COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 4, Pages 104–115
https://doi.org/10.17645/mac.v9i4.4217

Article

Protest Event Analysis Under Conditions of Limited Press Freedom: Comparing Data Sources

Jan Matti Dollbaum [1,2]

[1] SOCIUM Research Center on Inequality and Social Policy, University of Bremen, Germany;
E-Mail: dollbaum@uni-bremen.de
[2] Research Center for East European Studies at the University of Bremen, Germany

Submitted: 25 February 2021 | Accepted: 28 May 2021 | Published: 21 October 2021

Abstract
The investigation of long-term trends in contentious politics relies heavily on protest event analysis based on newspaper reports. This tends to be problematic in restricted media environments. To mitigate the effects of bias and (self-)censorship, researchers of protest in authoritarian regimes have experimented with other sources such as international media and dissident websites. However, even though classical news media are easier targets for repression, journalistic reports might still outperform other sources regarding the quality of information provided. Although these advantages and disadvantages are known in the literature, different types of sources have seldom been tested against each other in an authoritarian context. Using the example of Russia between 2007 and 2012, the present article systematically compares protest event data from English-language news agencies, dissident websites, and several local sources, first and foremost with a view to improving methodological knowledge. The analysis addresses broad trends across time and space as well as the coverage of specific regions and single protest events. It finds that although the data sources paint different pictures of protest in Russia, this divergence is systematic and can be put to productive use. The article closes with a discussion on how its findings can be applied in other contexts.

Keywords
authoritarian regimes; media freedom; opposition; protest event analysis; Russia

Issue
This article is part of the issue "Media Control Revisited: Challenges, Bottom-Up Resistance and Agency in the Digital Age" edited by Olga Dovbysh (University of Helsinki, Finland) and Esther Somfalvy (Research Centre for East European Studies at the University of Bremen, Germany).

1. Introduction

Protest event analysis is one of the most important and widespread methods to investigate social movements and protest cycles on a large scale; however, the insights drawn from it can only be as good as the data on which it is based. Traditionally, protest event analysis relies on journalistic reporting, which is usually reliable (if trustworthy sources are selected), but it nevertheless comes with certain biases concerning the selection of events and the reported details (Earl et al., 2004; Gladun, 2020).

In political environments where journalists cannot operate freely, the usefulness of protest event analysis data from news sources is questionable. At least, researchers must take additional precautions to circumvent or mitigate the additional biases that are introduced through (self-)censorship. Some protest data sets, therefore, rely only on international news agencies (Weidmann & Rød, 2019). This approach, however, has the disadvantage that it captures only the internationally visible fraction of the protest landscape, which may not be representative of protest as a whole. Other approaches include: (1) selecting as many different local sources as possible to minimize bias of individual sources—the so-called "blanketing strategy" (Beissinger, 2002); (2) using social media data (Zhang & Pan, 2019a); and (3) relying on dissident websites (Lankina, 2015; Robertson, 2013). While all of them have their own

advantages, they also have drawbacks: The blanketing strategy is highly resource intensive, social media posts often do not contain easily accessible information on important details such as topic and size (Zhang & Pan, 2019a), and activist data can be politically biased.

In this article, I undertake one of the first attempts to systematically compare different sources of protest event data, addressing broad trends over time and space, their overlap, and their thematic coverage. On the example of protest in Russia between 2007 and 2012, I compare data sourced from international news agencies (Weidmann & Rød, 2019), data extracted from dissident websites (Lankina, 2018; Robertson, 2013), and data from various local sources, including journalism and official accounts (Semenov, 2017). This is not an effort to expose flaws and biases in the different data sets. Instead, it is an attempt to better understand whether such biases systematically relate to the sources and structure of data sets, and to examine which data source is best for answering which type of research question.

Russia lends itself well to such an undertaking: First, in the period studied, it represented a paradigmatic case of a modern hybrid regime that combined democratic elements such as multi-party elections with obstruction, targeted repression, and manipulation—skewing the institutional playing field toward the political leadership (Levitsky & Way, 2010). This hybridity included the media sphere: While there was no official censorship, there were highly visible cases of political pressure through ownership changes and repression of journalists (McFaul & Stoner-Weiss, 2008). Reporters thus needed (and still need) to navigate a complex set of intentionally vague, unwritten rules which may result in self-censorship on sensitive topics such as protest. Moreover, there was considerable subnational variation of repression against activists and media (Dollbaum, 2020a; Petrov & Titkov, 2013). The Russian case can thus also be helpful to derive more general hypotheses on the interplay of press freedom, political activism, and the quality of protest event data.

Overall, the analysis finds that the data sources differ from each other in the picture they paint of protest in Russia. This divergence is, however, systematic, meaning that it can be put to productive use by matching one's data with the research question: While international media are best in capturing large protest waves, activist-based data quite consistently document regional trends. Finally, for case study research that focuses on single events, there seems to be no way around utilizing a diverse set of local sources.

2. The Evolution of Protest Event Analysis

Protest event analysis as a method evolved over several decades. Hutter (2014) identified four "generations" in this evolution: The first comprises early works by, for instance, Tilly et al. (1975) that established the method of using reports on protest events (usually newspapers) to map popular contention across time and space. Fueled by nascent debates on source selection, a second generation (Kriesi et al., 1995) paid greater attention to sources and, in addition, increasingly used the technique for cross-national comparisons. A third generation then more systematically addressed selection bias within newspaper sources themselves. Predictors for newspaper coverage include the number of protestors, the degree of disruption and violence, the presence of counter-demonstrations, and police involvement (Earl et al., 2004). Others have found that other additional factors such as the timing in a legislative process (Oliver & Maney, 2000), the sponsorship of a protest by established social movement organizations (McCarthy et al., 2008), and media attention to similar events (Hellmeier et al., 2018) also make coverage more likely. Similar distortions apply to television coverage (Wouters, 2013). A fourth phase then went beyond single protest events or aggregated protest cycles (Tarrow, 1993) as the unit of analysis, instead analyzing "contentious episodes" (Kriesi et al., 2019). Stages three and four also saw the increased use of machine learning to classify relevant reports (Bremer et al., 2020) or to automatically annotate whole articles (Hanna, 2017), a technique that has the potential to greatly reduce the resources necessary to conduct protest event analysis.

3. Protest Event Analysis in Unfree Media Environments

Problems concerning the sampling and the quality of information in the source material are exacerbated when the context is characterized by limited freedom of the press. The reported studies refer to democratic contexts where bias usually results from sources' or journalists' political orientation or from competition for attention. Repressive contexts add censorship and self-censorship as another layer of bias (McCarthy et al., 2008). A partial solution lies in what Beissinger terms "a 'blanketing strategy,' utilizing multiple sources and multiple types of information whenever they are available" (Beissinger, 2002, p. 476).

This idea also underlies two approaches that expand the range of source types used for protest event analysis. In the strictly controlled media environment of closed authoritarian regimes such as China, a viable alternative to news reporting may be to gather data from social media. In a pioneering study, Zhang and Pan (2019a) found that social media data does indeed help to increase the share of otherwise underreported rural protest, but so far it has been possible to automatically identify only a very small set of variables, excluding important information such as "size, action form, claims/issues, targets, organizers, and violence" (Zhang & Pan, 2019b, pp. 76–77).

The second approach involves activist projects that document protest. These often work through a network of activist-correspondents who monitor their local

environment while a website aggregates that information. In Russia, two such projects have been turned into data sets: Lankina Russian Protest Event Dataset (LAruPED) and Institute of Collective Action (IKD). In the context of restricted media freedom, these activist-based data sets have the potential to be simultaneously more efficient and less affected by self-censorship than media-based data. However, given that they come from politically interested sources, they may also oversample specific topics at the expense of others. The following section sets up a research design that systematically compares four data sets of different origins.

4. Data Sources and Research Design

The empirical analysis will be conducted in two stages. In the first stage, I compare the Mass Mobilization in Autocracies Database (MMAD) data set that is based on international news reporting to two data sets derived from activist websites. All three cover the whole of Russia in varying time periods, with an overlap from March 2007 through to March 2012. The second stage then compares all three data sets with the Contentious Politics in Russia (CPR) data set gathered by Andrei Semenov in two Russian regions. The MMAD data are available on their own website, all other data sets are being uploaded to the Discuss Data project.

The analysis is conducted to show the different pictures of a country's protest landscape that different data sets produce, in order to better understand possible distortions when relying on a single data set of known origin (e.g., activist-based or international media-based protest data). As the goal is not to arrive at a substantive insight on protest dynamics in Russia but, instead, to fully reveal these cross-source differences, in the analysis I consciously do not make the data sets more comparable to each other (for instance, by subsetting) but let the differences in sources and construction produce their full effects.

I now briefly describe each data set regarding the origin, coverage, and coding criteria, and summarize the differences in Table 1. Based on these characteristics, I then derive hypotheses on how their respective pictures of protest in Russia differ from each other.

4.1. Mass Mobilization in Autocracies Database

The MMAD has been developed by a team around Nils Weidman at the University of Konstanz. It covers all authoritarian regimes, identified in accordance with Geddes et al.'s research (2014). The data is based on three international news agencies: Associated Press, Agence France Press, and BBC Monitoring. While these are all English-language sources, BBC Monitoring also translates local news, considerably enlarging the overall coverage (Weidmann & Rød, 2019, pp. 42–44). Nonetheless, the authors make clear that the data set may introduce bias, "as these agencies typically cater to audiences in Western countries and primarily report on events that are of some interest to them" (Weidmann & Rød, 2019, pp. 41–42). MMAD includes only protest events with an identifiable political motive. The definition, however, is broad, including all "matters of, or relating to, the government or the public affairs of a country" (Keremoglu et al., 2020, p. 2). Events include, for example, protests against the monetization of social benefits or a spike in fuel prices—claims that would be labelled social or economic by other authors (see Lankina & Voznaya, 2015). A far greater restriction, which is likely to affect the composition of the database, is introduced by MMAD's minimum threshold of 25 participants for an event to be included. MMAD also covers pro-government rallies, which, however, are excluded here.

4.2. Lankina/Lankina Russian Protest Event Dataset

This data set is the first of two that is based on activist websites. To compile it, a team around Tomila Lankina coded all entries on protest that were posted on the website namarsh.ru between 2007 and 2016 (Lankina & Tertytchnaya, 2020). The site is funded by the opposition politician Garry Kasparov, who led the cross-ideological opposition group "Other Russia." Its content comes from a Russia-wide network of correspondents. As Lankina writes, "the website is run by a team of activists sympathetic to the cause of the opposition and therefore interested in ensuring wide reporting and the most comprehensive harvesting of published online reports on protest" (Lankina, 2015, p. 30).

In contrast to MMAD, this data set excludes pro-government events (Lankina & Tertytchnaya, 2020, p. 22). Also, it does not have a lower limit on the number of participants. However, the authors are clear that the source does not represent full coverage of all protests (Lankina, 2015; Lankina & Voznaya, 2015). For instance, as the data set comes from an opposition website, "it may well contain a liberal bias in favour of pro-democracy activism" (Lankina & Tertytchnaya, 2020, p. 24). However, as of yet, there are no systematic comparisons with other data sources that could indicate where any such biases lie.

4.3. Reuter and Robertson/Institute of Collective Action

The third data set was constructed by Reuter and Robertson (2015; see also Robertson, 2013), covering the period from January 2007 through to March 2012. Its origin is a website, the IKD, in which a group of sociologists with sympathies for left-wing oppositional causes compile weekly reports on protest events across the country. As with LAruPED, IKD has no lower limit on the number of participants (but it does not record this number). Hence, the two activist-based data sets are similar structurally, but their political outlooks are quite different. While Kasparov's "Other Russia" blends liberalism with nationalist elements introduced by one of its constituent groups (the National Bolsheviks), the

group behind IKD "seeks to unite different social groups it describes as being 'without a voice,' " including "leftist groups, labour unions, and environmental and youth organizations" (Reuter & Robertson, 2015, p. 241).

As evident from these short descriptions, the two activist-based data sets come from very different sources. LAruPEd can be expected to oversample political protest, while the IKD data are likely to prominently cover social protest (when the two are compared directly, this expectation is confirmed, see Supplementary File). It is therefore not to be expected that they would produce exactly the same picture of protest. Instead, it will be the task of the empirical analysis to estimate the degree to which they produce similar pictures of protest dynamics despite the potential differences in political outlooks of their author collectives.

4.4. Semenov/Contentious Politics in Russia

The final data set, CPR, was collected by Andrei Semenov and colleagues at the Center for Comparative History and Politics at Perm State University. Here, I use two publicly available portions of it, covering the regions of Perm and Tyumen. The former is based on an online search with the Integrum service that archives over 40,000 Russian online, press, radio, and TV sources (Semenov, 2017). The latter is based on a comprehensive search of two local online media as well as participant observation. Crucially, it also uses official data—not records on actual protest events collected by police (as in the case of Robertson's data from the later 1990s; see Robertson, 2013), but data on applications submitted for protest events (in Russia, as in many other countries, protests that exceed one person need to be registered with authorities, which provides an excellent potential source of data—even though in most cases these data are inaccessible). The CPR data have no demonstrator threshold. Both approaches use several types of sources, approximating Beissinger's (2002) "blanketing strategy" in different ways and should thus decrease bias that results from source selection and selective coverage.

Table 1 compares the characteristics of all four data sets. The Supplementary Files contain a discussion of how the (few) duplicates in the data sets were dealt with.

4.5. Hypotheses

In the first stage of the empirical analysis, MMAD, LAruPED, and IKD will be compared to each other regarding broad trends in the distribution of protest. Comparisons include the distribution over time, the share of protests events in the regions (as opposed to Moscow and St. Petersburg), as well the rank order of regions according to the number of protest events. The difference in sources, as well as MMAD's 25-person threshold, make it likely that, compared to the other two, these data paint a different picture of protest in Russia. First, media coverage, in general, follows "issue attention cycles" (Downs, 1972), which have a bearing on protest coverage (Oliver & Maney, 2000) and thus potentially affect fully media-based data sources like MMAD (see also Gladun, 2020). Moreover, Herkenrath and Knoll (2011) have found substantial differences in protest coverage when comparing international and national news media sources—differences that I expect to show in the analysis:

> H1) Since MMAD is based on international media and imposes a 25-person threshold, it will focus on larger and more visible events, which results in a different distribution of events when compared to LAruPED and IKD.

Moreover, data extracted from reporting in international news media likely overrepresent events in the capital "where most foreign journalists have their workplaces" (Wüest & Lorenzini, 2020, p. 49). Therefore:

Table 1. Comparison of data sets used in analysis.

Data set	Sources	Topic of events covered	Includes pro-government?	Minimum participant threshold	Number of events in covered period[1]	Number of regions with at least 1 event in covered period[2]
MMAD news reports	International	Political (broad definition)	Yes (but excluded for analysis)	25	1,152	76
LAruPED	Activist website	All	No	none	4,497	76
IKD	Activist website	All	No	none	5,593	81
CPR	Multiple local sources	All	No	none	458	—[3]

Notes: [1] The covered period ranges from 16 March 2007 (the start date of the LAruPED) through 5 March 2012 (the end date of the IKD data). [2] Before the annexation of Crimea in 2014, the Russian state counted 83 subnational subjects, including "oblasts," "krais," "autonomous republics," and others. For convenience, these are all summarized under the label "regions" here. [3] The CPR data is compared to LAruPED and IKD only on the regions of Perm and Tyumen, see Sections 6.1 and 6.2.

H2) MMAD will show a greater concentration of events in Moscow and St. Petersburg compared to LAruPED and IKD.

Similarity in source type, on the other hand, should increase convergence—even if the specifics of the sources vary:

H3) Because LAruPED and IKD both come from local activists they will, overall, show greater convergence with each other than either of them does with MMAD.

In the second stage, the activist-based data sets will be compared to the CPR data that likely come closer to full coverage of all protest as it approximates Beissinger's "blanketing strategy" (2002, p. 476) to different degrees, leading to the following hypotheses:

H4) The CPR data in Perm and Tyumen cover more protest events than even LAruPED or IKD.

H5) Since the Tyumen data include a greater variety of source types, when compared to LAruPED and IKD, they cover more unique events than do the CPR data on Perm when compared to LAruPED and IKD.

Finally, I code protesters' claims as reported in CPR to match LAruPED, assigning each event one or more of six broad categories: political, economic, social, legal, ecological, and cultural (Lankina & Voznaya, 2015, p. 332; see Supplementary Files for coding details). This offers the chance to compare the CPR data to LAruPED regarding the thematic coverage of protest. As Lankina and Tertytchnaya (2020) have pointed out, LAruPED may be biased in favour of political protest, leading to the final hypothesis:

H6) LAruPED oversamples political protest, which leads to a higher share of political protest compared with CPR.

5. Empirical Analysis I: General Trends

5.1. Development Over Time

First, I graphically explore the distribution of protest across time. As expected, Figure 1 shows that MMAD differs strongly from the other two, with protest peaking in late 2011 and early 2012. This was the time of the country-wide For Fair Elections (FFE) protests—the largest and geographically most widespread protests in post-Soviet history—which attracted comprehensive

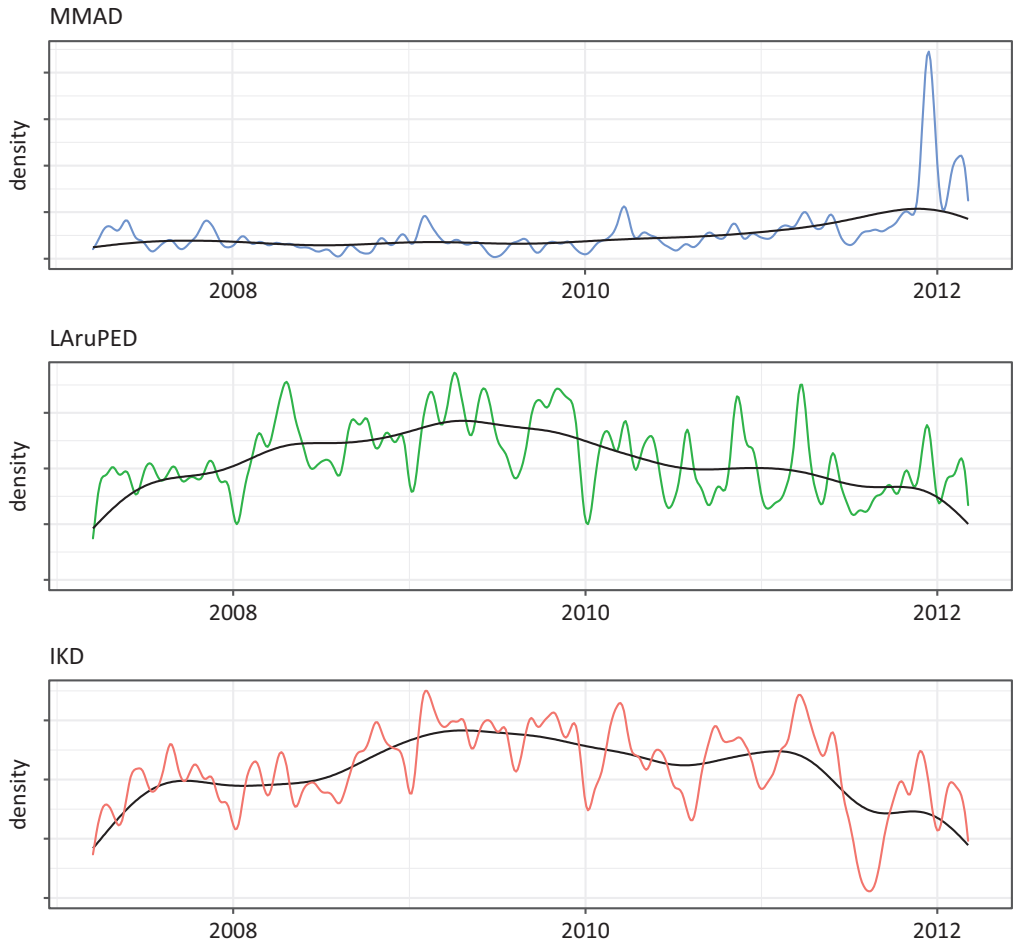

Figure 1. Protests events over time by data source, with density curve and trend line (March 2007–March 2012).

international media attention. But even when bracketing out this specific historical period, the distribution of MMAD diverges from that of LAruPED and IKD (see Figure A1 in the Supplementary Files). The latter two, by contrast, seem to depict broadly similar trends, confirming H1.

In the Supplementary Files, I show that the trend in the LAruPED data does not change markedly when the data are subsetted to only those events with 25 and more participants, which make them structurally more similar to the MMAD data (Figure A2). This suggests that differences between MMAD and the activist-based data sets do not just result from the different inclusion threshold.

5.2. Share of Regional Protest Events

The different trends observed above may, among other things, be because international media coverage underestimates the share of protest in the provinces (Gabowitsch, 2016). Figure 2 supports this assertion, demonstrating that MMAD features a substantially lower share of regional protest than both LAruPED and IKD. In 2011, however, the picture changes, with the share of regional protest in MMAD strongly increasing. This is likely once more connected to the FFE protests which, having begun in Moscow, quickly spread through the regions. This is supported when the period of the FFE protests is removed, which causes the share of regional protest in 2011 to drop substantially in all three data sets (see Figure A1 in the Supplementary File).

These observations suggest two things. First, the trend of concentration of protest in the capitals during the second half of the 2000s, as diagnosed by Robertson (2013), can be replicated with the LAruPED data (see also Lankina, 2015) and MMAD. This is a welcome finding as it points to the robustness of a broad trend that has become standard knowledge in protest research on Russia.

Second, there are two possible explanations for the sudden increase in the share of regional protest with the onset of the FFE protests. In one interpretation, the share surges because the FFE events are larger than previous regional protest so that more regional events cross MMAD's threshold of 25 participants. In the second explanation, the beginning of the protests wave triggered an "issue attention cycle," increasing the interest of international news agencies in regional protest. In other words, what used to be in the periphery of attention suddenly moved to the centre, which would mean that MMAD substantially underreported regional protest before the FFE protests. Both explanations are plausible. To test which has greater empirical support, I compute the average number of participants in regional protest events in LAruPED before and during the FFE period. If regional protest events did indeed become larger, this should be reflected in the LAruPED data as well.

The data show that regional protest events were indeed, on average, larger in the FFE period than they had been before: The median before December 4, 2011, is 100 participants. Between that date and the presidential elections on March 4, 2012, it doubles to 200. In part, this appears to be because more events cross the 25-participant threshold: Before the FFE protests, 24% of events are below that number; during FFE, that share drops to 9% (see table A1 in the Supplementary Files). Although these numbers do not give a full answer, they at least suggest that part of MMAD's strong rise in regional protest may be due to real changes in regional protest patterns rather than simply the shifting attention of international journalists.

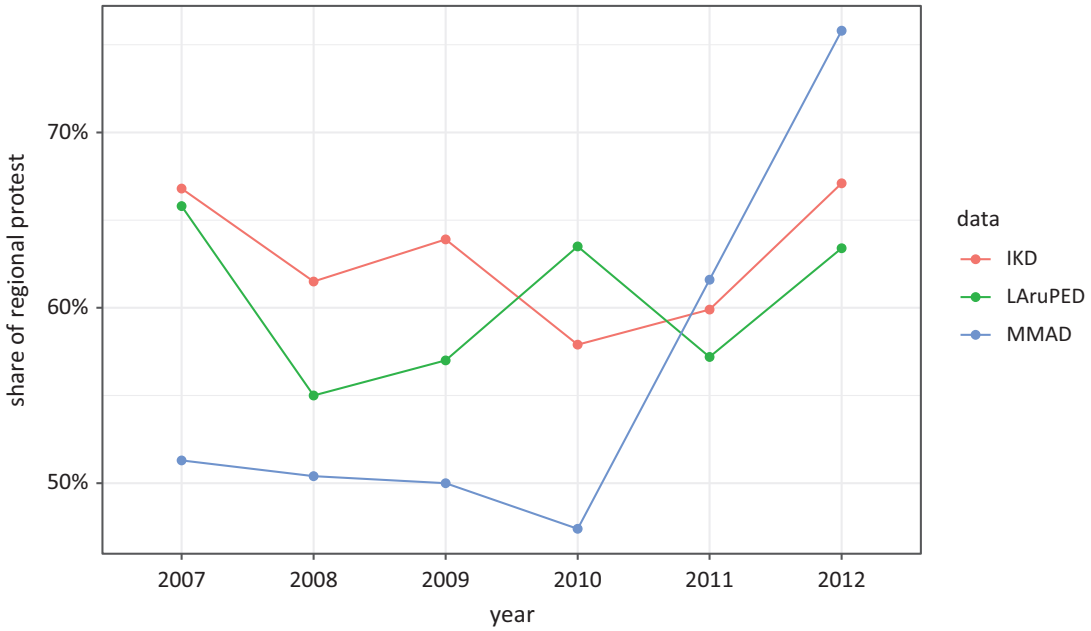

Figure 2. Share of protest events outside Moscow and St. Petersburg over time, by data source (March 2007–March 2012).

5.3. Coverage Across Regions

The two previous sections have shown broadly similar trends across the two activist-based data sets. Regarding regional protest, MMAD did not stray too far from the other two data sets, albeit on a different baseline and except for the high phase of the FFE protests. This leads to the question of whether these trends are only comparable for the country as a whole or whether they hold true across specific regions. Table 2 lists the regions with the highest event counts in the jointly covered period (out of 83 covered regions in total). Unsurprisingly, Moscow and St. Petersburg come out on top in all three data sets. Sverdlovsk, Kaliningrad, and Samara are also among the top 10. But there is also considerable variation: In LAruPED, Penza ranks fifth but is placed much lower in MMAD (rank 43) and IKD (rank 34). Similarly, Dagestan in the North Caucasus is surprisingly highly ranked in MMAD, which is mirrored neither in the LAruPED nor in the IKD data (ranks 35 and 38, respectively).

Do the data show entirely different pictures, or are these dissimilarities exceptions? To answer this question, I compute Spearman's rho of all of the 83 regions covered for all three pairs of data sets. Spearman's rho is better suited for this task than Pearson's r, because it compares the ranks of the regions rather than the absolute values, limiting outliers to the values of their rank. Figure 3 shows that the two activist-based data sets correlate strongly at *rho* = 0.79 ($p < .001$). The regions are not ranked identically—there is considerable variability as to how many events are covered per region—but they are usually in similar sections of the rank order. This indicates that broadly speaking, LAruPED and IKD document similar things. This is noteworthy and encouraging given the different political projects behind them. The same cannot be said for the correlations between MMAD and either of the other two, where Spearman's rho is at only moderate levels (0.48 and 0.59). This suggests that, while bearing some relation to the activist data, international media paint a rather different picture (certainly in part because of MMAD's inclusion threshold).

Having demonstrated that the MMAD data appear to cover protest in Russia's regions differently, I explore one of these differences to the activist-based data sets in greater detail. Table 2 shows Dagestan and North Ossetia, two regions of the North Caucasus, to be in MMAD's top 10. Both are ranked far lower in LAruPED and IKD. Table 3 compares the six core ethnic republics of the Russian North Caucasus—Chechnya, Dagestan,

Table 2. Regions with the highest number of events per data source (March 16, 2007–March 5, 2012).

MMAD			LAruPED			IKD		
Region	# of events	% of total	Region	# of events	% of total	Region	# of events	% of total
Moscow City	346	30.0%	Moscow City	1,288	29.6%	Moscow City	1,795	32.5%
St. Petersburg	152	13.2%	St. Petersburg	484	11.1%	St. Petersburg	309	5.6%
Dagestan	54	4.7%	Samara	176	4.0%	Leningrad	232	4.2%
Primorie	40	3.5%	Moscow Oblast	130	3.0%	Novosibirsk	166	3.0%
Sverdlovsk	37	3.2%	Penza	115	2.6%	Sverdlovsk	144	2.6%
Novosibirsk	27	2.3%	Sverdlovsk	104	2.4%	Samara	138	2.5%
Kaliningrad	23	2.0%	Kaliningrad	99	2.3%	Irkutsk	131	2.4%
Samara	20	1.7%	Voronezh	99	2.3%	Moscow Oblast	130	2.4%
Bashkortostan	19	1.6%	Primorie	96	2.2%	Kaliningrad	125	2.3%
North Ossetia	19	1.6%	Kirov	89	2.0%	Chelyabinsk	122	2.2%

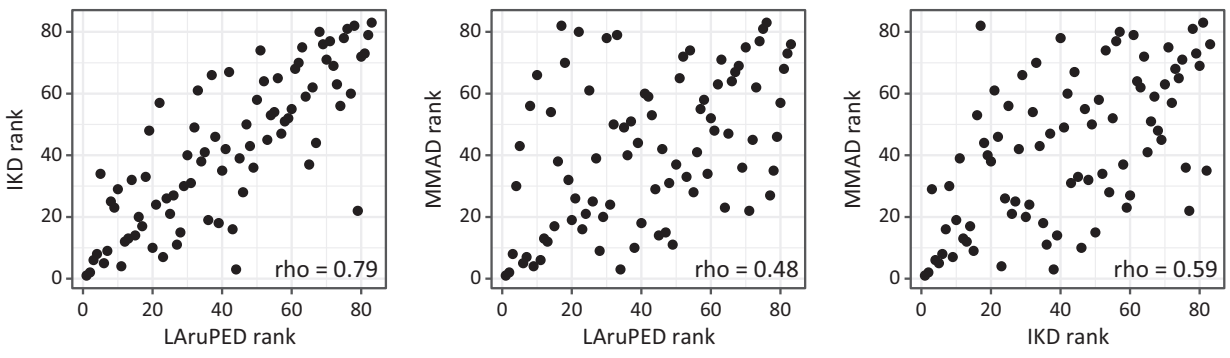

Figure 3. Rank correlations of the number of covered protest events per region, by combination of data sets (March 2007–March 2012).

Table 3. Protest events in North Caucasus by data source and year.

	2007	2008	2009	2010	2011	2012
MMAD	25 (31.6%)	31 (51.7%)	10 (14.3%)	11 (11.8%)	32 (12.4%)	5 (5.3%)
LAruPED	18 (4.6%)	16 (3.0%)	17 (2.7%)	21 (4.0%)	8 (2.0%)	3 (3.5%)
IKD	14 (2.7%)	19 (2.9%)	16 (1.8%)	17 (2.5%)	4 (0.7%)	1 (1.0%)

Notes: Cells show absolute numbers of reported events in the North Caucasus and the share among all regional protest (in brackets). North Caucasus defined as Chechnya, Dagestan, Ingushetia, Kabardino-Balkaria, Karachaevo-Cherkessia, and North Ossetia.

Ingushetia, Kabardino-Balkaria, Karachaevo-Cherkessia, and North Ossetia—demonstrating that this is a systematic difference. MMAD draws a much larger share of its regional protest events from that macro-region, particularly before 2011. Moreover, even though it covers far fewer regional protest events overall, in the Caucasus MMAD has higher *absolute* numbers in several of the years covered.

This finding allows two conclusions: On the one hand, the strong oversampling of the Caucasus compared to other regions indicates a bias resulting from reliance on English-speaking news, as the authors of MMAD note themselves. The North Caucasus being a region with a long history of instability and conflict, it might be that international news agencies have a particular focus on this macro-region. On the other hand, the comparison of absolute numbers also shows that the activist-based data sources under-report on the North Caucasus. Given the highly repressive context, where activist groups such as Other Russia (the source for LAruPED) have a poor standing, this is hardly surprising. But it points to an insight of broader relevance. Where the freedom of the press is limited, activist groups may be an important alternative source for protest data, but where freedoms are curtailed to such an extent that activist groups cannot associate freely, accurate protest coverage depends on either international journalists (who might enjoy a somewhat higher level of protection) or on local sources that are less formalized—and thus less easily targeted by repression—than the activist groups behind LAruPED and IKD. A look into the sources of the MMAD for the North Caucasus reveals that most come from the BBC World Monitoring service that translates local sources, making the latter the more likely explanation. At any rate, a hypothesis for further research could hence be that regime features may play an important role in the decision of which data to use for which questions.

6. Empirical Analysis II: Comparing Event Coverage

The previous section has looked at protest from a bird's eye perspective. I now turn to a brief comparison of the specific coverage of protest in two regions, comparing the activist-based data sets to Semenov's CPR data from Perm and Tyumen. Since Stage 1 has suggested that MMAD underreports regional protest in "normal times," i.e., in the absence of a cross-nationally diffusing protest wave, MMAD is excluded from this comparison.

6.1. Overlap in Event Coverage

A pressing question when comparing specific coverage is to what extent the data sets cover the same events. I approximate this by computing the overlap of the three data sets using the date as an indicator. When two data sets have an entry on the same date in the same region, these are counted as the same event. Certainly, this is a crude method. It would be preferable to use additional characteristics such as precise location or action form, which are, however, not given in the data. The Supplementary File contains at least a partial validity check of overlap between LAruPED and CPR; using the available information on specific events, I show that the date-based method is a useful approximation, but somewhat overestimates the overlap. This is a clear limitation. However, the crude method can still produce valid insights if the limitation is used productively: Calculating overlap only based on the date likely produces false positives, i.e., events that are classified as the same although they are not. At the same time, the method is unlikely to produce false negatives, i.e., events that are coded as different although they, in reality, are the same. Provided that the dates are assigned correctly, this procedure constitutes a most-likely case to detect convergence—meaning that any indication of difference under this scenario would likely translate into even larger differences if the events were identified in a more sophisticated way.

Figure 4 computes the overlap between the three data sets for both regions separately. For each region, it shows what portion of the overall date-region combinations (here treated as the same events) are covered only by either of the three data sets, by two of the three data sets, or by all of them. These Venn diagrams, therefore, visually illustrate the overlap between the data sets. The colouring, moreover, underscores the percentages displayed in each of the sections: White means a low share of protest events, red means a high share.

As the diagrams show, CPR covers by far the most events in both regions, proving that neither IKD nor LAruPED cover the full protest landscape. In Perm, 31% of all events are uniquely covered by CPR, in Tyumen that share stands at 65% (shown in the top section of the top circles in each diagram). This suggests that the Tyumen data, which contain official sources, are even more comprehensive than the Perm data that are based on media alone.

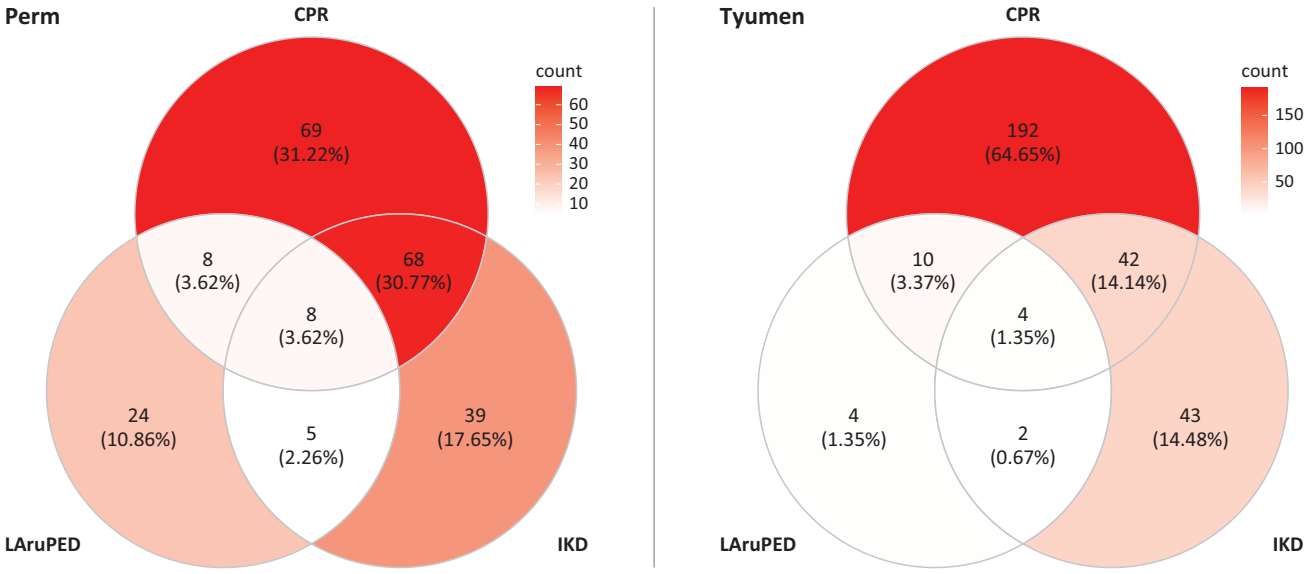

Figure 4. Overlap of three data sets, by region: Left panel shows Perm (March 16, 2007, through March 5, 2012); right panel shows Tyumen (January 1, 2008, through to March 5, 2012).

These findings confirm H4 and H5: In line with H4, CPR has a higher absolute number of covered events than LAruPED or IKD; in line with H5, when comparing the two cities, the CPR data from Tyumen have a larger share of unique events than they have in Perm, underlining the importance of diversifying the source base when attempting to approach full coverage. However, contrary to the assumption behind the two hypotheses, the figures show that CPR does not constitute full coverage either. Between 15% (Tyumen) and 30% (Perm) of all events are not covered by CPR. Therefore, even though the "blanketing strategy" generates more data than the activist-based approach, dissident websites appear to have a unique added value. The lower share in Tyumen suggests that this can be reduced by diversifying source types, but it cannot be eliminated.

Finally, the two different activist data sets appear to have quite different foci, as the overlap between LAruPED and IKD is rather small. Given that, in the country-wide perspective, the two have shown similar trends, the regions of Perm and Tyumen might be outliers. However, the observed divergence suggests that all researchers who are interested in specific regional protest events are well-advised to approach any single data set with caution.

6.2. Thematic Coverage

As the last step, I compare thematic protest coverage in the two regions between CPR and LAruPED. The latter is chosen because of its intuitive and easily adaptable coding scheme (Lankina, 2018), which was applied to CRP (see Supplementary Files). Table 4 displays protest in the two regions for each of LAruPED's six thematic categories as a share of the total number of coded topics. For Perm, the two data sets provide strikingly similar distributions. Except for economic protests (mostly against low salaries or wage arrears), the difference between the two data sets is marginal. Contrary to H6, LAruPED does not oversample political protest relative to a media-based approach. These results are especially noteworthy considering the previous analysis that showed only a small overlap between the two data sets on specific events. The results thus suggest that more efficient methods like LAruPED's *can* come to similar conclusions as the resource-intensive mining of local sources.

The numbers on Tyumen show, however, that this is not a given. Here, LAruPED clearly overreports political and underreports social and environmental protest in comparison to CPR. If these findings reflect more than a coincidence, one could tentatively conclude that the

Table 4. Topics of protest as a share of all, by region and data source.

Region	Data	political	economic	social	cultural	legal	environm.	# of events
Perm	LAruPED	27.7%	23.4%	29.8%	0.0%	8.5%	10.6%	45
Perm	CPR	27.6%	17.6%	31.8%	0.0%	12.9%	10.0%	167
Tyumen	LAruPED	52.4%	9.5%	23.8%	9.5%	4.8%	0.0%	20
Tyumen	CPR	39.2%	7.2%	30.1%	3.3%	11.0%	9.1%	295

Note: Perm covered March 16, 2007, through March 5, 2012; Tyumen covered January 1, 2008, through March 5, 2012.

activist nature of LAruPED seems to drive over-coverage of political events where the absolute number of events is low. In Perm, LAruPED contains 45 events, more than twice the number that it covers in Tyumen. Since this hypothesis cuts to the core of a tradeoff between resources and accuracy in protest event analysis data, it should be investigated more systematically in further research.

7. Discussion and Conclusion

This article gives an overview of four protest event data sets that use international media, dissident websites, and diverse local sources (including news reporting and official accounts). It pursued the question of whether the pictures that each data set paints of protest in Russia differ, and if so, in what way. One answer is bad news: what one learns about protest in Russia depends in part on the chosen data. The good news, however, is that these different pictures at least partially overlap, providing confidence that the data constitute more than statistical noise, and that, when carefully matching one's research question with the appropriate data, valid inferences can be drawn from all of them. To conclude, I differentiate three levels of analysis to briefly discuss the strengths and weaknesses that each type of data set provides with regard to specific research aims.

On the national level, protest event analysis can be used to identify "extraordinary times," periods of high protest activity that are the most likely to provoke a reaction from political authorities and are thus important when studying regime dynamics. The MMAD data may be best suited for this research interest: They identify the fewest events overall, but they clearly mark the protest wave of 2011–2012 (including its regional component), which arguably had the greatest effect on Russian politics of all protest periods in post-Soviet history (Dollbaum, 2020b; Greene & Robertson, 2019). The fact that MMAD spikes in 2011–2012 and thus clearly delineates the FFE period is, as comparison with LAruPED has shown, in part the result of more events crossing the 25-participant threshold. It might, in addition, also be driven by the greater attention of international news media. These two factors, then, make the MMAD data quite efficient in identifying periods of particular protest intensity that are most likely to have political consequences.

If one is interested in protest dynamics on a cross-regional level rather than on its overall effects on the regime as a whole (Gabowitsch, 2016), then the efficient MMAD approach distorts the picture. It might cover single, highly repressive regions better (see the results on the North Caucasus), but the general lesson is that, for studying regional trends, regional data are necessary. Here, the encouraging news is that LAruPED and IKD converge to a high extent, even though they do not cover the same specific events.

Finally, on a case study level, even the activist-based data strongly underreport absolute numbers. If single protest events are of interest, the analysis suggests that one cannot easily circumvent the cumbersome blanketing strategy which should, however, include activist data to reap their unique benefits. If, by contrast, the goal is not the single event but the distribution of protest topics, the comparison of Perm and Tyumen suggested that activist data may give a relatively accurate account—but only if they cross some threshold of absolute coverage. This last point, however, is an inductively generated hypothesis that needs to be investigated in further research. Table 5 summarizes the strengths and weaknesses of the different data set types.

Overall, the comparison gives reason for both confidence and caution. The most important insight is not that any data source outperforms another, but that researchers should invest time in matching their research goals to the data they use in pursuing them. This, of course, presupposes that different data sources exist to choose from.

This, finally, opens the question of how the findings—and conclusions drawn from them—travel to other contexts. In many other non-democratic regimes, the source base will be less generous than it is in Russia—for instance, because of a lack of well-established opposition

Table 5. Advantages and disadvantages of data sets compared in the analysis.

Data set	Sources	Advantages	Disadvantages	Level of analysis for best use
MMAD	International news reports	Identifies major protest episodes Available cross-nationally	Distorts cross-regional dynamics (especially of small protest events) Underreports absolute numbers	National/cross-national
LAruPED IKD	Activist-based	Data sets converge on regional dynamics Available cross-regionally	Underreport absolute numbers	Subnational comparative
CPR	Multiple local sources	Fairly comprehensive (but still not full coverage)	Resource intensive	Subnational case-study

projects like namarsh.ru or IKD or because of tighter restrictions on the online sphere. Under such conditions, it will not be possible to use different data sources for different research purposes. Nonetheless, findings from this study may inform the methodological discussion and the application of protest event analysis more broadly. Beyond providing empirical support for the requirement to match research question and data source, the findings suggest particular ways in which different types of data sources systematically differ in the way they cover protest. For instance, if the conclusions on the MMAD data are valid, then, in countries where data based on international sources are the only data available, it should be possible to identify periods of high protest intensity that are likely to trigger political responses (repression, concessions, etc.). Conversely, the analysis has shown that MMAD will be of less use when studying the subnational dynamics of protest. Moreover, the two activist-based data sets give a relatively similar picture of protest dynamics across regions even though they come from quite different activist groups, reducing fears of strong distortions introduced by such sources. Finally, however, the data show that no single source is sufficiently well-placed to provide a complete picture of protest in the studied period. This analysis, then, serves as a general reminder of the limits of our inferential capacity.

Acknowledgments

This publication was produced as part of the research project Comparing Protest Actions in Soviet and Post-Soviet Spaces—Part 2, which is organized by the Research Centre for East European Studies at the University of Bremen with financial support from the Volkswagen Foundation.

Conflict of Interests

The author declares no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Beissinger, M. R. (2002). *Nationalist mobilization and the collapse of the Soviet state*. Cambridge University Press.

Bremer, B., Hutter, S., & Kriesi, H. (2020). Dynamics of protest and electoral politics in the Great Recession. *European Journal of Political Research*, *59*(4), 842–866. https://doi.org/10.1111/1475-6765.12375

Dollbaum, J. M. (2020a). When does diffusing protest lead to local organization building? Evidence from a comparative subnational study of Russia's "For Fair Elections" movement. *Perspectives on Politics*. Advance online publication. https://doi.org/10.1017/S1537592720002443

Dollbaum, J. M. (2020b). Protest trajectories in electoral authoritarianism: From Russia's "For Fair Elections" movement to Alexei Navalny's presidential campaign. *Post-Soviet Affairs*, *36*(3), 192–210. https://doi.org/10.1080/1060586X.2020.1750275

Downs, A. (1972). Up and down with ecology: The "issue-attention cycle." *National Affairs*, *48*, 38–50. https://www.nationalaffairs.com/storage/app/uploads/public/58e/1a4/b56/58e1a4b56d25f917699992.pdf

Earl, J., Martin, A., McCarthy, J. D., & Soule, S. A. (2004). The use of newspaper data in the study of collective action. *Annual Review of Sociology*, *30*(1), 65–80. https://doi.org/10.1146/annurev.soc.30.012703.110603

Gabowitsch, M. (2016). *Protest in Putin's Russia*. Polity Press.

Geddes, B., Wright, J., & Frantz, E. (2014). Autocratic breakdown and regime transitions: A new data set. *Perspectives on Politics*, *12*(2), 313–331. https://doi.org/10.1017/S1537592714000851

Gladun, A. (2020). Protesting that is fit to be published: Issue attention cycle and nationalist bias in coverage of protests in Ukraine after Maidan. *Post-Soviet Affairs*, *36*(3), 246–267. https://doi.org/10.1080/1060586X.2020.1753428

Greene, S. A., & Robertson, G. B. (2019). *Putin v. the people: The perilous politics of a divided Russia*. Yale University Press.

Hanna, A. (2017). *MPEDS: Automating the generation of protest event data*. SocArXiv. https://doi.org/10.31235/osf.io/xuqmv

Hellmeier, S., Weidmann, N. B., & Rød, E. G. (2018). In the spotlight: Analyzing sequential attention effects in protest reporting. political communication, *35*(4), 587–611. https://doi.org/10.1080/10584609.2018.1452811

Herkenrath, M., & Knoll, A. (2011). Protest events in international press coverage: An empirical critique of cross-national conflict databases. *International Journal of Comparative Sociology*, *52*(3), 163–180. https://doi.org/10.1177/0020715211405417

Hutter, S. (2014). Protest event analysis and its offspring. In D. della Porta (Ed.), *Methodological practices in social movement research* (pp. 335–367). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198719571.001.0001

Keremoglu, E., Hellmeier, S., & Weidmann, N. B. (2020). *Coding instructions for the mass mobilization in autocracies database, version 3.0*. Mass mobilization in autocracies database. https://mmadatabase.org/about/documentation

Kriesi, H., Hutter, S., & Bojar, A. (2019). Contentious episode analysis. *Mobilization: An International Quarterly*, *24*(3), 251–273. https://doi.org/

10.17813/1086-671X-24-3-251
Kriesi, H., Koopmans, R., & Duyvendak, J. W. (Eds.). (1995). *New social movements in Western Europe: A comparative analysis*. University of Minnesota Press.
Lankina, T. (2015). The dynamics of regional and national contentious politics in Russia: Evidence from a new dataset. *Problems of Post-Communism*, *62*(1), 26–44. https://doi.org/10.1080/10758216.2015.1002329
Lankina, T. (2018). *Lankina Russian protest event dataset* [Data set]. http://eprints.lse.ac.uk/90298
Lankina, T., & Tertytchnaya, K. (2020). Protest in electoral autocracies: A new dataset. *Post-Soviet Affairs*, *36*(1), 1–17. https://doi.org/10.1080/1060586X.2019.1656039
Lankina, T., & Voznaya, A. (2015). New data on protest trends in Russia's regions. *Europe-Asia Studies*, *67*(2), 327–342. https://doi.org/10.1080/09668136.2014.1002696
Levitsky, S., & Way, L. A. (2010). *Competitive authoritarianism: Hybrid regimes after the Cold War*. Cambridge University Press.
McCarthy, J., Titarenko, L., McPhail, C., Rafail, P., & Augustyn, B. (2008). Assessing stability in the patterns of selection bias in newspaper coverage of protest during the transition from communism in Belarus. *Mobilization: An International Quarterly*, *13*(2), 127–146.
McFaul, M., & Stoner-Weiss, K. (2008). The Myth of the authoritarian model: How Putin's crackdown holds Russia back essay. *Foreign Affairs*, *1*, 68–84.
Oliver, P. E., & Maney, G. M. (2000). Political processes and local newspaper coverage of protest events: From selection bias to triadic interactions. *American Journal of Sociology*, *106*(2), 463–505. https://doi.org/10.1086/316964
Petrov, N., & Titkov, A. (2013). *Rejting demokratichnosti regionov Moskovskogo Centra Karnegi: 10 let v stroju* [The Carnegie Moscow Center's Rating or Regional Democracy: 10 years in the making]. Carnegie Endowment for International Peace.
Reuter, O. J., & Robertson, G. B. (2015). Legislatures, cooptation, and social protest in contemporary authoritarian regimes. *The Journal of Politics*, *77*(1), 235–248. https://doi.org/10.1086/678390
Robertson, G. B. (2013). Protesting Putinism: The election protests of 2011–2012 in broader perspective. *Problems of Post-Communism*, *60*(2), 11–23. https://doi.org/10.2753/PPC1075-8216600202
Semenov, A. (2017). From economic to political crisis? Dynamics of contention in Russian regions (2008-2012). *Österreichische Zeitschrift Für Politikwissenschaft*, *45*(4). https://doi.org/10.15203/ozp.1107.vol45iss4
Tarrow, S. G. (1993). Cycles of collective action: Between moments of madness and the repertoire of contention. *Social Science History*, *17*(2), 281–307. https://doi.org/10.2307/1171283
Tilly, C., Tilly, L., & Tilly, R. (1975). *The rebellious century*. Harvard University Press.
Weidmann, N. B., & Rød, E. G. (2019). *The internet and political protest in autocracies*. Oxford University Press.
Wouters, R. (2013). From the street to the screen: Characteristics of protest events as determinants of television news coverage. *Mobilization: An International Quarterly*, *18*(1), 83–105. https://doi.org/10.17813/maiq.18.1.y6067731j4844067
Wüest, B., & Lorenzini, J. (2020). External validation of protest event analysis. In H. Kriesi, J. Lorenzini, B. Wüest, & S. Hausermann (Eds.), *Contention in times of crisis: Recession and political protest in thirty European countries* (pp. 49–78). Cambridge University Press.
Zhang, H., & Pan, J. (2019a). CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, *49*(1), 1–57. https://doi.org/10.1177/0081175019860244
Zhang, H., & Pan, J. (2019b). The challenges of "more data" for protest event analysis. *Sociological Methodology*, *49*(1), 76–82. https://doi.org/10.1177/0081175019866425

About the Author

Jan Matti Dollbaum is a postdoctoral researcher at the University of Bremen. He studies social movements and their interaction with institutional politics across different regime types. His work has appeared in *Perspectives on Politics*, *Post-Soviet Affairs*, and *Social Movement Studies*, among others. Together with Ben Noble and Morvan Lallouet he is the author of *Navalny: Putin's Nemesis, Russia's Future?* (Hurst Publishers/Oxford University Press).