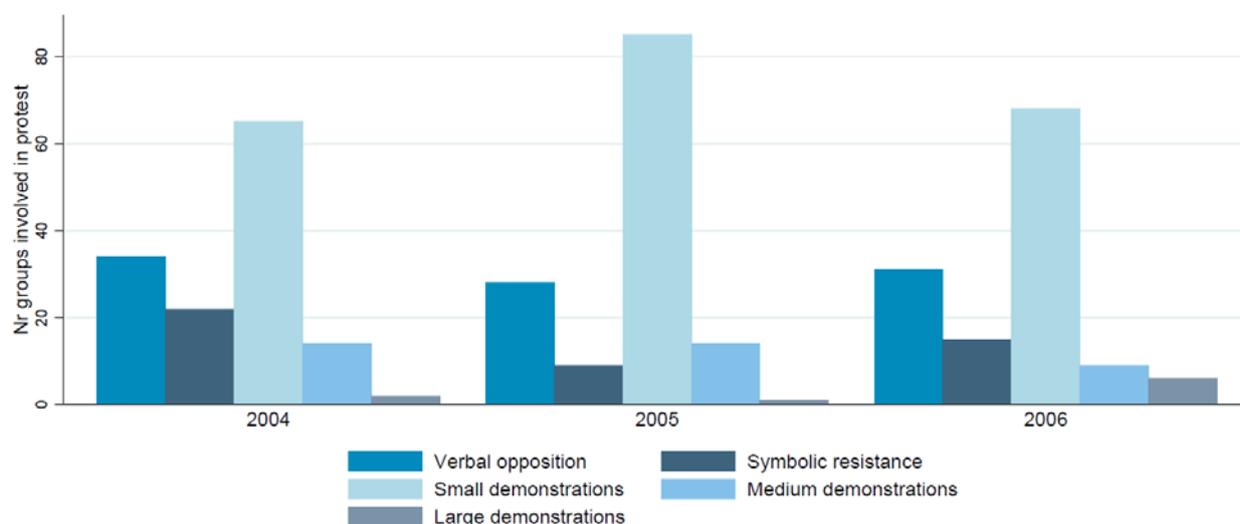
Source: author's elaboration based on data from Marshall et al. (2021) and CEPIL et al. (2021).

3.2 Ethnic tension and protests

Beginning with a note on ethnic tensions, as mentioned in the introduction, among the datasets selected, only the variable ethnic and religious tensions from GSoDI (which draws on the ICRG data) explicitly refers to this term. We compare this variable with the indicator on group grievances from the Fragile State Index (FSI), while referring to the limitations discussed in section 2.2. The Spearman correlation coefficient between the two variables is -0.642 , indicating some correlation between the two variables. To give an example, Syria, Israel, Sudan, Turkey, and the Democratic Republic of the Congo were the five countries with highest levels of group grievance in 2019, and in the same year, the last three countries also featured in the list of countries with worst scores of ethnic/religious tensions. The negative sign stems from the fact that higher values of the group grievances correspond to worst outcomes, whereas the opposite applies to ethnic and religious tensions.

Turning to data on protests, Figure 3 shows the number of groups involved in different types of protest, using the MAR variable at the group-year level, in the three years with available data. The largest number corresponds to small demonstrations followed by verbal opposition.

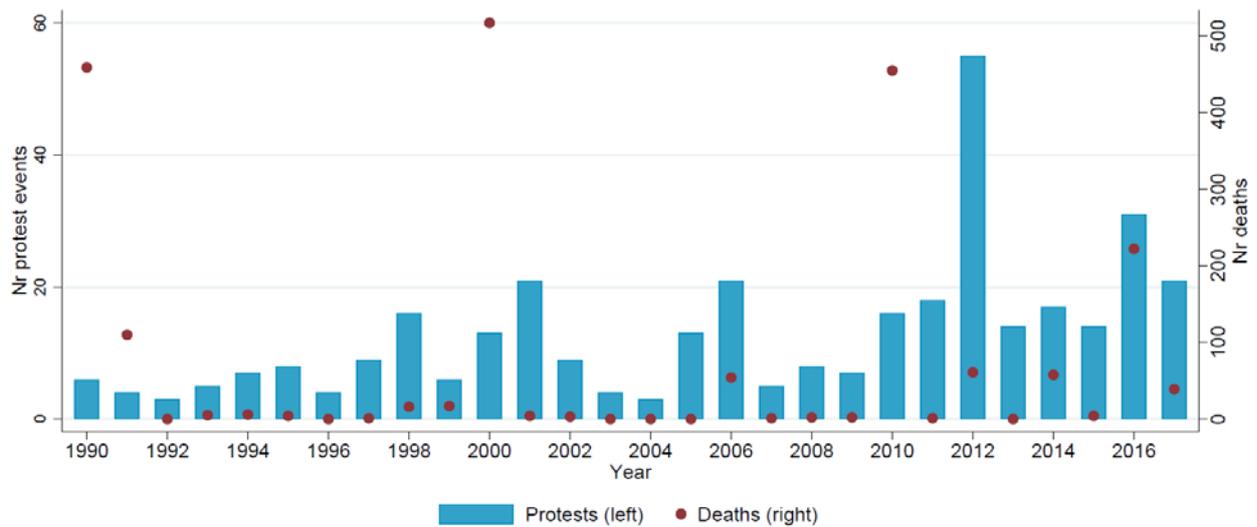
Figure 3: Number of groups involved in protest, by protest type, 2004–06



Source: author's elaboration based on data from CIDCM (2016).

Variables from SCAD provide data on the number of protests and estimated number of deaths covering a longer period, but I remind here that, geographically, the data covers only Africa, Mexico, Central America, and the Caribbean. Figure 4 shows that, with a few exceptions, the number of protests (left axis) was below 20 between 1990 and 2017. The most notable exception is in 2012, with almost 60 events, in part explained by the Arab Spring protests in Tunisia and Egypt. In terms of estimated number of deaths, the years 1990, 2000, 2010, and 2016 all show total numbers greater than 200. For comparison, data on protests involving identity militias from ACLED was also considered, but the number of years covered was much smaller, as was the number of events.

Figure 4: Number of protest events and best estimate of death numbers, 1990–2017



Source: author's elaboration based on data from Salehyan et al. (2012).

3.3 Final notes

The insights from the previous section together with this brief descriptive analysis suggest there is a range of different sources providing data on ethnic conflict, tensions, protests, and other violent and non-violent events. The GROW^{up} dataset offers comprehensive data on ethnic armed conflict, which can also be linked to other UCDP datasets on ethnicity and conflict. While, once updated, the MAR and accompanying datasets will also provide useful data, data on other types of violent and non-violent events, such as ethnic tensions and protests, is currently scattered. Moreover, the coverage of the datasets varies, and the information provided is not always consistent across sources.

References

- ACLED (2019a). 'Armed Conflict Location & Event Data Project (ACLED) Codebook, 2019'. Madison: Armed Conflict Location and Event Data Project. Available at: https://acleddata.com/acleddatanew/wp-content/uploads/2021/11/ACLED_Codebook_v1_January-2021.pdf (accessed 2 December 2021).
- ACLED (2019b). 'Armed Conflict Location & Event Data Project (ACLED) Methodology'. Madison: ACLED. Available at: https://acleddata.com/acleddatanew/wp-content/uploads/dlm_uploads/2019/04/Methodology-Overview_FINAL.pdf (accessed 2 December 2021).
- Bartusevicius, H. (2016). 'Introducing the Categorically Disaggregated Conflict (CDC) Dataset'. *Conflict Management and Peace Science*, 33(1): 89–110. <https://doi.org/10.1177/0738894215570423>
- Bertho, F. (2012). 'Presentation of the Institutional Profiles Database 2012 (IPD 2012)'. Les Cahiers de la DG Trésor 2013-03. Paris: Direction Générale du Trésor. Available at: <https://www.tresor.economie.gouv.fr/Articles/ba9e35d1-f930-4443-8284-66d321e77554/files/a0252cf0-ac76-4301-aa53-4579cfb733d5> (accessed 20 December 2021).
- Bormann, N.-C., L. Girardin, P. Hunziker, and M. Vogt (2021). 'GROW^{up} Research Front-End Documentation, RFE Release 2021'. Zurich: ETH Zurich. Available under 'Documentation' at: <https://growup.ethz.ch/rfe> (accessed 2 December 2021).

- Butcher, C., B.E. Goldsmith, and S. Nanlohy (2019). ‘Targeted Mass Killing Dataset Codebook’. Canberra: Australian National University.
- Butcher, C., B.E. Goldsmith, S. Nanlohy, A. Sowmya, and D. Muchlinksy (2020). ‘Introducing the Targeted Mass Killing Data Set for the Study and Forecasting of Mass Atrocities’. *Journal of Conflict Resolution*, 64(7-8): 1524–47. <https://doi.org/10.1177/0022002719896405>
- Cederman, L.-E., A. Wimmer, and B. Min (2010). ‘Why Do Ethnic Groups Rebel? New Data and Analysis’. *World Politics*, 62(1): 87–119. <https://doi.org/10.1017/S0043887109990219>
- CEPII, AFD, and MEF (2021). *Institutional Profiles Database* [Online]. Centre for Prospective Studies and International Information (CEPII), Agence française de développement (AFD), and Ministère de l’Économie et des Finances (MEF). Available at: <http://www.cepii.fr/institutions/EN/ipd.asp> (accessed 2 December 2021).
- CIDCM (2009). ‘Minorities at Risk (Mar) Codebook Version 2/2009’. College Park: Center for International Development and Conflict Management (CIDCM). Available at: http://www.mar.umd.edu/data/mar_codebook_Feb09.pdf (accessed 2 December 2021).
- CIDCM (2016). *Minorities at Risk Project* [Online]. College Park: CIDCM. Available: <http://www.mar.umd.edu/> (accessed 2 December 2021).
- Fearon, J.D., and D.D. Laitin (2003). ‘Ethnicity, Insurgency, and Civil War’. *The American Political Science Review*, 97(1): 75–90. <https://doi.org/10.1017/S0003055403000534>
- Fjelde, H., L. Hultman, L. Schubiger, L.-E. Cederman, S. Hug, and M. Sollenberg (2021). ‘Introducing the Ethnic One-Sided Violence Dataset’. *Conflict Management and Peace Science*, 38(1): 109–26. <https://doi.org/10.1177/0738894219863256>
- Forsberg, E., A. Duursma, and L. Grant (2012). ‘Theoretical and Empirical Considerations in the Study of Ethnicity and Conflict’. UCDP Paper 8. Uppsala: Uppsala University. Available at: <http://uu.diva-portal.org/smash/get/diva2:579300/FULLTEXT01.pdf> (accessed 2 December 2021).
- Fund for Peace (2017). ‘Fragile States Index and Cast Framework, Methodology’. Washington, DC: Fund for Peace. Available at: <https://fragilestatesindex.org/methodology/> (accessed 2 December 2021).
- Fund for Peace (2021). *Fragile States Index* [Online]. Washington, DC: Fund for Peace. Available at: <https://fragilestatesindex.org/> (accessed 3 December 2021).
- Gilley, B. (2004). ‘Against the Concept of Ethnic Conflict’. *Third World Quarterly*, 25(6): 1155–66. <https://doi.org/10.1080/0143659042000256959>
- Girardin, L., P. Hunziker, L.-E. Cederman, N.-C. Bormann, S. Rügger, and M. Vogt (2015). *GROW^{up} - Geographical Research on War, Unified Platform* [Online]. Zurich: ETH Zurich. Available at: <http://growup.ethz.ch/> (accessed 2 December 2021).
- Gurr, T.R. (2017). ‘Observations on the Study of Ethnic Conflict’. *Ethnopolitics*, 16(1): 34–40. <https://doi.org/10.1080/17449057.2016.1235345>
- Horowitz, S. (2008). ‘Mapping Pathways of Ethnic Conflict Onset: Preferences and Enabling Conditions’. *Ethnopolitics*, 7(2-3): 307–20. <https://doi.org/10.1080/17449050802243483>
- Hug, S. (2013). ‘The Use and Misuse of the "Minorities at Risk" Project’. *Annual Review of Political Science*, 16: 191–208. <https://doi.org/10.1146/annurev-polisci-060512-142956>
- International IDEA (2021). *The Global State of Democracy Indices* [Online]. Stockholm: International Institute for Democracy and Electoral Assistance (IDEA). Available at: <https://www.idea.int/gsod-indices/democracy-indices> (accessed 3 December 2021).
- Keefer, P., and S. Knack (2002). ‘Polarization, Politics and Property Rights: Links between Inequality and Growth’. *Public Choice*, 111: 127–54. <https://doi.org/10.1023/A:1015168000336>

- Marquardt, K.L., and Y.M. Herrera (2015). 'Ethnicity as a Variable: An Assessment of Measures and Data Sets of Ethnicity and Related Identities'. *Social Science Quarterly*, 96(3): 689–716. <https://doi.org/10.1111/ssqu.12187>
- Marshall, M.G. (2019). 'Codebook: Major Episodes of Political Violence (MEPV) and Conflict Regions, 1946-2018'. Vienna, VA: Center for Systemic Peace. Available at: <http://www.systemicpeace.org/inscr/MEPVcodebook2018.pdf> (accessed 3 December 2021).
- Marshall, M.G., B.R. Cole, D.R. Marshall, E. Elzinga, and G. Elzinga-Marshall (2021). *INSCR Data Page* [Online]. Vienna, VA: Center for Systemic Peace. Available at: <https://www.systemicpeace.org/inscrdata.html> (accessed 3 December 2021).
- Marshall, M.G., T.R. Gurr, and B. Harff (2019). 'PITF - State Failure Problem Set: Internal Wars and Failures of Governance, 1955-2018, Dataset and Coding Guidelines'. Vienna, VA: Center for Systemic Peace. Available at: <http://www.systemicpeace.org/inscr/PITFProbSetCodebook2018.pdf> (accessed 3 December 2021).
- Pettersson, T. (2021). 'UCDP Non-State Conflict Codebook Version 21.1'. Uppsala: Uppsala University. Available at: <https://ucdp.uu.se/downloads/> (accessed 3 December 2021).
- Pettersson, T., S. Davis, A. Deniz, G. Engström, N. Hawach, S. Höglbladh, M. Sollenberg, and M. Öberg (2021). 'Organized Violence 1989-2020, with a Special Emphasis on Syria'. *Journal of Peace Research*, 58(4): 809–25. <https://doi.org/10.1177/00223433211026126>
- Pischedda, C. (2020). 'Ethnic Conflict and the Limits of Nonviolent Resistance'. *Security Studies*, 29(2): 362–91. <https://doi.org/10.1080/09636412.2020.1722854>
- Polo, S.M.T. (2020). 'How Terrorism Spreads: Emulation and the Diffusion of Ethnic and Ethnoreligious Terrorism'. *Journal of Conflict Resolution*, 64(10): 1916–42. <https://doi.org/10.1177/0022002720930811>
- Raleigh, C., A. Link, H. Hegre, and J. Karlsen (2010). 'Introducing ACLED-Armed Conflict Location and Event Data'. *Journal of Peace Research*, 47(5): 651–60. <https://doi.org/10.1177/0022343310378914>
- Salehyan, I., C.S. Hendrik, J. Hamner, C. Case, C. Linebarger, E. Stull, and J. Williams (2012). 'Social Conflict in Africa: A New Database'. *International Interactions*, 38(4): 503–11. <https://doi.org/10.1080/03050629.2012.697426>
- Salehyan, I., and C. Hendrix (2017). 'Social Conflict Analysis Database (SCAD), Version 3.3, Codebook and Coding Procedures'. Austin: Robert Strauss Center for International Security and Law. Available at: https://www.strausscenter.org/wp-content/uploads/SCAD_33_Codebook.pdf (accessed 3 December 2021)
- Sambanis, N. (2001). 'Do Ethnic and Nonethnic Civil Wars Have the Same Causes?: A Theoretical and Empirical Inquiry (Part 1)'. *Journal of Conflict Resolution*, 45(3): 259–82. <https://doi.org/10.1177/0022002701045003001>
- Schubiger, L.I., and H. Fjelde (2019). 'The Ethnic One-Sided Violence Dataset Codebook'. Uppsala: Uppsala University. Available at: https://ucdp.uu.se/downloads/eosv/EOSV_Codebook.pdf (accessed 2 December 2021).
- Sundberg, R., K. Eck, and J. Kreutz (2012). 'Introducing the UCDP Non-State Conflict Dataset'. *Journal of Peace Research*, 49(2): 351–62. <https://doi.org/10.1177/0022343311431598>
- Svensson, I., and D. Nilsson (2018a). 'Disputes over the Divine: Introducing the Religion and Armed Conflict (RELAC) Data, 1975 to 2015'. *Journal of Conflict Resolution*, 62(5): 1127–48. <https://doi.org/10.1177/0022002717737057>
- Svensson, I., and D. Nilsson (2018b). 'Religion and Armed Conflict (RELAC) Data, Codebook, V.1.1'. Uppsala: Uppsala University. Available at: <https://files.webb.uu.se/uploader/1576/Codebook-RELAC.pdf> (accessed 3 December 2021)

Tufis, C.D. (2020a). 'The Global State of Democracy Indices Codebook, Version 4'. Stockholm: International Institute for Democracy and Electoral Assistance (IDEA).
<https://doi.org/10.31752/idea.2020.71>

Tufis, C.D. (2020b). 'The Global State of Democracy Indices Technical Procedures Guide, Version 4'. Stockholm: International Institute for Democracy and Electoral Assistance (IDEA). Available at: <https://www.idea.int/gsod-indices/sites/default/files/global-state-of-democracy-indices-technical-procedures-guide-v4.pdf> (accessed 3 December 2021).

Appendix

Table A1: Coverage and definitions used in the selected datasets

Dataset	Short description	Main definitions	Overall coverage: time (space)	Data sources and coding procedure
Armed Conflict Location & Event Data Project (ACLED)	Codes reported information on political violence events, demonstrations and select politically relevant non-violent events.	<p><i>Political violence</i>: ‘use of force by a group with a political purpose or motivation. (...) some non-violent instances – including protests and strategic developments – are included in the dataset to capture the potential precursors or critical junctures of a violent conflict.’</p> <p>Given the scope of this note, data can be restricted to groups with ‘Inter Code 4: Identity Militias’: ‘armed and violent groups organized around a collective, common feature including community, ethnicity, region, religion or, in exceptional cases, livelihood. Therefore, for ACLED’s purposes, identity militias include those reported as “tribal”, “communal”, “ethnic”, “local”, “clan”, and “religious” and “caste” militias. Events involving “identity militias” are often referred to as “communal violence” as these violent groups often act locally, in the pursuance of local goals, resources, power, security, and retribution.’</p>	1997–2021 (Global)	Information collected primarily from secondary source information including from news reports. Individual researchers scrutinize the information from reports; they are then aggregated and revised by the first coding reviewer, investigated and cross-checked by the second reviewer, and then event notes and details are inspected by the third and final reviewer.
Categorically Disaggregated Conflict (CDC)	Dataset categorizing conflict into ethnic vs. non-ethnic and governmental vs. territorial.	<p><i>Ethnic groups</i>: those ‘composed of people who share a distinctive and enduring collective identity based on shared experiences and cultural traits’ who ‘may define themselves, and be defined by others, in terms of any or all of the following traits: life ways, religious beliefs, language, physical appearance, region of residence, traditional occupations, and a history of conquest and repression by culturally different peoples’.</p> <p><i>Conflict</i>: ‘contested incompatibility that concerns government and/or territory where the use of force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths’ (UCDP); restricted to non-international conflicts.</p> <p><i>Ethnic conflict</i>: ‘conflict between two or more groups whose majorities represent different ethnicities’.</p>	1946–2010 (Global)	The UCDP/PRIOD dataset is used as a base; these conflicts are classified as ethnic or non-ethnic based on the author’s own reading of primary and secondary sources.

Dataset	Short description	Main definitions	Overall coverage: time (space)	Data sources and coding procedure
Ethnic conflict (Pischedda 2000)	Indicator of non-violent campaigns with ethnic character.	<i>Ethnic conflict</i> : coding of conflict events based on the identity of rebel groups' members. An event is coded as ethnic conflict if the rebel organization recruits from an ethnic group other than the state-controlling group and does not recruit from the state-controlling ethnic group; and also in the case of ethnic-conflict events of violent resistance against foreign occupation and against colonial rule.	1945–2006 (Global)	Based on NAVCO 2.0, using the Ethnic Power Relations (EPR) dataset to code the state-controlling ethnic group. A variety of case-specific sources (e.g., newspapers, scholarly works, and reports) and general sources on non-violent resistance and ethnic politics are used.
Ethnic One-Sided Violence (EOSV)	Information on the ethnic identity of civilian victims of direct and deliberate killings by state and non-state actors.	<i>Politically relevant ethnic groups</i> : 'ethnic groups [groups with a shared belief in a common descent and culture] whose interests are claimed to be represented by at least one political actor (for example, a political party) in the national political arena, or whose members are systematically excluded in certain domains of public politics.' <i>Ethnic victim</i> : 'reflects whether victims could be identified as belonging to a particular group'.	1989–2013 (Global)	Uses the documentation underlying the original coding of the One-Sided Violence (OSV) dataset by re-reading all individual media reports, non-governmental organization reports, and other case-specific documentation that support the original coding of the event, as well as other relevant international and country-specific sources.
Fragile States Index (FSI)	Overall index of fragile states and underlying indicators.	Group grievance: 'The Group Grievance Indicator focuses on divisions and schisms between different groups in society – particularly divisions based on social or political characteristics – and their role in access to services or resources, and inclusion in the political process. (...)'	2006–21 (Global)	Based on CAST's comprehensive social science approach. Triangulation of data from pre-existing quantitative datasets, content analysis, and qualitative expert analysis; and critical review to obtain final scores for the Index.
GROW ^{up} – Geographical Research on War, Unified Platform	Integrated data platform providing information on ethnic groups. Based on other component datasets: EPR Code Dataset, GeoEPR, ACD2EPR, EPR-TEK, EPR-ER, and EPR-ED.	<i>Politically relevant ethnic group</i> : an ethnic group is defined as a 'group of individuals sharing a common ethnicity' (i.e. 'any subjectively experienced sense of commonality based on the belief in common ancestry and shared culture'); it is politically relevant 'if at least one political organization claims to represent it in national politics or if its members are subjected to state-led political discrimination'. <i>Conflict</i> : UCDP definition (see row 'Categorically Disaggregated Conflict') <i>Ethnic conflict</i> : UCDP rebel organizations and EPR ethnic groups are linked based on both ethnic claims ('explicit or implicit ethnic claim by the rebel organization') and significant ethnic recruitment.	1946–2021 (Global)	Relevant variables obtained using three steps: (i) linking UCDP rebel organizations to EPR ethnic groups, based on both ethnic claims and significant ethnic recruitment (ACD2EPR); (ii) link rebel organizations to Internal and Internationalized Armed Conflicts (based on UCDP Actor Dataset); (iii) link each rebel-ethnic group dyad to information on intra-state conflicts (using the ACD Dataset). Sources and coding specific to each dataset.

Dataset	Short description	Main definitions	Overall coverage: time (space)	Data sources and coding procedure
GTD2EPR (Polo 2020)	Group-level dataset of ethnic and ethno-religious terrorism.	<p><i>Ethnic group</i>: EPR definition; see definition in row 'GROW'⁴⁰.</p> <p><i>Terrorism</i>: 'the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation.'</p> <p><i>Ethnic terrorism</i>: terrorist attacks perpetrated by organizations linked to an ethnic group; only domestic terrorist attacks are considered.</p>	1970–2009 (Global)	The author used a number of sources to code ethnic claims of GTD organizations and keyword searches through Lexis-Nexis. To construct the dataset, the author aggregated domestic terrorist attacks by organizations linked to a specific EPR group.
Institutional Profiles Database (IPD)	Measure of countries' institutional characteristics.	<p><i>Internal conflicts: ethnic, religious, regional</i>: answer to the question 'Intensity of internal conflicts of an ethnic, religious or regional nature' on a scale from 0 ('serious conflicts') to 4 ('no conflict').</p>	2001, 2006, 2009, 2012, 2016 (Global)	Based on the IDP questionnaire, completed by the Economic Services and French Development Agency's (AFD) offices using their own knowledge and also calling upon local expertise.
Minorities at Risk (MAR)	Monitors and analyses the status and conflicts of politically-active communal groups.	<p><i>Minority at risk</i>: ethno-political group that fulfils the following criteria: (i) 'Membership in the group is determined primarily by descent by both members and non-members. (...)' (ii) 'Membership in the group is recognized and viewed as important by members and/or non-members. The importance may be psychological, normative, and/or strategic.' (iii) 'Members share some distinguishing cultural features, such as common language religion and customs.' (iv) 'One or more of these cultural features are practiced by a majority of the members of the group.' (v) 'The group has at least 100,000 members or constitutes one percent of a country's population.'; (vi) 'For groups dropping below population thresholds after being included in the dataset, coding will continue for up to 10 years to determine if population rebounds. (...)'</p> <p><i>Conflict</i>: includes different indicators of group conflict behaviour.</p>	2004–06 (Global)	Coding is made primarily by trained graduate and undergraduate students using open-source information. All coding is then reviewed by senior editors and by the research director before public release. As much as is possible, coders rely on multiple sources for each code assigned.

Dataset	Short description	Main definitions	Overall coverage: time (space)	Data sources and coding procedure
Major Episodes of Political Violence (MEPV)	Dataset on magnitude of different types of political violence.	<p><i>Major episodes of political violence</i>: 'systematic and sustained use of lethal violence by organized groups that result in at least 500 directly-related deaths over the course of the episode'.</p> <p><i>Ethnic</i>: 'intrastate involving the state agent and a distinct ethnic group'.</p> <p><i>Violence and war</i>: 'use of instrumental violence without necessarily exclusive goals' and 'violence between distinct, exclusive groups with the intent to impose a unilateral result to the contention', respectively.</p>	1946–2018 (Global)	Each episode was researched, analyzed, and coded by the author.
PITF State Failure Problem Set (PITF)	Includes four distinct types of state failure.	<p><i>Ethnic wars</i>: 'episodes of violent conflict between governments and national, ethnic, religious, or other communal minorities (ethnic challengers) in which the challengers seek major changes in their status. (...)'</p> <p>Includes two minimum thresholds: 'each party must mobilize 1000 or more people'; 'there must be at least 1000 direct conflict-related deaths over the full course of the armed conflict and at least one year when the annual conflict-related death toll exceeds 100 fatalities'.</p> <p><i>Genocide/Politicide events</i>: 'involve the promotion, execution, and/or implied consent of sustained policies by governing elites or their agents -- or in the case of civil war, either of the contending authorities -- that result in the deaths of a substantial portion of a communal group or politicized non-communal group. In genocides the victimized groups are defined primarily in terms of their communal (ethnolinguistic, religious) characteristics. In politicides, by contrast, groups are defined primarily in terms of their political opposition to the regime and dominant groups.'</p>	1955–2018 (167)	Cases and coding based on information compiled from multiple sources; discrepancies in the historical records scrutinized and reconciled by analysts.
Religion and Armed Conflict (RELAC)	Dataset suitable for analyzing the causes, dynamics, and resolution of religious conflicts.	<p><i>Armed conflict</i>: UCDP definition (see row 'Categorically Disaggregated Conflict')</p> <p><i>Religious conflict</i>: 'codes [as religious] both whether the armed conflict is (1) fought over religious issues and (2) whether the main belligerents come from different religious identities.'</p>	1975–2015 (Global)	Rely primarily on data from the UCDP Encyclopaedia. UCDP collects the data first from the online database Factiva and from country-specific sources; then region experts help in resolving uncertainties; finally, coding is reviewed by project managers.

Dataset	Short description	Main definitions	Overall coverage: time (space)	Data sources and coding procedure
Social Conflict Analysis Database (SCAD)	Information on different social disturbances in Africa, Latin America, and the Caribbean.	The codebook includes short definitions of different types of violence. For the purpose of this note, the list of events can be restricted to tensions/disorders whose source stems from 'ethnic discrimination, religious issues' or 'religious discrimination, religious issues'. The codebook includes no further details about how these are coded.	1990–2017 (Africa, Mexico, Central America, and the Caribbean)	The primary source of information for this dataset comes from searches of Associated Press and Agence France Presse newswires, as compiled by the Lexis-Nexis news service. Searches and coding undertaken by the research staff.
Targeted Mass Killing (TMK)	Dataset on episodes of targeted mass killing, including genocide and politicide.	<i>Targeted mass killing</i> : 'is the direct killing of noncombatant members of a group by an organized armed force or collective with the intent of destroying the group, or intimidating the group by creating a perception of imminent threat to its survival. A targeted group is defined in terms of political and/or ethnic and/or religious identity.'	1946–2017 (Global)	Use different existing datasets. Collected a broad pool of potential cases, relying on existing datasets and historical research based on consultation of literature, historical accounts, and area and genocide studies experts, before applying the coding criteria.
The Global State of Democracy Indices (GSoDI)	Depicts democratic trends across a broad range of different attributes of democracy.	<i>Religious tension</i> : 'attributable to religious division, domination or suppression'. <i>Ethnic tension</i> : 'attributable to racial, nationality or language divisions'.	1975–2019 (Global)	Used four different types of sources: expert surveys, standards-based in-house coding, observational data, and composite measures. GSoDI produced by a team, including two external experts with the supervision of leading experts in democracy measurement.
UCDP Non-state conflict (NSC)	Database on non-state conflict.	<i>Non-state conflict</i> : 'the use of armed force between two organized armed groups, neither of which is the government of a state, which results in at least 25 battle-related deaths in a year.' Given the scope of this note, data can be restricted to conflict between groups with 'organization level 3': 'Groups that share a common identification along ethnic, clan, religious, national or tribal lines. These are not groups that are permanently organized for combat, but who at times organize themselves along said lines to engage in fighting. This level of organization captures aspects of what is commonly referred to as "communal conflicts", in that conflict stands along lines of communal identity.'	1989–2020 (Global)	Results from an automatic filtering and aggregation of the UCDP Georeferenced Event Dataset from incident/event level to the conflict/dyad-year level; and information gathering and coding of a number of extra variables at the aggregate conflict or actor level. The original reporting underlying the dataset is collected from different news sources and secondary sources, such as local media, NGO and IGO reports, field reports, and books.

Source: ACLED: ACLED (2019a, b). CDC: Bartusevicius (2016). Pischedda: Pischedda (2020). EOSV: Fjelde et al. (2021); Schubiger and Fjelde (2019). FSI: Fund for Peace (2017). GROW^{UP}: Girardin et al. (2015); Bormann et al. (2021). GTD2EPR: Polo (2020). IPD: Bertho (2012); CEPII et al. (2021). MAR: CIDCM (2009). MEPV: Marshall (2019); Marshall et al. (2021). PITF: Marshall et al. (2019); Marshall et al. (2021). RELAC: Svensson and Nilsson (2018a, b). SCAD: Salehyan and Hendrix (2017). TMK: Butcher et al. (2019); Butcher et al. (2020). GSoDI: Tufis (2020a, b); International IDEA (2021). NSC: Sundberg et al. (2012); Pettersson et al. (2021); Pettersson (2021).

Table A2: Descriptive statistics of the variables used in the analysis

Dataset – Indicator	Variable	Obs.	Mean	Std. dev.	Min	Max
<i>Country-year level</i>						
MEPV – Ethnic violence	mepv_ethviol	9938	.074	.413	0	6
MEPV – Ethnic war	mepv_ethwar	9939	.25	1.007	0	10
FFP – Group grievance	ffp_c3groupgrievance	2811	5.992	2.074	.5	10
GSoDI – Religious tensions/Ethnic tensions	gsodi_tensions	4723	.606	.226	0	1
IPD – Internal conflicts	ipd_conflict	541	2.416	1.282	0	4
<i>Group-year level</i>						
GROW ^{up} – Conflict onset (KO)	gu_onset_ko_flag	54496	0.005	0.071	0	1
GROW ^{up} – Ongoing territorial group-level conflict	gu_incidence_terr_flag	54496	0.019	0.137	0	1
GROW ^{up} – Ongoing governmental group-level conflict	gu_incidence_gov_flag	54496	0.01	0.1	0	1
MAR – Protest	mar_prot	845	1.221	1.459	0	5
<i>Event-year (dyad) level</i>						
CDC – Category of conflict	cdc_category	331	2.317	0.834	1	4
<i>Event-year level</i>						
PISC – Nonviolent campaign of ethnic character	pisc_ethnic_conflict	303	0.459	0.499	0	1
SCAD – Type of event	scad_etype	3607	7.063	2.514	1	10
SCAD – Number of deaths	scad_nddeath	3314	24.116	140.543	0	5000

Note: variable names are formed by the prefix corresponding to the name of the data source followed by original name of the variable.

Source: author's elaboration based on data from Marshall (2019); Marshall et al. (2021); Fund for Peace (2021); International IDEA (2021); CEPIL et al. (2021); Girardin et al. (2015); Bormann et al. (2021); CIDCM (2009, 2016); Bartusevicius (2016); Pischedda (2020); Salehyan and Hendrix (2017); and Salehyan et al. (2012).