4

Coding Protest Events in Autocracies

In this chapter, we begin our empirical investigation into the effect of Internet penetration on political protest. Clearly, our analysis is not the first of its kind. As discussed above, previous work has typically fallen into one of two categories. The first type of study has a relatively narrow geographic and temporal scope and usually focuses on individual protest episodes, such as the Tahrir Square protests in Egypt in 2011 (Tufekci and Wilson, 2012) or the Arab Spring as a whole (Hussain and Howard, 2013). While this approach can reveal interesting insights for a particular case, it is difficult to generalize these insights. The second type of study analyzes protest by comparing different countries and time periods (Brancati, 2014; Ruijgrok, 2016). These studies, however, typically work with aggregated, country-level indicators such as annual protest counts, which is why they cannot capture patterns *within* these countries. For this reason, we use a combination of both approaches. We study the relationship between Internet access and protest occurrence for a global sample of autocracies, but also analyze variation between their major cities for each case. This allows us to examine how the density of the Internet in particular places affects protest, but also how this effect varies across different national contexts. This disaggregated analysis obviously requires fine-grained data, not only for the outcome in question (protest), but also for the degree of Internet penetration as well as other variables in the analysis. In this chapter, we focus on the former before introducing our research design in the next chapter. We introduce a new dataset on protest in

autocracies—the Mass Mobilization in Autocracies Database (MMAD)—which contains information on individual protest incidents, including dates and locations. The data collection is based on information from media reports that identify instances of protest. The use of information from the media raises issues regarding the selectivity and accuracy of the reported information. Therefore, we briefly review existing data collection approaches to political protest before describing how our new coding effort deals with these challenges.

4.1 Existing Data on Political Protest

Political protest has been a frequently studied topic in the social sciences for decades, particularly in political science and sociology. For that reason, there are a number of data collections on protest at the national and local levels. Almost all of these collections are based on information reported in the media, from which a structured list of protest events was created. We can distinguish between *human-coded* datasets, where information from other sources is extracted entirely by humans, and *automatic* coding, where computer-based text analysis is used for this task. Human coding, while producing high-quality content and being extremely flexible in terms of the type of information that can be extracted, is very costly and time-consuming. Automated coding is good for categorizing news articles into various topics of interest or extracting actors from known actor lists ("actor dictionaries"), but may have shortcomings when it comes to the precision of the extracted information from the text (e.g. the number of fatalities or the date of the incident). We briefly review existing human- and computer-coded datasets before introducing the hybrid approach used for the MMAD, which combines the two approaches at different stages of the coding process.

4.1.1 Human Coding

When it comes to the study of protest in non-democratic environments, most of the large-N analyses on mass mobilization and regime stability utilize data from the *Cross-National Time-Series Data Archive* (CNTS, Banks, 2011). It contains count variables of domestic disturbances (riots, anti-government demonstrations) at the country-year level from 1815 until today. The dataset is unique with regard to its extensive spatial and temporal information and is widely used. However, the highly aggregated nature of the data makes micro-level analysis—such as the one presented in this book—impossible. The events are derived exclusively from a single newspaper, the *New York Times*, which can lead to gaps in reporting compared to collections

that draw from multiple sources (Salehyan et al., 2012).

Currently, there are a number of available datasets that contain both spatially and temporally disaggregated information on collective action events. These datasets are coded by humans who extract event-level information from media reports. An early dataset of this kind is Francisco's (2006) *European Protest and Coercion Data*, a collection of daily, city-level information about protests. Due to the focus on European states and the 1980-1995 period, however, the dataset includes almost exclusively democratic states, and is thus of limited use to students of autocracy. Recently, a number of data projects have emerged on collective action and violence in civil wars. One of these efforts is the *Geo-referenced Event Dataset* (GED) created by the Uppsala Conflict Data Program (Sundberg and Melander, 2013). The *Armed Conflict Location and Event Data* project (ACLED Raleigh et al., 2010) and the *Social Conflict in Africa Database* (Salehyan and Hendrix, 2012) have a slightly broader focus with regard to the type of action recorded and also include large-scale protests and riots. Finally, the recently released *Mass Mobilization* (MM) dataset contains geolocated protest events from 162 countries (Clark and Regan, 2016). The level of detail and scope of these datasets is unprecedented and enables a host of interesting research questions to be pursued. However, because their main focus is on more violent forms of collective action, or because their scope and coverage is limited, these datasets may not be an optimal choice for our study on Internet and protest in autocratic regimes.

4.1.2 Automatic Coding

Rather than relying on human coding, some researchers use computer-based methods to code protests from primary and secondary sources. Automated event coders such as the *Kansas Event Data System* (Schrodt and Gerner, 1994), *TABARI* (Schrodt, 2001), and the *VRA Reader* (King and Lowe, 2003) have been used for a number of data collection efforts on contentious politics. These projects are typically much more extensive regarding the type of events they cover and the sources they can process because automation makes the coding very resource-efficient. One of these machine-coded databases is *The World Handbook of Politics IV (WHIV)* (Jenkins et al., 2012), which contains more than 250,000 observations on 40 event forms between 1990 and 2004. Although tests show that machine coding is comparable to human coding if the same coding conventions are used (King and Lowe, 2003; Mikhaylov, Laver and Benoit, 2012; Ruggeri, Gizelis and Dorussen, 2012), the data generated by these collection efforts still have a number of downsides. For one, machine-coded datasets require an *a priori* specification of relevant actors in the form of actor dictionaries—pre-specified lists of politicians, groups, or organizations involved in the events coded.

This means that every news report using alternative names, or no names at all, will not be coded. Most importantly, traditional machine-coded event datasets do not provide spatial event coordinates, and are thus unsuitable for studying subnational protest dynamics.

A more recent machine-coded event dataset, the *Global Data on Events, Location and Tone* database (GDELT, Leetaru and Schrodt, 2013), does add spatial coordinates to events. It contains more than 200 million geo-referenced events from 1979 to 2012, divided into 20 main categories with 26 subcategories for "protest". Despite this seeming advantage, GDELT has some data quality issues which make it a less than ideal choice for certain analyses. As Hammond and Weidmann (2014) show, the automatic geo-referencing used in GDELT is often inaccurate and possibly even biased, which can lead to completely different findings compared to human-coded datasets. The *Integrated Crisis Early Warning System* event dataset (ICEWS, O'Brien, 2010) uses a similar coding approach to GDELT and earlier machine-coded data in the sense that each event consists of a particular action, the source actor of this action, and the target. As with all actor dictionary-based systems, it requires the prior specification of all involved actors and actions in the form of dictionaries, which constrains the coding to known and clearly identifiable entities.

Because of the drawbacks of fully automated coding with respect to the types of actions that can be captured and its poor precision in extracting information, the MMAD project presented here uses a combination of automation and human coders to extract events from news reports. The coding procedure differs markedly from existing projects in a number of ways, however. In the next section, we identify common problems in the coding of data from media reports and show how the MMAD addresses these.

4.2 Coding Protest Events from Media Reports

As described above, a typical event coding process turns the source material into a standardized list of protest events, each of which has a set of attributes (for example, the number of protesters, location, issue, etc.). While this process sounds straightforward and easy to implement, a number of potential problems can arise in doing so, particularly those relating to source selection, report selection, and information extraction and -aggregation. We first introduce these problems before discussing the solutions developed for the MMAD.

The source material for event coding consists of "reports" (or "articles") covering political developments in the countries of interest. Typically, these reports come from

news outlets such as newspapers, magazines, or news agencies, but can also stem from international or non-governmental organizations operating on the ground. Öberg and Sollenberg (2011) provide an excellent overview of the process by which actual events are reported by news outlets and eventually reach the consumer, which we do not discuss in detail here. The number of available media sources is considerable, and limited resources typically require coders to include only some of them. Thus, the *source selection problem* refers to the challenge of having to select a limited number of sources for a coding project. This selection will of course vary with the scope of the project: for example, if we are interested in coding protest in a single country, we may be best served by relying on local newspapers, as they are likely to provide the best coverage of these events.

However, the use of local sources can be problematic when coding protests *across* countries, which is the aim of our MMAD project. First, the availability of news stories can differ widely between countries. While some countries have excellent and regular local newspaper coverage, in others coverage may be spotty. This can lead to widely differing numbers of protest events that can be recorded, simply because there are few (or no) sources reporting them in the first place. Ultimately, this makes it difficult to compare codings across countries, since we do not know whether a low number of protest events is the result of little actual protest activity, or a consequence of low reporting. A second, practical problem in using local sources is language. Whether done by humans or computers, the processing of language requires language-specific skills or software. In a large cross-national project, however, it is simply not feasible to analyze sources in many different languages. For these reasons, the standard approach for most cross-national event coding projects—including MMAD—is to rely on international, English sources only. Yet this restriction does not fully solve the source selection problem, as there are still a large variety of English-language news sources to choose from that need to be narrowed further. Below, we return to this problem and describe our strategy for solving it.

Even with a fixed set of sources from which to draw protest data, we cannot proceed straight to the coding of events. This is because it is difficult to select relevant reports from the respective sources, i.e. those that actually cover political protest. In many cases, database searches are limited to simple keywords and return news reports containing terms such as "protest" or "riot". The resulting set of reports is usually very large, since many of the search terms do not unambiguously refer to political protest. Only few of these articles are relevant for the coding of political protest—for MMAD, their proportion is less than 5%. Thus, the second problem we face when coding protest events from news sources is the *report selection problem*—in other words, the problem of choosing articles that are relevant for the coding of political protest. In order to reduce the large number of irrelevant reports, MMAD uses an

automated approach based on computational text analysis and machine learning. We describe this approach below.

Given a set of relevant articles, a (human or machine) coder then faces the task of creating a standardized list of protest events from them. This final stage of the coding process poses additional challenges: The third problem of event coding is the *information extraction problem*, which refers to the identification of those parts of a news report that contain information about the key variables of interest. For example, the sentence "A group of 300 people rallied in Bishek on December 4 to protest the ban of the ruling party" contains information about the date, location, number of protesters, and the issue of the protest, all of which should eventually be recorded in the protest database. The coder's task is to identify these pieces of information, convert them into the format used by the database, and enter them in the correct field in the coding form.

Beyond information extraction, a final and fourth problem remains to be solved. During ongoing episodes of protest, it is rarely the case that only a single source reports about a particular protest event. Quite the opposite, we frequently receive reports about a particular protest event from multiple sources, and sometimes even multiple reports from the same source. These different reports often contain conflicting information, for example, about the number of protesters or the level of violence. How do we turn this information into a single entry in our final database? In order to do this, we need to solve the problem of aggregating multiple reports into a single event coding, i.e. the *aggregation problem*. To build on the above example of different sources reporting different numbers of participants: Should we consider one of the sources to be more trustworthy, and thus prefer its estimate over others? In a scenario where multiple sources agree on the number of participants, the reported number is probably more reliable compared to a situation where the estimates differ widely across sources. Existing event coding projects are largely opaque when it comes to this problem. If anything, we know the number of reports a particular event is derived from, but not how the information was eventually aggregated into the final coding for this event. For MMAD, we developed a revised coding process that attempts to resolve this issue (see below).

In sum, when creating an event database from news sources, there are four problems we need to address: (i) source selection, (ii) report selection, (iii) information extraction, and (iv) information aggregation. When researchers devise solutions to these problems, they need to balance several requirements for their event database:

- **Completeness** of coverage, or the degree to which the dataset includes the events of interest

- **Feasibility**, or the extent to which the coding effort can be completed with the available resources
- **Transparency and reliability**, or the degree to which the coding process can be understood and replicated by others, with similar results.

The first two requirements are at odds with one another. While coverage is certainly never complete in the sense that we can never record all events that fit our coding criteria, the selection of the sources (the first problem) and/or relevant articles from these sources (the second problem) has a considerable impact on the number of events we fail to record. As mentioned above, relying on local sources can in some cases improve the completeness of our dataset, but it fails to satisfy the second requirement in that the effort quickly becomes unfeasible. This is similar for the selection of reports. If we fail to drop news articles that are irrelevant for our coding, the volume of articles to be processed would require an amount of manual labor that quickly exceeds the limits of a typical research project. Similarly, the third requirement, transparency and reliability, can conflict with feasibility. In order to make it possible to understand why an event was coded with particular attributes, the coding decisions should ideally be documented in detail, along with the respective parts of the original report they were derived from (the third problem). In particular, coders would have to document how they integrated potentially conflicting information from different sources into single events (the fourth problem). Due to feasibility reasons, this is hardly ever possible. Still, we can improve on existing event coding approaches by implementing a number of innovations for event coding, which we describe in the following sections.

4.2.1 Source Selection

Above, we discussed the problem of selecting media sources for event coding. Since media outlets cater to different audiences and have varying geographic scope, the choice of an outlet matters tremendously for how many protests we "see" (Jenkins and Maher, 2016). To some extent, the problem can be addressed by relying on news *agencies* rather than newspapers. News agencies produce media content for a large number of outlets in different countries, which is why they are typically much broader and inclusive in their coverage as compared to national or local media. Moreover, the fact that these agencies are much bigger in size makes it possible for them to have reporters in many different locations around the globe, which increases scope and depth of coverage. Still, even if we rely solely on English-language news agencies, we will unavoidably have some bias in coverage, as these agencies typically cater to audiences in Western countries and primarily report on events that are of

some interest to them. The choice of sources is especially important in projects spanning the non-democratic regions of the world, where media coverage by Western media is often sparse. Nam (2006) illustrates the problem by comparing protest data on Korea and Burma collected in KINDS—a local news database—and LexisNexis. Unsurprisingly, the results show a significant increase in data comprehensiveness using the local news database. However, as discussed above, for larger coding projects the use of local news databases is often unfeasible due to cost, language, and time constraints.

In order to gauge the extent of the source selection problem and to select a good combination of sources for MMAD, we performed an extensive trial coding. The trial coding was performed on Kyrgyzstan for the period of October 2004 through June 2005 searching *all English-language sources* in LexisNexis, restricting the search to articles from that country and using a broad set of search terms.[1] This is the same search string that was later used during the active coding phase, though with a more limited number of news sources. Kyrgyzstan is a suitable sample candidate for two main reasons. First, it is a relatively liberalized autocracy where—despite its US military base and strategic location near Afghanistan, China, and Russia—we expect limited coverage in Western media. Second, in March 2005 there was a mass uprising that forced Akayev, president since independence, to flee the country. We therefore expect Kyrgyzstan to be the center of attention in late March 2005 with coverage up until then to be limited, allowing us to determine coverage by different sources both for periods of calm and unrest. All news reports were screened for events that fit the MMAD definition of mass mobilization events. The resulting data consisted of 602 event reports from 73 unique sources, which together describe 193 protest events in the nine-month period.[2]

The trial coding reveals a number of interesting patterns. We examined the news sources that identify the largest number of events: the Associated Press (AP), the Agence France Press (AFP), and BBC Monitoring. While we would assume that Western news agencies cover similar events, this is not necessarily so. According to Figure 4.1 (left panel), the AP covers the smallest number of events, and many events captured by the AFP are not recorded by the AP. Individually, the AP and AFP cover only 25% and 39% of all recorded events, respectively. Both the AP and AFP together only identify less than half of all events we obtain when examining all sources. This suggests that relying on large news agencies alone leads to the omission of a significant number of events, which is why we need other sources to

[1] Articles containing at least one of the following terms: "protest", "demonstration", "rally", "campaign", "riot", "picket".

[2] In order to aggregate the event reports to events, we used information about the date, location, and side of the protesters (e.g. pro-government, anti-government). For example, three event reports with the exact same date, location, and side designation constitute one event.

achieve higher coverage. Figure 4.1 (left panel) shows that including BBC Monitoring boosts coverage dramatically. BBC Monitoring is a service provided by the BBC that translates and aggregates news from a large number of local news sources worldwide (BBC, 2017). For this reason, it provides much more detailed coverage of events that typically do not make it into the international news. BBC Monitoring alone provides coverage of 56% of recorded protest incidents. At the same time, however, the overlap between the agencies is only partial, which means that omitting even one of them would make us lose a considerable number of events.

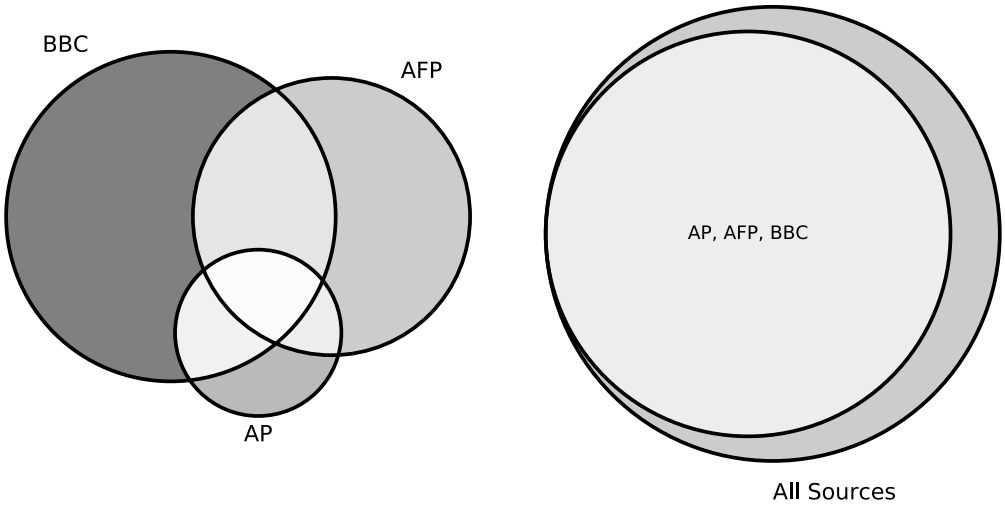

Figure 4.1: Coverage of protest events by different sources.

Figure 4.1 (right panel) shows the combined coverage of our three sources (AP, AFP, and BBC World Monitoring). Together, they cover around 84% of all events reported during the coding period. Overall, this is a good result, in particular since the number of articles from these three sources alone is quite small relative to the total number of articles from all sources: In the nine-month period of the trial coding, searching all sources returns 2,711 hits, while restricting the search to the AP, the AFP, and BBC Monitoring reduces the number to 1,023. In other words, 84% of the events come from 38% of the material. About 59% (1,606 of 2,711) of the reports using all sources are from March 2005 while 53% of all event reports were published the week leading up to the March 24 storming of the presidential palace. In comparison, 48% (495 of 1,023) of the reports searching AP, AFP, and BBC Monitoring are dated March 2005, while 29% of event reports using these sources were reported in the week leading up to March 24. This indicates that the coverage is more balanced using AP, AFP, and BBC Monitoring compared to using the complete set of sources.

While it is clearly difficult to make strong general claims regarding the source selection problem from a single case, we believe that the results from our trial coding can

provide some guidance regarding the selection of sources for MMAD. In general, we cannot expect to obtain a complete set of events using media sources alone (Jenkins and Maher, 2016). Rather, we need to accept the fact that gaps in the data will remain, although our results indicate that certain strategies help to minimize these gaps. For the three sources we examined in detail, there is a high trade-off between the number of source materials (articles) and the amount of coverage. This trade-off is likely even higher in countries with better coverage by Western media, for example Egypt and Iran. For countries outside the international spotlight, the inclusion of BBC Monitoring leads to much better coverage of protest events that are not picked up by international media, as in Kyrgyzstan. In short, the use of international news agencies such as the AP and AFP combined with local reports provided by BBC Monitoring seems to strike a balance between coverage and feasibility for our coding project. Thus, the coding of MMAD is based entirely on these three sources. Still, simple keyword searches return large numbers of irrelevant articles, which is why we need to address the report selection problem.

4.2.2 A Machine Learning Approach to Filtering News Reports

Now that we have chosen the news sources for event coding, we need to select articles covering political protest from these sources. One way to do this is by using article topics and categories defined by database providers, as for example Lexis-Nexis. However, such an approach does not satisfy scientific standards of transparency and replicability. Opaque and non-replicable methods are used to assign these tags, which makes it difficult to assess whether they can produce a reasonably complete selection of articles. Consequently, these proprietary methods were avoided and simple keyword searches were used (in combination with the respective country names) to retrieve the relevant articles for a given country. These terms had to be kept rather general in order to avoid losing too many relevant articles due to too-narrow criteria. For the MMAD, articles were retrieved by searching for *protest*, *demonstration*, *rally*, *campaign*, *riot*, or *picket*. Not surprisingly, the number of false positives (selected articles that do *not* cover political protest) in the simple keyword search is extremely high. In the first coding phase for the MMAD project, we performed an extensive human coding of roughly 250,000 articles from our three sources, covering 19 autocratic countries from different regions of the world to avoid regional biases. Figure 4.2 shows the distribution of relevant and irrelevant articles across these countries in this initial set.

Overall, the proportion of relevant articles obtained through a simple keyword search is only around 2%. In other words, a huge share of the work by human coders would involve sifting through articles that ultimately provide no information

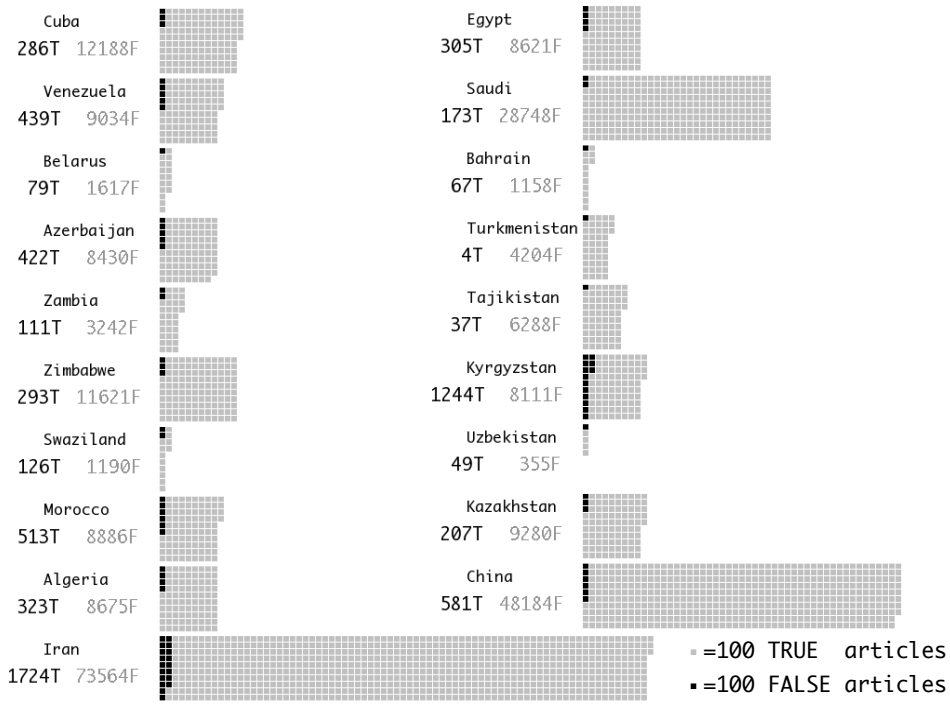

Figure 4.2: Share of articles describing political mass protest from the first coding phase (Figure from Croicu and Weidmann, 2015). The black squares correspond to 1,000 articles each that describe mass protest, the grey squares represent 1,000 articles that do not describe mass protest and thus irrelevant for the coding.

on political protest and are thus irrelevant for the project. This suggests that a coding process based entirely on human coding would not be feasible for the entire set of around 70 countries in the MMAD. Therefore, a machine learning approach was developed to eliminate a large share of the irrelevant articles, while keeping most of the relevant ones. For this task, the application of computer-assisted text classification seems reasonable, since the task is not to extract individual pieces of information from a news report (which, as discussed above, is fraught with problems), but rather to assess the general relevance of the article based on the presence of certain terms and combinations of terms. Using a *supervised* machine learning approach, the 250,000 articles from the first coding phase were used as a training set in which the computer "learned" how to detect relevant articles covering instances of political protest. Once this learning process is complete, the trained classifier can be used on new articles to determine whether they are relevant and should be passed on to humans for coding.

The classifier designed for MMAD follows standard approaches in automated text coding and relies on a set of frequent words and word combinations. Since the distri-

bution of irrelevant and relevant articles is extremely skewed (only 2% are relevant), existing off-the-shelf algorithms had to be modified (Croicu and Weidmann, 2015). The final result of Croicu and Weidmann's effort is a classifier that eliminates more than 60% of all irrelevant articles while retaining around 90% of the relevant articles when evaluated "out-of-sample" on countries that it was not trained on (for detailed results, see Croicu and Weidmann, 2015). Given the large number of irrelevant articles, this trade-off seems acceptable. Thus, articles obtained through simple keyword searches were pre-filtered using machine learning before being passed on to human coders for the final step of the coding. In the next section, we describe how this final step is implemented for the coding of the MMAD.

4.2.3 Separating Events and Reports

An event coding process translates raw information from media reports into a simplified, usually numeric representation of events that allows for large-N analyses of the coded cases.[3] We refer to this source information as "reports". Reports rarely come in a form that is convenient for coding. Instead, two steps need to be performed during the coding process: First, the relevant pieces of information need to be extracted from the report—for example, information about the location and time of an incident, the issue of a protest, and the number and type of participants. This is the information extraction problem introduced above. Second, once the relevant information has been determined across a set of reports, it needs to be aggregated into a set of events, which constitute the final dataset. This is the aggregation problem.

Figure 4.3 provides a stylized illustration of the coding process. For now, consider only the two dashed boxes on the left and right. The left box shows a set of two reports, which together constitute the source material for a coding project. For example, this can be a collection of news articles obtained from a news archive such as Lexis-Nexis. The box on the right shows the final dataset, which in our case is a set of individual events. Usually, data projects do not specify how exactly they get from the source material to the final dataset. While almost all datasets specify the type of information that needs to be available before an event is coded, the fact that both information extraction and aggregation are performed by the coder without precise guidelines leaves us with two problems: First, we do not know the precise formulation of certain types of information in a report. For example, was the location reported as a precise city, or as a village outside a city? What was the precise label the news report used for a group of protesters? Without transparent coding rules, it is essentially up to the coder to map a particular piece of information in a report to the

[3]This section uses material from Weidmann and Rød (2015).

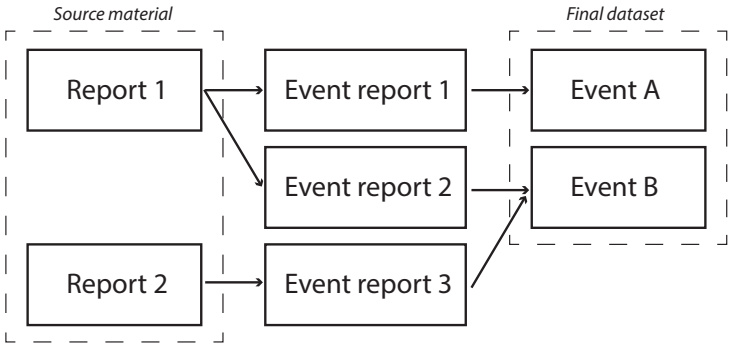

Figure 4.3: Reports, event reports, and events (from Weidmann and Rød, 2015). Reports can contain information about multiple events, thus generating multiple event reports (report 1 and event reports 1 and 2). Events, on the other hand, can be based on different event reports (event B and event reports 2 and 3).

corresponding entity (i.e. a location or group) and thus create the data the dataset is based on. A similar but perhaps even more severe problem arises when the coder aggregates different pieces of information. For example, if two reports are about the same event, but mention different numbers of protesters, which one is used in the final coding? How do we know that multiple sources, rather than one source, were used to code the event? So far, existing coding projects (human or automatic) have not tackled this problem directly, which can lead to serious problems (Jenkins and Maher, 2016).

In order to remedy these problems, we propose separating information extraction and aggregation into two steps. Essentially, the idea is to introduce an intermediate type of output from the coding process, called an "event report". As the name suggests, an event report is an individual statement of an event derived from a news report. It contains fields for all the relevant information needed to eventually code an event. Thus, an event report is the output of the information extraction step which serves as input for the aggregation step to generate the final list of events. Figure 4.3 illustrates this. From the source reports (left) we extract a set of event reports (center), which are later aggregated into the events that constitute the final dataset (right). While many reports will only contain information about a single event (for example, report 2 generates only one event report, no. 3), this procedure is able to deal with more complex reports: report 1 mentions two events, which result in event reports 1 and 2. Once we have generated the set of event reports, we need to aggregate them to obtain the final dataset. Again, many events will be based on only one event report, as is the case for event A in Figure 4.3. However, in cases where there are multiple event reports for one event, the coder will have to aggregate them into a single event. Since the extracted information is provided in a standardized form in the event record, this process can largely be automated,

thus making it extremely cheap and transparent. We will provide an example of this below.

How does this procedure solve the problems discussed above? First, it makes the information extraction step much more transparent. By using the event report(s) that an event is based on, a user can find out, for example, what phrase in the report was used to pinpoint the location of an event. This applies to other types of information as well, such as the number and type of protesters or the issue of the protest. Also, the user has full information about, and can even control, the aggregation process. For example, it is possible to change the way that participant numbers from the event reports are aggregated into a single number, or even to weigh information by source. Last, the event records can serve as training data for automatic text coding of event data. To date, these routines perform information extraction and aggregation in a single step, similar to human coding. This leads to exactly the same concerns described above, particularly regarding information aggregation. In contrast, using the intermediate stage of event reports for training computational classifiers can support new efforts to automate information extraction and the aggregation of these reports, and thus improve the transparency of automated coding techniques.

A short example from the Tulip Revolution in Kyrgyzstan helps to illustrate this. Table 4.1 displays nine event reports with divergent information on two variables included in the MMAD: the number of participants and the level of security force involvement (ordinal, values in parentheses).[4] The last column displays the news source. All of the event reports in Table 4.1 took place in the city of Osh on March 21, 2005. It is immediately apparent that the information in the event reports diverges both across reports from the same source and across sources: the number of participants differs in all three reports from the AP and all five reports from BBCM. Also, the estimate given by the AFP (`hundreds`) is very different from the AP estimates (`1,000` and `2,000`) and three of the BBCM estimates (`1,000`, `several thousand`, and `3,000`). In addition, there are two reports without participant number estimates. There is similar uncertainty regarding security force involvement in the protest. In fact, the information ranges from `not present` (0) to `physical intervention` (2). Four event reports indicate the presence or intervention of security forces, four do not mention security force involvement at all, and one asserts that forces were absent. Without transparent guidelines, it is not clear how different coders would have aggregated these event reports in a conventional event dataset.

Once we have extracted these event reports from the selected news reports, we need to aggregate them to the level of individual events. Table 4.2 shows this for our above example. For the sake of illustration, we employ two alternative aggregation rules.

[4] Other variables omitted for the sake of illustration.

Number of participants	Security force involvement	Source
hundreds		AFP
2,000	present (1)	AP
1,000		AP
	present (1)	AP
	present (1)	BBCM
1,000	not present (0)	BBCM
several thousand		BBCM
3,000	physical intervention (2)	BBCM
200		BBCM

Table 4.1: Event reports for Osh (Kyrgyzstan), March 21, 2005.

The first uses the *average* number of participants across all news reports (1,440) and the *most frequent value* for security force involvement (present, 1).[5] However, users who prefer other aggregations can do so easily, as the second line shows. Here, we use the *maximum* reported number of protesters (3,000), and the *maximum* level of security force involvement (physical intervention, 2). Of course, other aggregation rules are possible and can easily be applied by the user. For instance, one could compute confidence intervals around the aggregated numbers. In addition, one can focus on other variables in the aggregation process. For example, one could give preference to more recent reports by using the date and time a report was released (not shown in the table).

Aggregation rule	Number of participants	Security force involvement
average(#part), mode(sec. force inv.)	1,440	present (1)
max(#part), max(sec. force inv.)	3,000	physical intervention (2)

Table 4.2: Alternative event codings for Osh (March 21, 2005), according to two different aggregation rules.

In short, our procedure adds a new type of output to the coding process: the list of event reports. For the MMAD, these event reports are distributed alongside the finished list of events, which allows users to explicitly incorporate any uncertainty in the news reports into their analyses (Cook and Weidmann, 2017) or to study patterns of higher or lower reporting (Hellmeier, Weidmann and Geelmuyden Rød, 2018). However, before generating a final set of protest events, the different reports need to be aggregated, as demonstrated in the example above. Before introducing the design of our analysis in the next chapter, we briefly describe the scope and variables

[5]For the sake of illustration, the reported average number of participants omits the verbally specified numbers (`hundreds` and `several thousand`).

contained in the MMAD and give a few examples of protest episodes covered in our data.

4.3 The Mass Mobilization in Autocracies Database

In this section, we give a short overview of the database, starting with its scope: autocratic regimes.

4.3.1 Defining Autocracy

How can we tell democracies and autocracies apart? Political scientists agree that, at a minimum, democracies must fill executive and legislative offices through competitive elections (Przeworski et al., 2000*a*). Beyond this, however, there are different definitions, some arguing that the distinction between democracy and autocracy is gradual (Diamond, 2002; Freedom House, 2015; Marshall, Gurr and Jaggers, 2014), while some conceive of them as distinct categories (Przeworski et al., 2000*a*; Cheibub, Gandhi and Vreeland, 2010; Hadenius and Teorell, 2007). We follow one of the most popular latter approaches: Geddes, Wright and Frantz (2014*a*) considers a regime to be autocratic if the government came to power either (i) by using means other than a direct, reasonably fair competitive election, or (ii) through a competitive election but changed the rules while in office to prevent future elections from being competitive.

Regimes of the first type seize power through military coups or popular uprisings, for example. An example of the second type of regime is Chavez's presidency in Venezuela, where a democratically elected government prevented future elections from being competitive. This, of course, raises the question: under which circumstances does an election fail to meet a reasonable level of fairness and competitiveness? According to Geddes, Wright and Frantz (2014*a*), this is the case if large opposition parties are not allowed to participate, there is extensive repression of opposition leaders or supporters, or vote fraud alters the outcome of the election (see Geddes, Wright and Frantz, 2014*b*, 6 for details). Importantly—and this is where the definition deviates from many others—this excludes cases where no government exists or the government does not control the territory (e.g. Somalia since the end of Siad Barre's regime in 1991) and countries occupied by a foreign power (e.g. Iraq after the U.S. invasion from 2003 to 2005).

The Geddes, Wright and Frantz (2014*a*) regime classification identifies countries as autocracies for certain time periods. In other words, countries that used to be democratic can later become autocracies (such as Peru in 1992), but autocracies

can also drop out of the autocratic sample by transitioning to democracy (such as Portugal in 1974 and Mexico in 2000). Our coding project and the analysis follow this definition and include those countries that are considered autocratic during the relevant time periods. In its initial version used for this book, the dataset covers the years 2003-2012, although more recent years will be added in the future. The MMAD includes data from 70 different countries, whereby some are covered only for a certain part of the coding period. The chapter appendix shows the corresponding list of countries/time periods included in the dataset.

4.3.2 Political Protest

Having defined the type of political regime to be included in the MMAD, we now need to take a closer look at political protest, the object of study of this book. The public discourse about protest in autocratic countries is often dominated by more recent events such as mass protests in the Middle East during the Arab Spring or the anti-regime demonstrations against President Maduro in Venezuela. However, these dramatic events conceal the fact that political protest occurs much more frequently, although often on a smaller scale. So, what constitutes political protest? In short, we focus on overt events that (i) are directed against the government, (ii) involve a large number of people, (iii) take place in a public space, and (iv) do not explicitly aim to use violence. Let us take a closer look at the four elements of this definition.

Anti-regime. Political protests can be directed at different political actors or institutions. Here, we will deal with the most frequent and—for our purpose—most relevant type of protests: those that address the government ("anti-regime"). If successful, this type of protest can seriously destabilize and even topple autocratic regimes. In its typical form, anti-regime protest is associated with maximalist demands such as replacing the government, but this does not apply in all cases. In fact, we frequently see anti-regime mass protests with much more specific demands, such as lower wages for public officials, the release of arrested opposition members, or the reversal of commodity price increases. As a consequence, these demands are not necessarily directed toward the domestic central government; rather, they can address regional or even local governments and governmental institutions.

A large number of people. By definition, political protest is a collective endeavor by a large number of people. In this book, we therefore analyze protest events involving at least 25 people. This excludes a number of dissident activities carried out by single individuals or small groups. The power of large numbers of participants

derives from the fact that large protests are more visible and are thus more likely to attract the attention of national and international audiences, and therefore also the government. While more powerful, larger protests are also more difficult to coordinate, which is one of the reasons why modern communication technology is believed to have a potentially strong impact here.

A public gathering. Protest is a public activity. This means that it excludes acts of dissent that are carried out in private, such as "blackouts" where activists turn off the electricity in their homes, or when activists display flags or paint their homes in political protest (Schock, 2005). These activities take many forms and are carried out frequently, but are less likely to receive attention and thus have a smaller impact.

No systematic use of armed force. It is also important to distinguish protest from other, more violent forms of political contest, such as civil war. While both are similar in the sense that a political opposition confronts a government outside of regular political channels, civil wars by their nature are characterized by the systematic use of military force on both sides (Gleditsch et al., 2002). This is not the case for protest, which does not necessarily rely on armed force. Hence, our definition excludes armed dissent such as terrorist attacks and rebel violence. Importantly, however, we do not exclude events based on the level of violence; protests can turn violent, and can be violently repressed by the autocratic government.

By focusing on these four main criteria, we leave out others such as the level of organization. Protests differ in the extent to which they rely on existing political organizations, or rather emerge as spontaneous instances of collective action. In reality, however, many protests involve both organized groups and unorganized citizens, which is why we do not exclude one or the other. The MMAD contains variables for the date, location, actors, number of participants, level of violence by participants and by security forces. The full list of variables included in the database is available in the codebook that accompanies the dataset. A key feature of the database is the geo-referencing of protests to particular cities. During the coding process, coders assign the corresponding location based on the GeoNames database (`http://www.geonames.org`), a free gazetteer of geographic entities around the globe. This way, each event report is assigned a unique city from GeoNames, which later helps us to identify corresponding reports from other sources about the same city and on the same date.

4.3.3 Some Examples

The following three short examples help to illustrate the information contained in the MMAD: (i) the events leading up to the "Tulip Revolution" in Kyrgyzstan in 2005, (ii) the anti-regime protests before, during, and after the Iranian presidential elections in 2009, and (iii) the Arab Spring uprisings and their aftermath in Egypt in 2011. For these illustrations, reports in the MMAD were aggregated to events based on their reference to the same day and the same city. The examples highlight the spatial and temporal precision of the data, which is key for the disaggregated analysis of protest behavior we present in the remainder of this book.

Kyrgyzstan 2005 In the period after it gained independence from the Soviet Union, Kyrgyzstan was often hailed for its relatively liberal political environment (Anderson, 2013). Multiparty elections and economic reform created a stark contrast to the closed autocratic regimes in the neighboring states (e.g. Turkmenistan, Uzbekistan). In 1996, however, President Askar Akayev increased his power through constitutional amendments. Moreover, in the period following these changes, fraudulent election practices and persecution of political rivals ensured that Akayev and his loyalists remained in power (Freedom House, 2002). After the "Tulip Revolution" ousted Akayev in 2005, Kurmanbek Bakiyev came to power and maintained the nepotistic rule of his predecessor. Five year later, in 2010, Bakiyev was ousted in a new uprising after the killing of dozens of protesters led to a popular backlash. Since the ouster of Bakiyev, there has been little stability in Kyrgyzstan, as predatory elites rotate between being in office and being the opposition (Freedom House, 2016*b*).

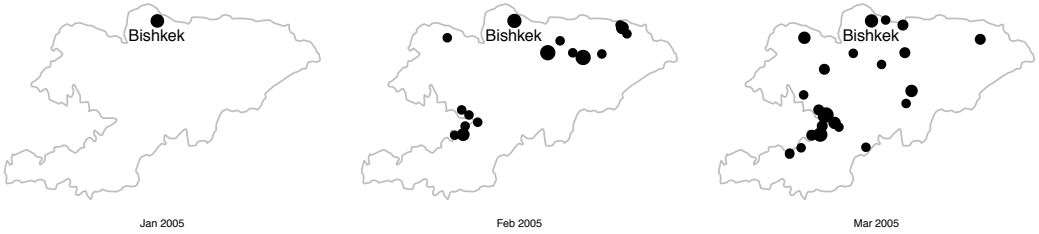

Figure 4.4: The "Tulip" Revolution (2005). January: 3 protests, 1 location. February: 26 protests, 16 locations. March: 78 protests, 24 locations.

Figure 4.4 maps the events leading up to the "Tulip Revolution" in Kyrgyzstan (Radnitz, 2006). The popular uprising is commonly referred to as one of the "Color Revolutions", a term used to describe successful non-violent mobilization against authoritarian regimes following disputed elections (Chenoweth and Stephan, 2011; Beissinger, 2007; Bunce and Wolchik, 2006; Way, 2008). Similar uprisings also overthrew incumbents in Serbia ("Bulldozer Revolution", 2000), Georgia ("Rose Revolu-

tion", 2003), and Ukraine ("Orange Revolution", 2004-2005). Like the other "Color Revolutions", the uprising in Kyrgyzstan was the result of stolen elections (Tucker, 2007). In February 2005, parliamentary elections marred by election fraud removed a number of influential local elites from power. As a result, these elites mobilized anti-government protests across the country (Temirkulov, 2010; Lewis, 2008). On March 24, demonstrators stormed the presidential palace and forced sitting President Akayev to flee the country.

On the maps, the black dots are anti-government demonstrations. The size of the dots corresponds to the number of events in each location. The left map shows that there was little activity in January, with protests restricted to the capital (Bishkek). In February (middle map), fraudulent parliamentary elections led to widespread protests. The MMAD records protests in 16 different cities in February. However, the number of events in each location was relatively low, peaking at four in Bokombayevskoye and Kochkor. In March (right map), political violence escalated as protesters stormed and occupied government buildings throughout the country (24 cities in total). The southern capitals of Osh and Jalal-Abad were the epicentres of protest (Temirkulov, 2008) with 12 and 17 recorded anti-government demonstrations, respectively. In fact, protests only spread to Bishkek after security forces raided an occupied building in Jalal-Abad and killed several protesters. Prominent opposition politicians, such as former Prime Minister Kurmanbek Bakiyev and founding member of the Fatherland Movement (Ata-Jurt) Roza Otunbayeva, joined the Bishkek protests and led the overthrow of President Akayev.

Iran 2009 Since the overthrow of Shah Mohammad Reza Pahlavi in 1979, Iran has been governed by religious leaders. The Supreme Leader Ali Khameini is the highest authority in the country, with the power to appoint and dismiss highly ranked members of the government, the judiciary, and the military. The office of Supreme Leader is not subjected to popular elections, and power has only been passed once in the 37 years since the regime was established when the first Supreme Leader, Ruhollah Khomeini, died in 1989. However, Iran holds presidential and legislative elections. Candidates for elected political offices must be approved by the Supreme Leader and other religious leaders (the Guardian Council and the Assembly of Experts). This approval ensures that no candidate can fundamentally oppose the regime, yet candidates commonly campaign under conservative or reformist agendas, and the resulting elections have brought about some policy implications. For example, under the period of reformist rule by Mohammad Khatami, restrictions on freedom of expression and gender separation were relaxed (Freedom House, 2010).

The 2009 presidential elections pitted the conservative incumbent Mahmoud Ah-

madinejad against reformist opposition candidate Mir-Hossein Mousavi. The pre-election phase was perceived to be rigged in favor of Ahmadinejad, in particular by attempting to control communication via cell phones and the Internet (CNN, 2009). Observers have argued that the election fraud was part of a larger crackdown on the reformists by the religious leadership (Freedom House, 2010). Official election results published on June 12, 2009 proclaimed victory for Ahmadinejad with 62% of the votes against Mousavi's 34%. The announcement of the results sparked large-scale protests that quickly turned violent. The protest movement, claimed to be the largest since the Iranian Revolution, is often referred to as the Green Movement. Green was originally the color of Mousavi's campaign and became a symbol of the resistance against the incumbent administration. A number of studies have identified Twitter as a catalyst of the protests (Grossman, 2009; Morozov, 2009).

Figure 4.5: Iranian presidential election protests (2009). May: 8 protests, 6 locations. June: 20 protests, 8 locations. July: 8 protests, 2 locations.

Figure 4.5 maps the MMAD events from May to July 2009 in Iran. Before the election in May (left map), the number of protests was quite low: eight events in six cities. In June (middle map), the month of the election, anti-regime protests spread, occurring in eight cities and totaling 20 events, with more than half of these (12) occurring in Tehran. Moreover, many protests featured violence from both protesters and government security forces. In July (right map), the protests subsided, and the data record eight events in two cities. This also marks a shift in the tactics employed by the Green Movement, who started boycotting goods, scribbling anti-regime slogans on banknotes, and marking properties of militiamen with the color green (Wright, 2009).

Egypt 2011 After the overthrow of King Fuad in 1952, Egypt was governed by military officers. Hosni Mubarak, the fourth officer in a row to become president of the republic, ruled for 30 years (1981-2011). Following a brief period of civilian rule after Mubarak's resignation, military officers regained power in 2013. Since then, repression of political activities and persecution of opposition have increased. For

example, the victors of the 2012 election—the Muslim Brotherhood—were declared a terrorist organization shortly after the coup. Similar action has also been taken against non-religious political rivals (Freedom House, 2016*a*).

In early January 2011, protests calling for regime change in Tunisia quickly spread to Egypt. A few weeks later, the ousting of Tunisian President Ben Ali intensified hopes of a successful overthrow in Egypt. The Egyptian protests were launched under an umbrella of grievances related to e.g. corruption, power abuse, unemployment, and fraudulent elections. Moreover, a number of observers have highlighted the role of social media for protest mobilization in Egypt (Lotan et al., 2011; Khondker, 2011), while others have downplayed its importance (Anderson, 2011). On January 25—the so-called "Day of Revolt"—thousands of people took to the streets in cities across the country. Protests persisted over the next weeks, prompting the government to impose a curfew and increase military presence in Cairo. In addition, political reforms were promised. On February 10, protests intensified once more when Mubarak stated his intent to stay in office. However, the day after it was announced that he had resigned from his post as president. During the upheavals, hundreds of people were killed and thousands were injured.

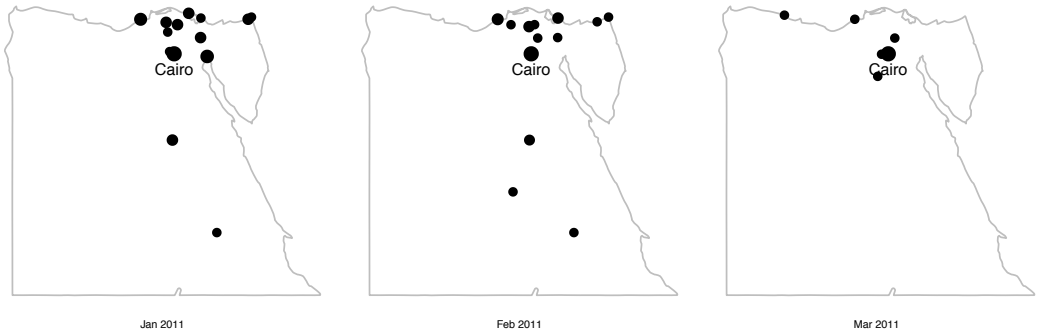

Figure 4.6: Egyptian uprising (2011). January: 35 protests, 14 locations. February: 38 protests, 14 locations. March: 19 protests, 7 locations.

Figure 4.6 plots the events in Egypt from January to March 2011. In January (left map), 35 protests in 14 different locations are recorded in the MMAD. 31 of these protests occurred on or after the "Day of Revolt" (January 25). Nine of the protests occurred in Cairo, five in Suez, and four in Alexandria. In February (middle map), the number of recorded protests and locations is very similar to January. However, protests in Cairo intensified. In total, 17 protests were recorded in the capital in February. After the overthrow of Mubarak, the number of protests subsided to about half. In March (right map), the MMAD records 19 protests in seven different locations, with most of these (13) occurring in Cairo.

4.4 Conclusion

We began this chapter with a review of the two main approaches for event coding: traditional coding by humans and computer-based automated coding through natural language processing. Each of them has its downsides: human coding is resource-intense and potentially less reliable, while automated coding is less precise and (at present) unable to capture the detailed information on protests that scholars require for their analysis. Regardless of whether we use human or automated coding, there are four challenges we need to overcome: the problem of selecting sources, the problem of selecting relevant reports from these sources, the problem of extracting information from the selected reports, and the problem of aggregating different reports to individual events.

The chapter described in detail how the MMAD project addresses these challenges: a set of three news agencies were selected as sources, since together they cover around 85% of all events reported. Articles from these sources were filtered using a machine learning classifier that separates relevant from irrelevant articles. The articles classified as relevant were then processed by human coders. During the coding process, the coders extracted reported information on specific characteristics of the protest incidents from each article that covers a political protest. Rather than aggregating these different reports about a single event, the database retains the individual reports. Not only does this make the event coding process much more transparent and lead to a more comprehensive database for this book, it also allows for a number of new applications of the data. For example, Cook and Weidmann (2017) show that the use of report-level data leads to better and more accurate results in quantitative analyses of protest compared to the use of event-level data. Hellmeier, Weidmann and Geelmuyden Rød (2018) used the MMAD to study patterns of media attention, which they measure as the number of reports per event. Thus, the more complex coding procedure and more involved use of the dataset add considerable scientific value.

Chapter Appendix

List of Variables in the MMAD

For more details about the coding of these variables please refer to the MMAD Codebook (see `http://mmadatabase.org/`).

- Event date (type: date): Date of incident.

- Location (type: string): City of incident, according to the GeoNames database of place names (`http://www.geonames.org`).

- Actors (type: string): Actors involved in the protest incident. In a general sense, this string variable captures the label given by the article to the actors involved. If more than one actor is given in the article, we separate them with a semicolon.

- Number of participants (type: string): Estimate of the number of participants. Can be either an *integer number* or a *phrase*.

- Issue (type: string): Reported issue / motivation for incident. Described by one or two terms using the original wording in the event report. More than one issue can be reported for each incident, separated with a semicolon.

- Side (type: dichotomous): Takes the value 1 if incident was anti-government. Anti-government is understood in a broad sense. It is not necessary that protesters demand the resignation of the central government, but that they are protesting actions made or sanctioned by it. This includes national, regional, and local authorities' actions, since the hiring and firing of state employees at all levels of government rests on the people in charge. If 0, the protest is explicitly pro-government, staged to show support for the government or the government's actions. If NA, protest was directed at a domestic public or private non-governmental institution.

- Scope (type: categorical): Indicates which level the protest is directed at. 0 = national, 1 = regional / state, 2 = local. If NA, protest was directed at a *domestic* public or private institution. The assumption here is that local and regional protest is also anti- or pro-government.

- Level of violence by protest participants (type: ordinal): Ordinal level of violence from protest participants. NA = no report on the level of violence from protest participants, 0 = explicit report of no violence, 1 = reports of property damage or clashes with civilians or security forces, 2 = reports of people injured, 3 = reports of people killed. Protesters blocking roads or railroads do not qualify as exerting violence, unless there are explicit reports that participants were damaging cars, equipment, or exerting physical violence against bystanders. Moreover, self-immolation does not count as violence because the action is not directed at other people.

- Level of official security forces engagement (type: ordinal): Ordinal level of official security force involvement. NA = no report on the level of official security force involvement, 0 = explicit report of no presence, 1 = reports of

presence, 2 = reports of physical intervention. Includes crowd dispersal, arrests, and beatings but excludes lethal intervention, 3 = reports of lethal intervention.

List of Country-periods in the MMAD

Country	Start	End
Afghanistan	2009-08-20	2012-12-31
Algeria	2003-01-01	2012-12-31
Angola	2003-01-01	2012-12-31
Armenia	2003-01-01	2012-12-31
Azerbaijan	2003-01-01	2012-12-31
Bangladesh	2007-01-11	2008-12-29
Belarus	2003-01-01	2012-12-31
Botswana	2003-01-01	2012-12-31
Burkina Faso	2003-01-01	2012-12-31
Burundi	2003-01-01	2003-04-30
Cambodia	2003-01-01	2012-12-31
Cameroon	2003-01-01	2012-12-31
Central African Republic	2003-03-15	2012-12-31
Chad	2003-01-01	2012-12-31
China	2003-01-01	2012-12-31
Congo	2003-01-01	2012-12-31
Cuba	2003-01-01	2012-12-31
Democratic Republic of the Congo	2003-01-01	2012-12-31
Egypt	2003-01-01	2012-06-30
Eritrea	2003-01-01	2012-12-31
Ethiopia	2003-01-01	2012-12-31
Gabon	2003-01-01	2012-12-31
Gambia	2003-01-01	2012-12-31
Georgia	2003-01-01	2003-11-23
Guinea	2003-01-01	2010-01-16
Guinea-Bissau	2003-01-01	2003-09-14
Haiti	2003-01-01	2004-02-29
Iran	2003-01-01	2012-12-31
Ivory Coast	2003-01-01	2012-12-31
Jordan	2003-01-01	2012-12-31
Kazakhstan	2003-01-01	2012-12-31
Kuwait	2003-01-01	2012-12-31
Kyrgyzstan	2003-01-01	2012-12-31

Laos	2003-01-01	2012-12-31
Liberia	2003-01-01	2003-08-11
Libya	2003-01-01	2012-12-31
Madagascar	2009-03-17	2012-12-31
Malaysia	2003-01-01	2012-12-31
Mauritania	2003-01-01	2012-12-31
Morocco	2003-01-01	2012-12-31
Mozambique	2003-01-01	2012-12-31
Myanmar	2003-01-01	2012-12-31
Namibia	2003-01-01	2012-12-31
Nepal	2003-01-01	2006-04-24
North Korea	2003-01-01	2012-12-31
Oman	2003-01-01	2012-12-31
Pakistan	2003-01-01	2008-08-18
Russia	2003-01-01	2012-12-31
Rwanda	2003-01-01	2012-12-31
Saudi Arabia	2003-01-01	2012-12-31
Singapore	2003-01-01	2012-12-31
South Sudan	2011-07-09	2012-12-31
Sudan	2003-01-01	2012-12-31
Swaziland	2003-01-01	2012-12-31
Syria	2003-01-01	2012-12-31
Tajikistan	2003-01-01	2012-12-31
Tanzania	2003-01-01	2012-12-31
Thailand	2006-09-19	2007-12-23
Togo	2003-01-01	2012-12-31
Tunisia	2003-01-01	2012-12-31
Turkmenistan	2003-01-01	2012-12-31
Uganda	2003-01-01	2012-12-31
United Arab Emirates	2003-01-01	2012-12-31
Uzbekistan	2003-01-01	2012-12-31
Venezuela	2005-12-04	2012-12-31
Vietnam	2003-01-01	2012-12-31
Yemen	2003-01-01	2012-12-31
Zambia	2003-01-01	2012-12-31
Zimbabwe	2003-01-01	2012-12-31

Table 4.3: Country-periods in the MMAD