Introducing xSub:
A New Portal for Cross-National Data on Sub-National Violence

Yuri M. Zhukov∗  Christian Davenport†  Nadiya Kostyuk‡

August 24, 2017§
(DRAFT)

Abstract
Researchers today have access to an unprecedented amount of geo-referenced, disaggregated data on political conflict and violence. Unfortunately, these new data sources lack a consistent event typology and use disparate units of analysis. As a result, findings are rarely comparable across studies, and we are unable to answer basic questions like “what does conflict A tell us about conflict B”? With this in mind, we introduce xSub – a unified framework and repository for micro-level event data on armed conflict and political violence. The goal of xSub is to reduce the barriers to comparative sub-national research, and empower researchers to quickly construct custom, analysis-ready datasets. Currently, xSub features 337 subnational datasets on political violence and protests in 139 countries, from 21 sources, including both large data collections and data from individual scholars. To facilitate comparisons across countries and sources, xSub organizes these data into consistent event categories and actors, and aggregates them into common units of analysis, by space (country, province, district, PRIO grid cell, electoral constituency) and time (year, month, week). This article lays outs the logic of the project and illustrates an example of its use, by investigating the impact of repression on dissent across 149 subnational datasets.

Keywords: sub-national conflict, event data, geographical disaggregation, violence, protest, repression, contention, disaggregated conflict, micro-foundations

∗Assistant Professor of Political Science, University of Michigan
†Professor of Political Science, University of Michigan
‡Ph.D. Candidate in Political Science and Public Policy, University of Michigan
§We thank Yioryos Nardis and Andrew Versalle for excellent programming and technical support.
Over the last two decades, social scientists have produced a tremendous amount of data on political conflict and violence.\textsuperscript{1} Large-scale data projects (e.g. Schrodt, Davis and Weddle, 1994, Gerner et al., 2002, Raleigh et al., 2010, Salehyan et al., 2012, Sundberg and Melander, 2013) – and more specialized studies of individual countries (e.g. Davenport and Ball, 2002, Davenport and Stam, 2006, Lyall, 2010, Verpoorten, 2012, Price, Gohdes and Ball, 2014, Shapiro and Weidmann, 2015, Zhukov, 2016\textsuperscript{a}) – have produced vast quantities of georeferenced, disaggregated data, collected from press reports, social media and government archives, using both manual and automated techniques. These data have fueled a new wave of sub-national conflict research, employing novel research designs at highly-granular levels of analysis. Yet they have also deepened five inter-related shortcomings in the literature: 1) most studies that use sub-national data nevertheless conduct their analyses at a highly-aggregated, macro level; 2) most micro-level studies focus on a single or limited set of countries; 3) cross-dataset comparisons are rare; 4) operational definitions of violence vary; and 5) there have been no consistent units of analysis, which might otherwise enable direct comparisons of results. As a result of this situation, sub-national conflict research has produced a series of contradictory findings, impeding the accumulation of knowledge, and leaving unanswered basic questions about the causes, dynamics and consequences of conflict and violence.

xSub – pronounced “cross-sub” – is a database of databases for cross-national research on sub-national violence. To facilitate comparisons across conflicts and studies, xSub features 337 datasets on violence and protests from 139 countries and 21 data sources, organized into consistent categories and units of analysis, by space (country, province, district, grid cell, electoral constituency) and time (year, month, week), by actors (government, rebel, civilians, others) and actions (any violence, indiscriminate force, selective force, or protests). In addition to conflict, xSub includes local data on demographics, geography, ethnicity, weather and other popular covariates. These data are available at x-sub.org.

Scholars can use xSub to examine a wide range of substantive areas. These include the relationships between climate variability and conflict, terrain and insurgency, ethnicity and communal fighting, elections and protests, as well as studies of spillover effects, conflict duration, recurrence and – as we demonstrated below – repression and dissent. In addition to facilitating meta-analyses across hundreds of conflicts, xSub is equally well-

\textsuperscript{1}We define political conflict as a dispute between two or more political actors (e.g. governments, challengers, third parties) over the pursuit, maintenance or distribution of power. We define political violence as the use of physical force to resolve an underlying political conflict.
suited for the study of micro-dynamics of political unrest within individual countries.

The article proceeds as follows. Section 1 surveys the current sub-national conflict data landscape, and introduces the rationale for a “database of databases” like xSub. Section 2 explains our data collection procedures, definitions and operationalizations, and the mechanisms we used to ensure data quality and consistency. Section 3 provides examples of how scholars can use these data, through illustrative meta-analysis of repression and protest across 149 subnational datasets. The conclusion outlines future development and our vision for widespread interest and participation in xSub.

Why xSub?

Traditionally, research on civil conflict (i.e., genocide, revolution, counter-revolution, insurgency, counter-insurgency, terrorism, counter-terrorism, protest, protest policing, negative sanction/civil liberties restriction and human rights violation) has maintained a cross-national focus, tracking macro-level trends across states. This work has provided incredible insights about why conflicts begin, become increasingly lethal, employ particular tactics, switch tactics, endure over time, terminate and reoccur in some countries as opposed to others.

Despite developing highly refined theories, measurement and estimation strategies, cross-national scholarship has come under increasing scrutiny. One criticism maintains that, given the coarse nature of the measures employed in cross-national research, macro-level work has shed little light on exactly what is taking place on the ground. A different criticism – and one that we pursue in the current project – is that, while prior research has focused on explaining variation between nation-states, it has generally overlooked variation within them.

To address this criticism, a growing movement in sub-national research has disaggregated actors and actions to a more granular local level, with shorter temporal units. This new wave of micro-level research has also led to incredible insights into the determinants of political conflict and violence, advancing our understanding of when and where conflict is likely to begin, escalate, and vary in terms of type, severity and persistence.

At the same time, sub-national research has met its own limitations. To date, most studies have largely focused on variation within a single country or conflict (e.g., genocide in Rwanda, counter-insurgency in Guatemala, civil war in Afghanistan, protest policing in the United States). This has proven to be problematic for existing scholarship. With
few attempts to generalize beyond the idiosyncrasies of an individual case, subnational conflict research has produced a series of contradictory findings, impeding the accumulation of knowledge. As a result, social scientists have been unable to answer the kinds of questions that activists, policymakers and ordinary citizens care about most, such as when and where violence is likely to occur, how it is likely to unfold, and – importantly – whether the lessons of previous conflicts apply to new ones.

As a result, to a greater extent than macro-level work, disaggregated, sub-national research on civil conflict has faced steep problems of external validity, perpetuating disagreements over how idiosyncratic these findings really are. The discovery of a positive relationship between, say, repression and dissent in country A is not evidence of a similar relationship in country B, yet most sub-national analyses stop at the borders of country A, without attempting to validate results with data from other conflicts. Also unknown is how robust empirical results are across sources and measurement strategies – whether media-generated data agree with archival sources, or whether manually-classified event reports tell the same story as data collected with automated techniques, like natural language processing and machine learning. Problems and questions abound.

The chief barrier to generalizability is not a lack of data on violence or covariates in country B – in many cases these data exist and are in the public domain. The problem is that no one has yet undertaken the entrepreneurial effort to merge and combine these disparate sub-national conflict datasets into a unified, analysis-ready format, with consistent theoretical constructs, definitions, measurement and levels of analysis. Without such an effort, the field cannot move forward. This is where the current project enters.

As conceived, xSub addresses the problems above explicitly. xSub is a web-based program that pulls together dozens of existing sub-national databases, and aggregates the relevant conflict events and covariates to consistent units of analysis across countries and conflicts. As a public good, xSub will significantly reduce the barriers to comparative sub-national research, empowering researchers to better situate their theoretical investigations and quickly construct custom, analysis-ready datasets. Similar initiatives have long existed for macro-level cross-national conflict data, most notably the EUGene and NewGene software packages developed by Bennett, Poast and Stam. In contrast, no resource of this kind currently exists for sub-national conflict data.
A need for multi-national multi-data conflict research

To take stock of the rapidly growing literature on subnational conflict, we surveyed the universe of topically-related studies published between 2006 and 2017 in top political science and international relations journals, including *American Political Science Review* (APSR), the *American Journal of Political Science* (AJPS), the *International Organization* (IO), *World Politics*, *Journal of Peace Research*, and *Journal of Conflict Resolution* – 392 articles in all. We organized these articles by topic, geographic and temporal scope, unit of analysis, and empirical method used. Appendix Table ?? enumerates the individual studies and summarizes their attributes.

Our survey revealed five common problems in the literature.

1. **Most sub-national studies don’t (really) use sub-national data.** While the development of multi-national disaggregated datasets like GED and ACLED have facilitated cross-national comparisons of sub-national trends, 94 percent of the studies that use such data continue to aggregate sub-national events to the country level, eschewing within-country geographic variation altogether. Unless researchers have strong theoretical reasons to focus only on macro-level patterns, this practice suggests that the field has under-utilized the potential of these new resources.

2. **Most sub-national studies focus on a single or limited set of countries.** At the other end of the spectrum, efforts to explain within-country variation often come at the expense of generalizability. 46 percent of sub-national studies focus on a single country, and 56 percent focus on a single region – like Sub-Saharan Africa or the Middle East. The tendency of sub-national empirical work to focus on individual countries and regions has led to a geographic fragmentation of the discipline. For instance, 64 percent of studies that examine the relationship between climate and conflict use data on Africa, as do all studies on reporting bias. By contrast, Africa represents less than one percent of studies on indiscriminate violence, less than one

---

284% of these studies were published after 2010, and the median study was published in 2013. These studies use a variety of empirical approaches – from ethnographic field work and interviews to regression analysis and computational modeling – to study conflict onset, duration, intensity, and diffusion varying across types (e.g., civil war, terrorism, genocide, human rights violation and protest). Their theoretical focus covers a wide range of causes and consequences of conflict, including climate variability, displacement, economic shocks, ethnic geography, elections, indiscriminate violence, information and communication technologies, occupation policies, political exclusion, reporting bias, repression and dissent, state capacity, among others (Table ??). Their spatio-temporal scope covers all major regions of the world (Figure 2), from 1400 to the present day.
percent of studies on repression and dissent, and zero studies on occupational politics (Table 1).

3. **Cross-dataset comparisons are very rare.** 65 percent of the literature has used pre-existing conflict datasets, while the remaining 33 collected original data. Yet different datasets employ different sampling and measurement strategies, and rely on different types of sources (e.g., media, NGO’s and/or governments) – choices that can be highly consequential for theory-testing and statistical inference (Tilly, 1978, Eck, 2012, Davenport and Moore, 2015, Salehyan, 2015). Despite these concerns, only 9 of the studies we surveyed use cross-dataset validation, and only three of these articles actually discuss data quality as a potential problem for conflict research (Eck, 2012, Hammond and Weidmann, 2014, Ward et al., 2013).

4. **Operational definitions vary greatly.** Among the many things that vary across data sources are typologies of violence and actors under examination. Theoretically, most studies (91 percent) focus on violence by government agents and the opposition, sometimes disaggregating further by tactics (e.g. indiscriminate vs. selective). Large data collections like ACLED, GED and SCAD provide highly-specific descriptions of actors – down to the name of the name of the protest movement, rebel unit or military commander. Some, like GED, also provide brief, qualitative textual descriptions of individual events. While these resources offer infinitely customizable choices for researchers, they also pose non-trivial challenges: determining which actors belong to pro-government or opposition factions requires some familiarity with individual cases, while the extraction of information on tactics from textual descriptions can be computationally costly. These challenges have limited the scalability of sub-national research projects to multiple countries, and made results less comparable across countries.

5. **There are no consistent units of analysis.** While cross-national analyses have generally converged on country-years or, to a lesser extent, dyad-years as common units of analysis, there is no similar “industry standard” in sub-national work. Figure 1 summarizes this variation in units of analysis, across topics. The larger the size of the dot, the more studies from our sample cover a certain topic at the given unit of analysis. Research on some topics – like indiscriminate violence and economic shocks – has been quite diverse in spatial and temporal scales. Others – like foreign
aid and elections – have favored more specific units, like country-years, and avoided sub-national units commonly used in other domains, like artificial grid cells. This inconsistency has precluded comparisons across countries and datasets. Without a standard “menu” of spatio-temporal aggregation options, it is not surprising that even most studies that use subnational data still aggregate them to the national level prior to analysis.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Africa</th>
<th>Americas</th>
<th>Asia</th>
<th>Europe</th>
<th>Multinational</th>
<th>Global</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Variability</td>
<td>18</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>Conflict Specifics</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>18</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>Data Quality</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Diffusion</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>Displacement</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>Economic Shocks</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>Elections</td>
<td>8</td>
<td>5</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Ethnic Geography</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>28</td>
<td>1</td>
<td>46</td>
</tr>
<tr>
<td>Foreign Aid</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Group Characteristics</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Indiscriminate Violence</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Information Communication Technology (ICT)</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Occupation Policies</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Political Exclusion</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Public Opinion</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Reporting Bias</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Repression and Dissent</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>State Capability</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>31</td>
<td>1</td>
<td>53</td>
</tr>
</tbody>
</table>

We address these drawbacks by applying a unified coding and aggregation scheme to 21 data sources, thereby allowing researchers to quickly assemble analysis-ready sub-national conflict datasets and easily compare their findings across countries and sources. While xSub does not eliminate the need for researchers to be aware of the limitations of the data sources they are using, our approach will make these data opportunities more accessible, allowing users to replicate their analyses in different empirical settings, and eliminate some of the greatest barriers to the accumulation of knowledge.

What is xSub?

xSub is a web-based “database of databases” for quantitative disaggregated research on conflict. The project’s online repository (www.x-sub.org) currently includes 337 datasets on the location, dynamics and intensity of political conflict and violence, in 139 countries (1942-2016), from 21 data sources, organized into consistent categories and units of analysis. It also includes several popular covariates on local weather, ethnicity, demographics,
and geography. As such, xSub is well-suited for both single-country and cross-national analyses of conflict onset, type, diffusion and (de)escalation, as well as more specialized investigations of the intensity of conflict and violence, causes and consequences of civilian victimization, and strategic interactions between governments, rebels, and third-parties.

XSub data are customizable by country and units of analysis, and are formatted for integration with standard statistical software packages (e.g. Stata, R), and Geographic Information Systems (GIS). Figure 2 shows the geographic distribution of data in xSub, with darker colors indicating more data sources for a given country. The current section details xSub’s data sources, event typology, units of analysis and included covariates.

Data Sources

xSub features event data on conflict and violence from 21 sources, including widely-used large-scale data collections and studies of individual countries and regions. At the

---

3 Each file contains a spatial unit ID that can be used to merge the event data to GIS spatial geometry files, like ESRI polygon shapefiles (.shp), and R SpatialPolygons (.RData).

4 These sources currently include Armed Conflict Location and Event Data Project (Raleigh et al., 2010), Empirical Studies of Conflict Project (Wigle, 2010, Berman, Shapiro and Felter, 2011, Calderón et al., 2015, Bueno de Mesquita et al., 2013), Political Instability Task Force’s Worldwide Atrocities Dataset (Schrodt and Ulfelder, 2016), the Social Conflict Analysis Database (Salehyan et al., 2012), UCDP’s Georeferenced Event Dataset (Sundberg and Melander, 2013), American Bar Association’s Darfur data (Totten, 2006), Beissinger’s and Zhukov’s datasets on the former Soviet Union (Beissinger, 2002, Toft and Zhukov, 2015, Zhukov, 2014, 2016b, Baum and Zhukov, 2015).
atomic level, the data are individual events, with information on the location and date of an incident of violence or protest, the actors involved, and tactics they used. We use these events to construct local event counts, at various levels of analysis (see section 2.4 below).

**Actors**

xSub organizes violent events from each data source and country by the actors who initiated them, grouped into four categories: governments (Side A), opposition groups (Side B), civilians (Side C), and third parties not directly challenging or working with the government (Side D):

Side A: The *government* category includes agents of an incumbent political regime, pro-government militias, activists, and third parties acting on an incumbent’s behalf (e.g., military contractors, foreign troops providing military assistance). It excludes mutinous factions of the military, and supporters of an ousted regime.\(^5\)

Side B: The *opposition* category includes rebels, dissidents, revolutionaries, anti-government militias, third parties acting on the rebels’ behalf, and other armed groups directly challenging the government. It also includes anti-government rioters and protesters that employ violent or non-violent tactics.

---

\(^5\)For instance, from 1996 to 2001, the Taliban held power in Afghanistan and is coded in our data archive as government; post 2001, we code it as a rebel group.
Side C: The *civilians* are unarmed victims of violence by any side of the conflict. This group includes individuals who abstain from willful participation in contentious behavior and violence, and generally do not enter the dataset as initiators of conflict events.\(^6\)

Side D: The final category (*other*) includes local militias, tribes, self-defense units, and other non-state actors *not* directly challenging or working with the government. This group also includes factions we were unable to neatly classify into the first three groups, such as inter-communal groups and criminal organizations.

xSub’s data sources differ in the kinds of actors they include and the typology they use to organize them. For example, the Social Conflict Analysis Database (SCAD) lies on one side of the spectrum, and provides unstructured information on actors for each event, with highly-detailed descriptions such as “airline workers,” “protesting women,” “Mungiki gang in Kenya” or a “Prokot Militia” (Salehyan et al., 2012). The Empirical Studies of Conflict (ESOC) data on drug-related violence and homicides in Mexico lies on the opposite end, and contains no information on perpetrators (Calderón et al., 2015). Between these extremes, some sources group actors’ affiliations into pre-defined categories.

If the data source did not contain an “off-the-shelf” classification, but only a detailed description of involved actors, we constructed source- and country-specific dictionaries mapping actors to each of our four categories.\(^7\) The xSub website provides detailed Actor Dictionaries for the four largest sources in our data archive: Armed Conflict Location & Event Data Project (ACLED), Uppsala Conflict Data Program Georeferenced Event Dataset (UCDP GED), Political Instability Task Force (PITF) Worldwide Atrocities Dataset, and Social Conflict Analysis Database (SCAD).

**Actions**

We categorized events by these actors into four types of actions:

1. *Any* use of force

2. *Indiscriminate* use of force, including indirect fire, shelling, air strikes, chemical weapons

---

\(^6\)For instance, we classify unarmed individuals protesting the government as members of the opposition (Side B), and classify armed civilians in local self-defense battalions as members of the fourth group (Side D), as described below.

\(^7\)If the data contain no information on actors, we referred to the original source (in most cases either the published article or the author/s). If we were not able to obtain such information, we labeled *Initiator* and *Target* as “Other.”
3. *Selective* use of force, including direct fire, arrest, assassination

4. *Protests*, both violent and non-violent

While some data sources contain information on the types of actions they include, most do not. Where a source provided no details, we either coded such violence as *Any* or referred to the original article or the author. Where a data source instead provided a textual description of each event, we constructed a custom action dictionary, and used natural language processing to categorize the event into one of the four action categories. The xSub codebook, available at [x-sub.org](http://x-sub.org), details our event typology. Table 2 summarizes the typology.

<table>
<thead>
<tr>
<th>Tactic</th>
<th>Initiator</th>
<th>subtotal:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Side A (government)</td>
<td></td>
</tr>
<tr>
<td>Any violence</td>
<td>SIDEA_ANY</td>
<td>ACTION_ANY</td>
</tr>
<tr>
<td>Indiscriminate violence</td>
<td>SIDEA_IND</td>
<td>ACTION_IND</td>
</tr>
<tr>
<td>Selective violence</td>
<td>SIDEA_SEL</td>
<td>ACTION_SEL</td>
</tr>
<tr>
<td>Protests</td>
<td>SIDEA_PRT</td>
<td>ACTION_PRT</td>
</tr>
<tr>
<td></td>
<td>subtotal:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIDEA_ANY</td>
<td></td>
</tr>
</tbody>
</table>

**Units of Analysis**

xSub aggregates event counts and covariates to users’ preferred spatial and temporal units of analysis. The geographic units include (1) *countries*, (2) *provinces*, or first-order administrative divisions, like U.S. states, (3) *districts*, or second-order administrative divisions, like U.S. counties, (4) *PRIO-GRID cells*, from a regularly-spaced and equally-sized 0.5 x 0.5 decimal degree lattice, and (5) *electoral constituencies*, the size of which is generally between that of a district and a province.\(^8\) Temporal units include (1) years, (2) months, and (3) weeks. For example, if the user selects “Space: district” and “Time: week” from the data page on [x-sub.org](http://x-sub.org), xSub will generate repeated weekly observations for each district, with the total number of violent incidents at each time period (broken down by actor and tactics), and local average statistics for weather and other covariates. Figure 3

---

\(^8\)Geographic units used by the Constituency-Level Elections Archive (CLEA): [http://www.electiondataarchive.org](http://www.electiondataarchive.org).
illustrates how one’s choice of spatial units affects the geographic distribution of violence. Figure 4 does the same for temporal units. In both cases, more fine-grained resolutions yield more variation, but also greater sparsity (i.e. lower event counts).

Spatial aggregation is not always straightforward. Matching events to spatial units requires geographic coordinates, which not every raw data source contained. If the original source did not include coordinates for each event, but only an address line (e.g. name of city or village, like Beissinger 2002), we used the latter information to geo-code events, with a suite of mapping APIs from Google, Yandex and MapQuest. We geo-coded events to the most precise location possible.9

Covariates

In addition to violence and protest event counts, xSub includes numerous variables that are frequently used to examine political conflict and violence: local statistics on demographics (e.g. population density), geography (e.g., elevation, roads), ethnicity (e.g. nationalities and linguistic groups residing in a given site) and weather (e.g., local temperature and rainfall). To make these covariates consistent and maximally comparable across countries and sources, we drew this information from global and publicly-available GIS datasets, like the Geo-referencing of Ethnic Groups dataset (Weidmann, Rød and Cederman, 2010) and the Global Land Cover Characterization (GLCC) dataset. The codebook, available at x-sub.org, contains detailed descriptions of these covariates. Some of our covariates (e.g., climate variables) date back to 1900. While some of these are time-variant (e.g., weather), others are static (e.g., elevation).

Illustration of use: Repression and dissent

As a demonstration of how scholars might use xSub, we investigate the empirical relationship between overt manifestations of violent repression and overt manifestations of dissent. One of the longest-standing topics in research on civil conflict is the contentious interaction between governments and challengers: how the actions and tactics of one side influence the actions and tactics of another, and whether escalation on one side sparks reciprocal steps from the opponent (Davenport, 2007). The dominant view in the literature is that repression tends to inflame dissent – by cracking down on protesters and other op-

9For instance, Beissinger 2002 contains Soviet-era names and spellings of locations in former Soviet republics and cities. Where appropriate, we updated these names using the current spelling.
Traditionally, research on government-challenger interactions has maintained a cross-national focus, tracking macro-level trends across conflicts, but overlooking variation within them (Hibbs, 1973, Muller, 1985, Buhaug, 2006, Slettebak, 2012, Braithwaite, Braithwaite and Kucik, 2015). A growing movement in sub-national research has disaggregated actors and actions to a more granular local level, but mostly within a single country or conflict (Beissinger, 2002, Berman et al., 2011, Maystadt and Ecker, 2014). Due to plurality of idiosyncratic sub-national research designs and data sources, the field has produced contradictory findings about when and where these confrontations are likely to occur, how they are likely to unfold, and whether the experience of past conflicts apply to new conflicts in other countries and regions. xSub permits us to bridge the gap between cross-national and sub-national research on repression and dissent, and examine whether sub-national patterns in individual countries are part of a general trend.

To take stock of whether repression (i.e., government violence) increases or decreases subsequent opposition activity, we use xSub to conduct a meta-analysis across
hundreds of sub-national datasets. For each country and data source in xSub, we fit the following core model specification:

\[ \text{Dissent}_{it} = \text{Repression}_{it-1} \beta_1 + \text{Repression}_{it-1}^2 \beta_2 + \alpha_i + \gamma_t + u_{it} \]  

(1)

where \( \text{Dissent}_{it} \) is the number of protests in locality \( i \) at time \( t \) (ACTION_PROTEST in xSub), \( \text{Repression}_{it-1} \) is the number of government uses of force in \( i \) during the previous time period \( t-1 \) (SIDEA_ANY). We include locality fixed effects \( \alpha_i \) to account for unobserved local factors influencing both repression and dissent, and time fixed effects \( \gamma_t \) to account for common shocks over time. To make estimates maximally comparable across countries and data sources, we report standardized coefficients (i.e. the impact of standard deviation increase in repression on standard deviation changes in dissent).

Our interest is in how the two \( \beta \) coefficients vary across conflicts. The relationship between repression and dissent is strictly inflammatory if \( \beta_1 > 0, \beta_2 \geq 0 \), indicating that increases in repression are followed by linear (\( \beta_2 = 0 \)) or exponential (\( \beta_2 > 0 \)) increases in protests. If \( \beta_1 < 0, \beta_2 \leq 0 \), then the relationship is negative, with dissent declining after repression. If \( \beta_1 > 0, \beta_2 < 0 \), then the relationship is “upside-down U”-shaped, where increases in repression are correlated with increases in protests, but the rate of increase gradually declines and potentially reverses as repression escalates. Finally, \( \beta_1 < 0, \beta_2 > 0 \) would indicate the opposite, “U-shaped” relationship.

We used the PRIO GRID cell as our spatial unit of analysis, and the month as our temporal unit of analysis. To ensure sufficient variation for model estimation, we limited our sample to datasets with at least 10 incidents of government violence and 10 protests. This approach narrows our empirical domain to 149 individual conflict datasets (out of 337 in xSub) from 8 sources, including four large collections (ACLED, GED, PITF, SCAD) and four individual studies.\(^\text{10}\) Figure 5 reports \( \beta_1 \) and \( \beta_2 \) estimates across these datasets, along with a weighted mean.\(^\text{11}\)

The results show overwhelming evidence for a curvilinear relationship between repression and dissent (\( \beta_1 > 0, \beta_2 < 0 \)). On average, a standard deviation increase in repression increases protests by \(.14 \) (95% confidence interval: \(.11, .18 \)) standard deviations,\(^\text{14}\)

\(^\text{10}\) The individual studies are Bueno de Mesquita et al. (2013) [ESOC_PakistanBFBS], Baum and Zhukov (2015) [yzLibya], Toft and Zhukov (2012) [yzCaucasus2000], Zhukov (2016) [yzUkraine].

\(^\text{11}\) Formally, the weighted mean is \( \bar{\beta} = \sum_d w_d \beta_d \), where \( \beta_d \) is the coefficient from dataset \( d \), and \( w_d \) is a model weight, here proportional to sample size of \( d \) (\( N \times T \)). We estimated confidence intervals through 1000 bootstrap replicates of \( \bar{\beta} \).
but also decreases the growth in protests by .08 standard deviations (95% CI: -.11, -.04). If we complete the square \((-\frac{\beta_1}{2\beta_2}\)), we can estimate that the slope of the curve should begin to change from positive to negative when governments escalate violence to .96 standard deviations (95% CI: .50, 2.1) above the mean. What these numbers mean varies across conflicts. In the North Caucasus region in 2000-2012, violence by Russian security forces began to suppress dissent after 19 operations per locality-month. During unrest in Ukraine in 2014, meanwhile, protests declined after just 3 local operations by government forces.

Table 3 summarize the coefficient estimates by region. The relationship is curvilinear in a plurality of conflicts, in every region except Oceania (47% overall). By contrast, the relationship was inflammatory in just 13% of conflicts, and suppressive in 9%.

Table 3: Estimated Repression-Dissent Relationships, by Region. Numbers represent number of datasets in which coefficient estimates were significant, and had the specified signs.

<table>
<thead>
<tr>
<th>Region</th>
<th>Inflammatory ((\beta_1 &gt; 0, \beta_2 \geq 0))</th>
<th>Suppressive ((\beta_1 &lt; 0, \beta_2 \leq 0))</th>
<th>Upside-down U ((\beta_1 &gt; 0, \beta_2 &lt; 0))</th>
<th>U-shaped ((\beta_1 &lt; 0, \beta_2 &gt; 0))</th>
<th>Null ((\beta_1 = \beta_2 = 0))</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eastern Africa</td>
<td>6 (16.7%)</td>
<td>7 (19.4%)</td>
<td>12 (33.3%)</td>
<td>1 (2.8%)</td>
<td>10 (27.8%)</td>
<td>36</td>
</tr>
<tr>
<td>Middle Africa</td>
<td>2 (14.3%)</td>
<td>0 (0%)</td>
<td>8 (57.1%)</td>
<td>2 (14.3%)</td>
<td>2 (14.3%)</td>
<td>14</td>
</tr>
<tr>
<td>Northern Africa</td>
<td>3 (16.7%)</td>
<td>2 (11.1%)</td>
<td>9 (50%)</td>
<td>0 (0%)</td>
<td>4 (22.2%)</td>
<td>18</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>2 (22.2%)</td>
<td>2 (22.2%)</td>
<td>4 (44.4%)</td>
<td>0 (0%)</td>
<td>1 (11.1%)</td>
<td>9</td>
</tr>
<tr>
<td>Western Africa</td>
<td>2 (5.7%)</td>
<td>1 (2.9%)</td>
<td>22 (62.9%)</td>
<td>0 (0%)</td>
<td>10 (28.6%)</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>15 (13.4%)</td>
<td>12 (10.7%)</td>
<td>35 (49.1%)</td>
<td>3 (2.7%)</td>
<td>27 (24.1%)</td>
<td>112</td>
</tr>
<tr>
<td>Americas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caribbean</td>
<td>1 (25%)</td>
<td>0 (0%)</td>
<td>1 (25%)</td>
<td>0 (0%)</td>
<td>2 (50%)</td>
<td>4</td>
</tr>
<tr>
<td>Central America</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>2 (50%)</td>
<td>0 (0%)</td>
<td>2 (50%)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1 (12.5%)</td>
<td>0 (0%)</td>
<td>3 (37.5%)</td>
<td>0 (0%)</td>
<td>4 (50%)</td>
<td>8</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South-Eastern Asia</td>
<td>1 (16.7%)</td>
<td>0 (0%)</td>
<td>1 (16.7%)</td>
<td>0 (0%)</td>
<td>4 (66.7%)</td>
<td>6</td>
</tr>
<tr>
<td>Southern Asia</td>
<td>2 (18.2%)</td>
<td>2 (18.2%)</td>
<td>4 (36.4%)</td>
<td>1 (9.1%)</td>
<td>2 (18.2%)</td>
<td>11</td>
</tr>
<tr>
<td>Western Asia</td>
<td>1 (12.5%)</td>
<td>0 (0%)</td>
<td>5 (62.5%)</td>
<td>0 (0%)</td>
<td>2 (25%)</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>4 (16%)</td>
<td>2 (8%)</td>
<td>10 (40%)</td>
<td>1 (4%)</td>
<td>8 (32%)</td>
<td>25</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>2 (66.7%)</td>
<td>0 (0%)</td>
<td>1 (33.3%)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>2 (66.7%)</td>
<td>0 (0%)</td>
<td>1 (33.3%)</td>
<td>3</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melanesia</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (100%)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (100%)</td>
<td>1</td>
</tr>
<tr>
<td>All</td>
<td>20 (13.4%)</td>
<td>14 (9.4%)</td>
<td>70 (47%)</td>
<td>4 (2.7%)</td>
<td>41 (27.5%)</td>
<td>149</td>
</tr>
</tbody>
</table>
Figure 5: **META-ANALYSIS OF REPRESSION AND DISSENT.** Quantities reported are standardized coefficient estimates and 95% confidence intervals for the model in Equation 1, across 8 datasets.
Limitations and extensions

The goal of the above analysis has been purely illustrative, demonstrating how one may use xSub to assess the correlation between repression and dissent at the local level, and how this correlation varies – in direction and magnitude – across conflicts and data sources. These models cannot – and are not intended to – identify a causal effect. A more rigorous analysis with xSub might consider additional sources of variation and bias, like the endogeneity of repression (Hibbs, 1973, Oliver and Myers, 2002, Ritter and Conrad, 2016, Rozenas, Schutte and Zhukov, 2017), interdependence across countries and actors (Carey, 2006), tactical shifts (Moore, 1998), and spatial autocorrelation (Warren, 2015).

Further building on this example, xSub allows scholars to examine potential heterogeneities in the repression-dissent nexus. For instance, because some instances of repression take place in civil war contexts and others do not, an empirical focus on protests may overlook variation in other forms of opposition activity, like armed insurrection. Because some forms of repression are very lethal and others are not, we may also expect indiscriminate artillery shelling to have a very different effect on dissidents than arrests and detentions. Due to space constraints, we did not differentiate by conflict type or form of government violence in our analysis here. However, xSub enables users to explore these and many other more granular patterns.

Conclusion

This article introduces a unified framework and data repository for the sub-national study of conflict and violence. xSub seeks to address five problems in the empirical conflict literature: (1) a tendency to under-utilize sub-national data; (2) a tendency of sub-national studies to focus on a single or limited set of countries; (3) a lack of cross-dataset comparisons; (4) inconsistent operational definitions of violence; and (5) inconsistent units of analysis. xSub helps correct these shortcoming by bringing together sub-national data on hundreds of conflicts from multiple sources, and organizing these events into a consistent set of categories and spatio-temporal units. In so doing, xSub significantly reduces the barriers to comparative sub-national research, empowering researchers to quickly construct custom, analysis-ready datasets, pre-loaded with several popular covariates.

Looking ahead, our hope it to establish xSub as a platform for scholars to contribute and distribute their own, original data. One of the reasons for fragmentation in sub-national research is that many individual data collection efforts are project-specific: schol-
ars assemble a new dataset for a one-off paper, and – apart from posting a replication archive – never re-use those same data again. Rather than allow a dataset to “die” with a paper, xSub enables researchers to give their data a second life, in the hands of new researchers, asking a wholly new set of empirical questions.

References


Carey, Sabine C. 2006. “The dynamic relationship between protest and repression.” Polit-


