Coverage Biases in Off-the-shelf Protest Datasets

Killian Clarke
Georgetown University
Killian.Clarke@georgetown.edu

The revolutionary movements that comprised the 2011 Arab Spring catalyzed a major new research agenda in Middle East political science, which sought to make sense of the drivers, contours, and dynamics of protest across the region. The methodologies embraced in this scholarship have been diverse and eclectic, but one of the most widely used approaches has been to analyze protest event datasets. As I explain in a recent article in *Mediterranean Politics*, which forms the basis for this short essay, these datasets are inventories of contentious events that meet certain criteria and that occur in a delimited time and place.[1] The researcher collects information about each event (e.g., the timing, location, size, demands, and participants) and then examines variation across these dimensions in an effort to understand the features and characteristics of a particular mobilizational wave. Though event analysis had been used occasionally by MENA-focused scholars before the Arab Spring,[2] the technique became much more popular after the uprisings, deployed to study mobilization in places as diverse as Tunisia, Algeria, Morocco, Egypt, Lebanon, Syria, and Iraq.

One of the thorniest challenges in collecting and analyzing event data is selecting which sources will be consulted to build the dataset. Newspapers are the most commonly used source, with researchers relying on reporters' descriptions of protests in their news articles to collect and code their data. Some datasets are also built using social media data, human rights reports, and/or government archives. However, no combination of sources, no matter how exhaustive or comprehensive, reports on *every* event that occurs, leaving researchers to reckon with the fact that their event datasets are inevitably a selective sample of the broader mobilizational whole. If the sample of events in the dataset is somehow not representative of that broader whole – because the sources used are reporting on certain types of events more than others – then the researcher runs the risk of drawing incorrect conclusions.[3] For example, if the researcher is using a newspaper that reports on large, violent events more than small, non-violent ones (which are potentially considered less 'newsworthy') then the resulting event dataset will paint a picture of an uprising or a movement that is far more violent and explosive than it may have actually been.

[1] Killian Clarke, "Which Protests Count? Coverage Bias in Middle East Event Datasets," *Mediterranean Politics*, online first, 2021.

[2] For example: Joel Beinin, "A Workers' Social Movement on the Margin of the Global Neoliberal Order, Egypt 2004–2009," in *Social Movements, Mobilization, and Contestation in the Middle East and North Africa*, ed. Joel Beinin and Frederic Vairel (Stanford: Stanford University Press, 2011), 181–201; Adria K. Lawrence, *Imperial Rule and the Politics of Nationalism: Anti-Colonial Protest in the French Empire* (New York: Cambridge University Press, 2013).

[3] For more on coverage biases in event datasets see: John D. McCarthy, Clark McPhail, and Jackie Smith, "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991," *American Sociological Review* 61, no. 3 (1996): 478–99; Pamela E. Oliver and Daniel J. Myers, "How Events Enter the Public Sphere: Conflict, Location, and Sponsorship in Local Newspaper Coverage of Public Events," *American Journal of Sociology* 105, no. 1 (1999): 38–87; Jennifer Earl et al., "The Use of Newspaper Data in the Study of Collective Action," *Annual Review of Sociology* 30 (2004): 65–80.

The challenges of sourcing in event analysis are particularly acute for scholars doing research on a non-English speaking part of the world. In the aftermath of the Arab Spring, many Middle East political scientists insisted that the only event data that would even come close to approximating the true patterns of mobilization on the ground would have to be sourced from Arabic-language newspapers. The alternative was to use MENA event datasets that had already been built as part of broader data collection projects – e.g., the Armed Conflict Location & Event Data Project (ACLED), the Social Conflict Analysis Database (SCAD), or the GDELT Project.[4] But these datasets were primarily based on English-language sources, like the international wire services AP, AFP, and Reuters, and many Middle East scholars argued that these sources were likely to have serious biases in which protests covered. Though it would require significant investment of time and resources, they insisted that instead scholars should build and analyze *original* event datasets using local newspapers that write and report in Arabic.

In this short essay I evaluate these claims by comparing one locally-sourced event dataset focused on Egypt to two off-the-shelf datasets that rely primarily on English language sources (ACLED and SCAD).[5] The datasets cover a particularly eventful period of Egyptian history – the eighteen months directly preceding Abdel Fattah al-Sisi's counterrevolutionary coup in July 2013, when Egypt was awash with protest and unrest. The locally-sourced data come from the major Egyptian daily newspaper *al-Masry al-Youm*.[6] The comparisons reveal not only that the off-the-shelf datasets contain far fewer events than the locally-sourced one, but also that their datasets appear to be biased in the types of events they include: they tend to capture a larger proportion of events during more intense political periods, and they overcount large, urban, violent, and political events.

We can begin by comparing raw protest counts in the three datasets. I consider all contentious events that occurred in Egypt from January 1, 2012 to July 3, 2013.[7] My dataset, based on *al-Masry al-Youm*, captures 7,522 events that meet this description. SCAD uses the wire services AFP and AP to source its data, and identifies 593 events over the same period. ACLED uses a broader range of sources, including wires as well as international news websites like the BBC and Egyptian English-language newspapers. Its dataset includes 1,014 contentious events over this period.[8] These numbers imply that off-the-shelf datasets are capturing somewhere between 13% and 8% of the events identified in a single local-language source. To put this in perspective, the SCAD researchers have

[4] Clionadh Raleigh et al., "Introducing ACLED: An Armed Conflict Location and Event Dataset," *Journal of Peace Research* 47, no. 5 (2010): 651–60; Idean Salehyan et al., "Social Conflict in Africa: A New Database," *International Interactions* 38, no. 4 (September 1, 2012): 503–11.

[5] The analyses are derived from Clarke 2021.

[6] For more on the data collection strategy see Clarke 2021.

[7] A contentious event was defined as a public, collective, and voluntary endeavor involving a group of people in a specific place trying to influence the actions or policies of some authority. It includes protests, demonstrations, strikes, marches, sit-ins or occupations, roadblocks or blockades, boycotts, petitions, and mass attacks. This definition is draw from: Doug McAdam, Sidney Tarrow, and Charles Tilly, *Dynamics of Contention* (New York: Cambridge University Press, 2001).

[8] For more information about how I operationalized contentious events in each dataset, and on how I constructed the variables in the analyses below, see Clarke 2021.

argued, based on a different methodology, that their dataset covers 76% of all the events that occur in Africa.[9]

Figure 1: Monthly Event Counts (Author data, ACLED, and SCAD), Jan 2012 – Jun 2013

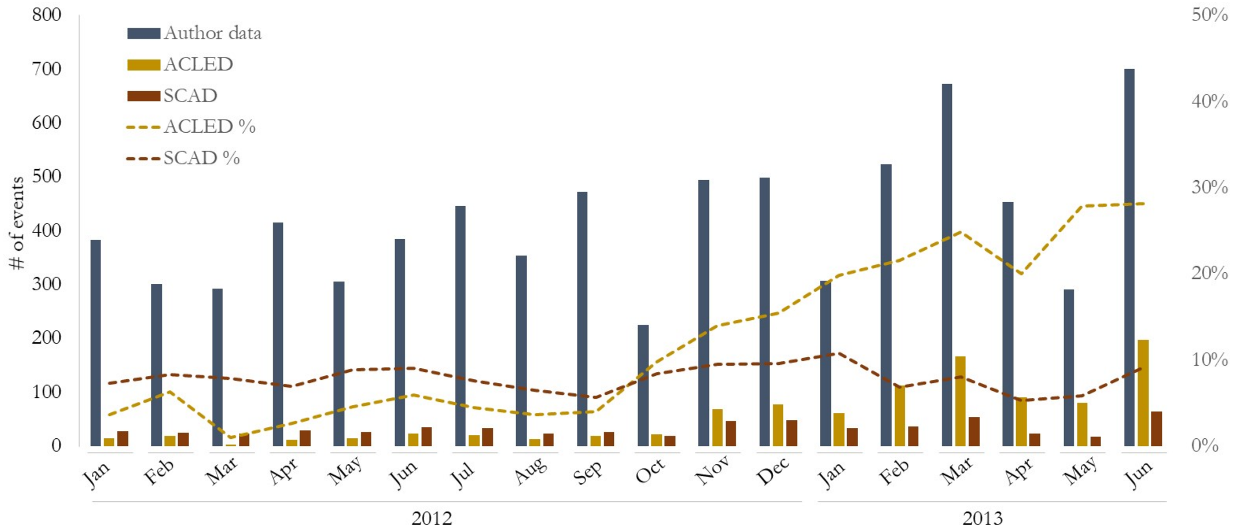
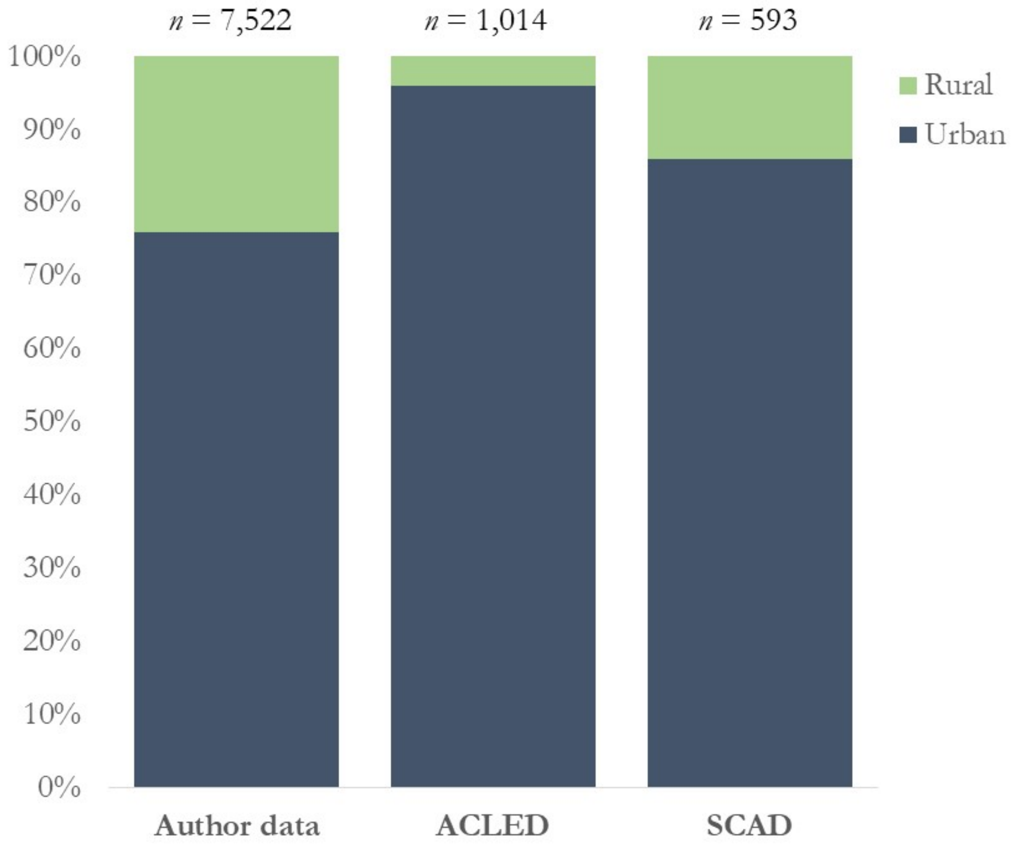


Figure 1 shows the monthly count of events for all three datasets. The figure also plots two lines representing the percent of monthly events in SCAD and ACLED compared to the monthly counts in my dataset. For SCAD the share ranges from 5% to 11%, and for ACLED it ranges from 3% to 28%. We also see in these trend lines signs of certain temporal biases. The ACLED dataset, especially, begins to capture a far higher percent of events after October 2012 and through the first half of 2013. This trend coincides with an increase in national news attention on Egypt, as the government headed by President Mohamed Morsi found itself engulfed in crisis and the counterrevolutionary movement to oust him gained momentum. The figure suggests that during such periods of heightened political tension and increased international scrutiny, ACLED's reliance on English-language and international news sources may result in it overstating the degree of unrest and contention in a country.[10]

[9] Cullen S. Hendrix and Idean Salehyan, "No News Is Good News: Mark and Recapture for Event Data When Reporting Probabilities Are Less Than One," *International Interactions* 41, no. 2 (March 15, 2015): 392–406.

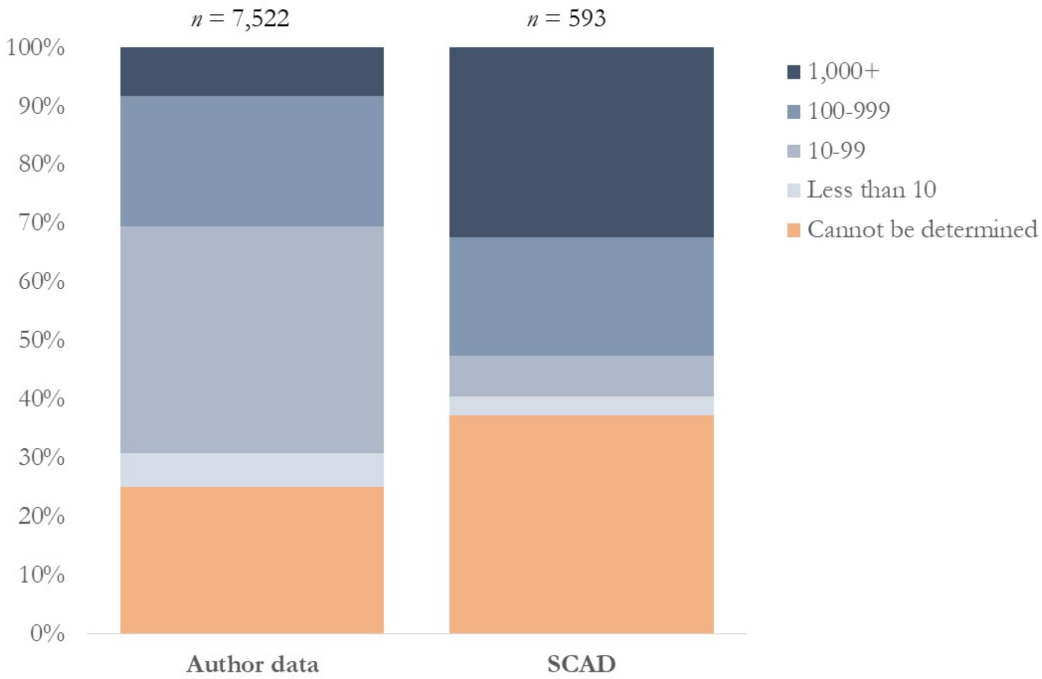
[10] Interestingly, SCAD's data to not exhibit the same temporal biases. This may be because it relies on a more standard set of sources, whereas ACLED's source base fluctuates according to the temporal period.

Figure 2: Urban Share of Events (Author data, ACLED, and SCAD)



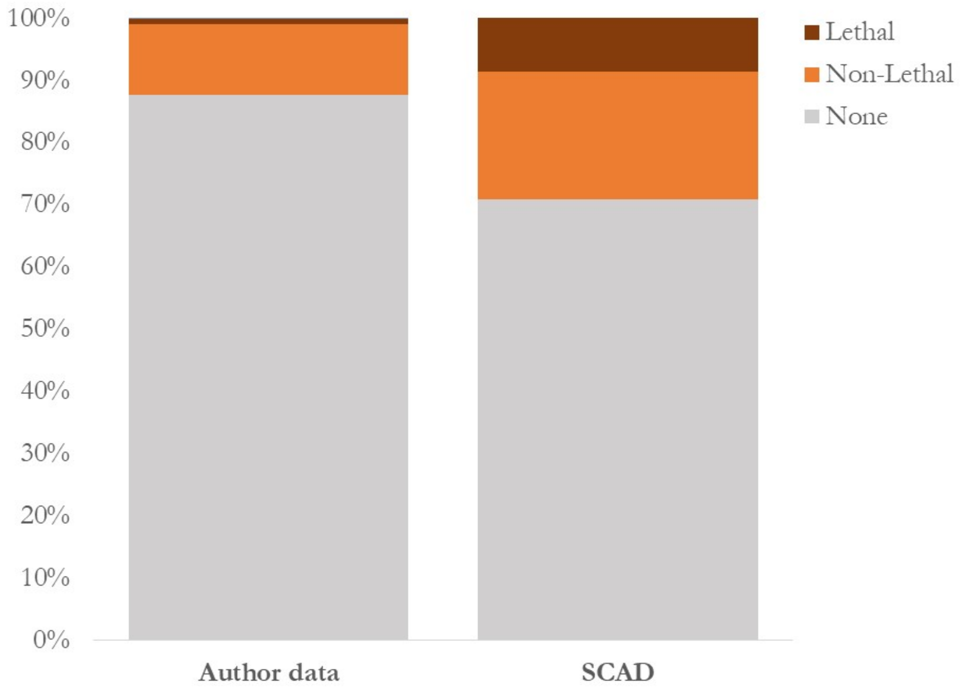
Next, I examine the distribution of events in the three datasets according to various protest characteristics: location, size, violence, and demands. I look first at the location of events, as scholars have found that newspapers and wire services are often biased in their coverage toward events that occur in cities. Figure 2 shows the distribution of events in the three datasets that occurred in urban versus rural locations. All three datasets include a large share of urban events, which partly reflects the simple fact that protests often occur in cities. However, whereas 24% of events in my dataset occur in rural locations, rural events make up only 4% of ACLED's data and 14% of SCAD's data, suggesting that both datasets may be over-counting urban events.

Figure 3: Distribution of Events by Number of Participants (Author data and SCAD)



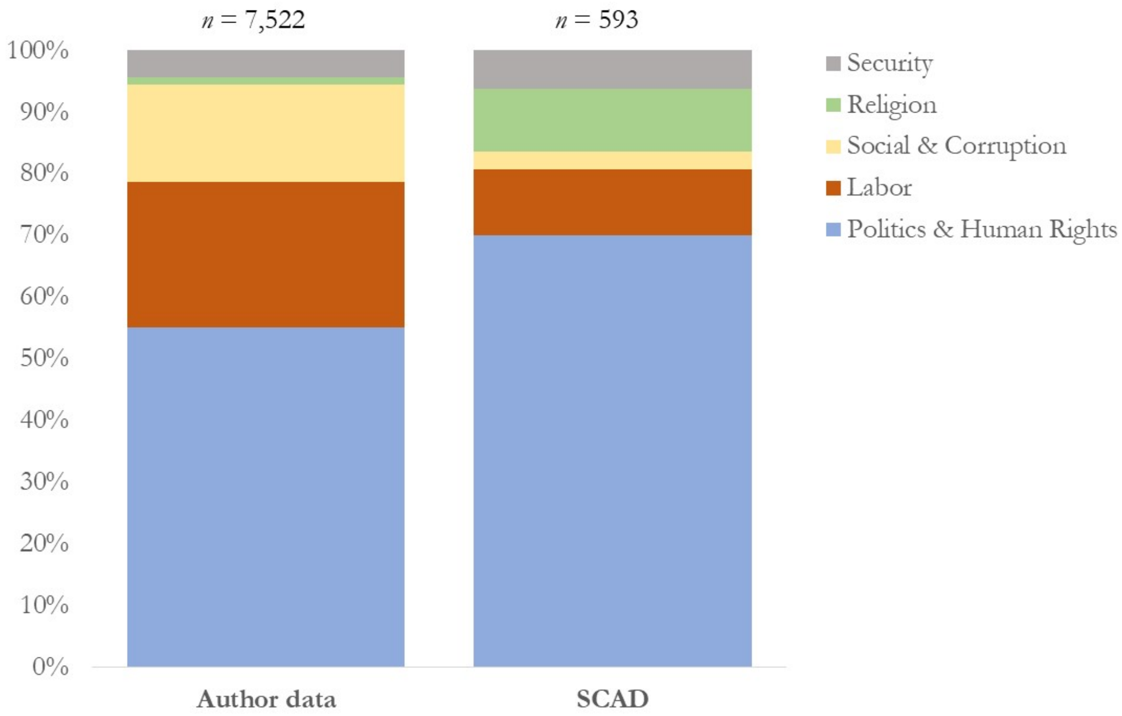
The next three figures compare my data only to SCAD, which includes a larger number of event-level variables than ACLED's data (e.g., size, repression, and demands). Figure 3 compares the distribution of events in the two datasets according to the number of participants, since larger events tend to receive more coverage in newspapers than smaller ones. Events were divided into five categories: those of more than 1,000 participants, 100-999 participants, 10-99 participants, less than 10 participants, and events where the reporting did not indicate a participation number. The figure reveals that SCAD's use of wire services for its sourcing does indeed result in a bias toward larger events: only 10% of its events include less than 100 participants, versus 44% in my dataset. In addition, SCAD has a higher share of events (37% versus 25% in my dataset) for which the number of participants was not reported, presumably because wire services include less rich and detailed information on protests.

Figure 4: Distribution of Events by Repression Level (Author data and SCAD)



Reporters are more likely to report on protests that involve more violence, which are generally regarded as more newsworthy than nonviolent protests. Figure 4 reveals that, indeed, SCAD's data is biased toward violent events – specifically, events that were repressed. The figure plots the distribution of events according to the level of repression they encountered: lethal repression, non-lethal repression, or no repression. The SCAD dataset disproportionately includes events that experienced repression (29% versus 12% in my dataset), especially those involving lethal repression (9% versus 1% in my dataset).

Figure 5: Distribution of Events by Demand Type (Author data and SCAD)



Finally, I compare the distribution of events according to the main demand that was raised. Though there are some issues of commensurability across the two datasets, I am able to group demands into five broad categories: politics & human rights, labor, social & corruption, religion, and security. As Figure 5 reveals, SCAD disproportionately includes events with demands related to religion and politics & human rights, and it tends to undercount events involving labor or social demands. These biases are explicable based on what we know about the reporting priorities of international wire services, which write for foreign audiences that are likely to be more interested in political, human rights, and religious issues than in labor strikes or social protests over issues like electricity provision, education, and corruption.

These findings have important implications for scholars interested in using event data to study protest in the Middle East. Particularly when doing sub-national analysis or examining the contours and dynamics of a single uprising, movement, or mobilizational cycle relying on event datasets whose sources are systematically biased is likely to lead to incorrect conclusions and a skewed representation of reality. For example, some have argued that research on the Arab Spring has devoted too much attention to the spectacular displays of protest in large public squares like Tahrir in Cairo, at the expense of less well-covered but equally important manifestations of contention in smaller cities like Suez and Port Said, industrial factories and workplaces, and rural settings.[11] Use of protest data that are themselves skewed toward large cities and political events is only likely to exacerbate such problems.

[11] For example: Jillian Schwedler, "Comparative Politics and the Arab Uprisings," *Middle East Law and Governance* 7, no. 1 (April 23, 2015): 141–52.

Further, in the article-length version of this essay I show that use of different datasets may lead to disparate and irreconcilable conclusions in statistical analyses.[12] In a simple pair of regressions modeling the determinants of protest repression I find different results depending on which dataset is used; SCAD's data would lead us to the conclusion that protests in small cities outside Cairo are most likely to be repressed, whereas my dataset suggests that it is protests in Cairo that are most repressed. These findings make sense when we consider that SCAD's wire sources are likely only reporting on events outside of Cairo when they are particularly violent and intense.

Ultimately, then, these biases have real implications for the kinds of conclusions we are able to draw about protest in the region. While off-the-shelf datasets may still be helpful for studying protest in a broader comparative setting – e.g., comparing protest waves across multiple countries or looking at trends over many years or decades – for more fine-grained, within-case analyses scholars are better off relying on datasets that use local, Arabic-language sources.

[12] Clarke 2021.