UNIVERSITY OF CALIFORNIA

Los Angeles

Machine Learning Methods of Protest Detection

A dissertation submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Anton Sobolev

2019

ABSTRACT OF THE DISSERTATION

Machine Learning Methods of Protest Detection

by

Anton Sobolev
Master of Science in Statistics
University of California, Los Angeles, 2019
Professor Chad J. Hazlett, Chair

This thesis develops a model that estimates the number of participants in a protest event. My central assumption suggests that those engaged in the same collective action should have similar trajectories during the event. Individual trajectories of those who are away should not express the same patterns in trajectories. I developed a Convolutional Neural Network (CNN) model with two input layers. The first layer contains data on the trajectories of easily identified subgroups of protesters and easily identifies subgroups of nonprotesters. The second layer indicates if a specific part of each trajectory is not observed directly but, instead, is imputed. The resulting classifier identifies citizens who share the same trajectory during the protest with an accuracy of 94 percent.

The proposed algorithm significantly outperforms these alternatives: Logistic Regression, Random Forest, and Artificial Neural Network. The advantage of CNN in this setting comes from the fact that this architecture applies feature-recognition filters to all segments of the trajectory and thus is robust to potential shifts among various trajectories. These shifts exist because protesters do not appear at the same time or in the same place. The resulting estimates are validated using estimates of protest events from the national press.

The dissertation of Anton Sobolev is approved.

Erin K. Hartman

Jeffrey B. Lewis

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2019

to Russia, which will one day be free

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful to the chair of committee, Chad Hazlett, for his tireless guidance. I am also grateful to my thesis committee members, Erin Hartman and Jeff Lewis, for their helpful advice.

I would like to express my special thanks of gratitude to Zachary Steinert-Threlkeld and Keith Chen. I have benefited enormously from talking to them.

I thank my fellow UCLA students for their help. Specifically, I am grateful to Maxim Ananyev, Natalia Lamberova, and Aaron Rudkin.

VITA

2014–2019	Teaching Assistant at UCLA
2013–2019	Ph.D. student in the UCLA Political Science Department
2010–2013	Lecturer at Higher School of Economics, Moscow
2008–2013	Research Fellow at Higher School of Economics, Moscow
2009–2011	MA student at Higher School of Economics, Moscow
2005–2009	BA student at Higher School of Economics, Moscow

PUBLICATIONS

Dagaev, Dmitry, Natalia Lamberova, and Anton Sobolev. “Stability of Revolutionary Governments in The Face of Mass Protest.” European Journal of Political Economy, 2019, 60: pp. 2–20.

Sobolev, Anton, and Alexei Zakharov “Civic and Political Activism in Russia.” The New Autocracy: Information, Politics, and Policy in Putin’s Russia, edited by Daniel Treisman, Brookings Institution Press, Washington, D.C., 2018, pp. 249–276.

Lazarev, Egor, Anton Sobolev, Irina Soboleva, and Boris Sokolov. “Trial by Fire: a Natural Disaster’s Impact on Support for the Authorities in Rural Russia.” World Politics, 2014, 66(4): pp. 641–668.

Yakovlev, Andrei, Anton Sobolev, and Anton Kazun. “Means of Production VS Means of Coercion: Can Russian Business Limit the Violence of Predatory State?,” Post-Soviet Affairs, 2014, 30(1): pp. 171–194.

Remington, Thomas, Anton Sobolev, Irina Soboleva, and Mark Urnov. “Social and Economic Policy Trade-Offs in the Russian Regions: Evidence from Four Case Studies,” Europe-Asia Studies, 2013, 65(10): pp. 1855-1876.

Smyth, Regina, Anton Sobolev, and Irina Soboleva. “Well-Organized Play: Symbolic Politics and the Effect of the Pro-Putin Rallies,” Europe-Asia Studies, 2013, 60(2): pp. 24-39.

CHAPTER 1

Introduction

Contentious collective action is one of the key objects of study across social sciences. The size of the protest crowd is an essential factor in the theoretical models of many fields of study. They include democratization (Acemoglu and Robinson, 2005; Gandhi and Przeworski, 2006), social movement (Koopmans and Rucht, 2002; Olzak, 2004), resource mobilization (Jenkins and Eckert, 1986), revolution (Tilly, Tilly and Tilly, 1975), informational cascades (Kuran, 1989; Lohmann, 1994; Weyland, 2010), and many others.

To test these theoretical models, researchers need credible estimates of the number of participants in the collective action. The primary source of these estimates has been traditional media; in particular, newspapers. However, these estimates often lack credibility for three main reasons. First, media outlets can choose which events to cover. Second, newspapers publish secondary estimates-numbers reported either by the organizers or the police. All of them can have incentives to misrepresent the number of protesters. Third, both organizers and the police usually lack the accurate technology to count the participants.

The progress in communication technologies helps newspapers address issues of selection bias. The adoption of social media helps protest organizers promote their events, while newspapers actively use online content to monitor essential developments. At the same time, the issue of technology that accurately estimates the size of crowds remains unaddressed.

Recently, some scholars developed alternatives to newspaper-based approaches for generating protest event data (see Steinert-Threlkeld, 2019, for a review).

They try to make use of the increased availability of data generated by Internet users. That data often includes social media posts and photos of the protest. In contrast to traditional approaches, the new approaches provide an explicit and clear methodology behind the estimates.

This thesis attempts to develop a model that estimates the number of participants at a protest. It assumes that individuals engaged in the same collective action should have similar trajectories during the protest event. Trajectories of those who are far away should not share similar trajectories. I have developed a Convolutional Neural Network model with two input layers. The first input layer contains data on the trajectories of easily identified subgroups of protesters and easily identifies subgroups of nonprotesters. The second input layer indicates if specific parts of each trajectory are not observed directly but rather imputed. Together, these two input layers can identify those who share the same trajectory during the protest with 94 percent accuracy and 95 percent precision. The proposed algorithm significantly outperforms these alternatives: Logistic Regression, Random Forest, and simple Neural Network.

CNN outperforms others, I argue, because this architecture applies feature-recognition filters to all segments of the trajectory and is thus robust to potential shifts among various trajectories. These shifts can be temporal or spatial. Temporal variations occur since citizens can join the protest event at different times; spatial shifts are caused by citizens who join the protesting crowd further down the route. The CNN feature extraction step aids the model in recognizing that two individuals had similar trajectories in the presence of such shifts. The algorithm is able to perform classification (protesters or nonprotesters) by looking for low level features of the trajectories such as edges and curves. It then builds up to more complex features through a series of convolutional layers (e.g., a zig-zag sequence). The model to identifies the presence of the class feature disregarding its distance from the starting point of the trajectory.

The resulting estimates are validated using estimates of the protest events from the national press.

The remainder is organized as follows. The next section reviews traditional as well as recently introduced methods of quantifying protests. Section 3 develops a general approach to protest detection, its assumptions and implementation. Section 4 introduces the results, and Section 5 is a conclusion.

CHAPTER 2

Review of existing methods

2.1 Traditional methods of quantifying protest

Newspapers. Most of the time, scholars rely on newspapers to collect records of collective protests. Typically, a social scientist uses libraries and archives to check the files of newspapers for a specific period and place. The first systematic collection of such data was introduced in (Tilly, Tilly and Tilly, 1975). The authors produced a dataset of violent collective actions in Italy, Germany, and France by using local presses from 1830 to 1930. Tarrow (1989) combined data on contentious collective action events in Italy's national news from the second part of the 20th century. In the United States, The New York Times is the most cited source of protest data. For instance, Olzak (2004) used its records to collect data on collective actions related to ethnic politics at the end of the 19th century.

Newspapers are also used in cross-country studies. For example, Hutter and Kriesi (1995) conducted a comparison of protest activities in European countries from 1975 through 1990 by using records in their national newspapers. The New York Times Index was the first research project to introduce large-scale, cross-national, time-series data on protests around the globe (Walton and Ragin, 1990). To improve the credibility of the protest numbers, some scholars suggested using multiple sources to determine the facts of the events (Myers and Caniglia, 2004).

Using a newspaper as a source for records raises several concerns. First, editors can choose which events to cover, which introduces selection bias. There are two points of view concerning the direction of this bias. One group of authors suggests that the vast majority of protests are, in fact, small events that never make the news (Maney and Oliver, 2001; Myers and Caniglia, 2004). Thus, the frequency and average size of protests could be systematically underestimated or overestimated. This approach should underestimate the degree of variation. According to the other view, newspapers cover both small and large protests (Koopmans and Rucht, 2002). The issue, however, arises from the fact that scholars often aggregate protest events by month or by year. Small protests mostly contribute to the frequency of collective actions, while large protests contribute to the number of participants. Most of the quantities of interest (e.g., the average size of the protest) appear to be misleading.

Another bias refers to the fact that editors of traditional media favor new events (Koopmans and Rucht, 2002; Steinert-Threlkeld, 2019). Repetitive, collective actions such as strikes are likely to be underreported as well.

Although both biases described above are important, the essential issue of using press records concerns the validity of protest size estimates. Usually, newspapers report secondary numbers produced either by the organizers or by the police. Both the former and the latter can have incentives to misrepresent the number of protesters. The Million Man March, a demonstration held in Washington, D.C., on October 16, 1995, is the most well-known example of a disparity in estimates. While civic activists estimated the crowd size at 2 million people, the local police reported 400,000. The latter provoked a wave of resentment and led to the introduction of a bill that explicitly forbids police from calculating crowd size (Watkins, 2001).

Hutter (2014) introduces a large-scale project of cross-national protest event data collection. The coding protocol has a set of strict requirements. For example, to be included in the collection, a protest episode should appear in at least three media outlets. Coders are also required to cover the primary source of the record. That approach allows the use of a wide range of characteristics of the event, including the type of mass action, the civil society organizations involved, the targets, the claims, the number of arrests, and, most importantly, the total number of participants. At the same time, this approach has several essential disadvantages. It is highly expensive when it comes to time, labor, and the requirement to carefully read back issues of newspapers. As a result, data collection is restricted to specific periods and countries.

Supposedly, the progress in communication technologies helps newspapers address issues of selection bias. With the adoption of social media, protest organizers have been promoting their events online, while both national newspapers and local ones actively use online content to monitor events. At the same time, the issue of valid estimates of the protest size remains unaddressed.

Crowd counting. To count each person in a crowd is the most straightforward method of estimating the size of a protest. However, it is almost impossible to implement it in most cases. Contentious collective actions usually take place in open spaces such as streets and town squares. Citizens can freely join, leave, or rejoin the crowd, thus making a direct count infeasible. In some cases, organizers or the police install turnstiles to precisely count the number of people entering an event. These cases, however, are rare.

To deal with the task of estimating protest size, scholars usually use indirect methods of crowd-counting. An interesting approach is considered in Budros (2011) in which the author studied the effects of collective events on the abolition of slavery in the U.S. North. He hypothesizes that emancipation was affected by anti-bondage protests organized by Quakers. To measure the scale of protest activity, the author makes use of the Quakers' custom to petition at their political gatherings. Signatures on a petition thus effectively revealed the size of the meeting.

Another way to estimate the size of a protest suggests obtaining dimensions of the gathering place and the physical density of demonstrators. McPhail and McCarthy (2004) introduce three requirements to calculate proper estimates. First, these calculations require information on the carrying capacity of the spaces in which people assemble. They point out that most of the urban areas have places where people typically gather. Estimates of the carrying capacity of a specific location can usually be obtained from local police, scientists, or journalists. Second, correct estimates of a protest size require knowing the density of occupation. The researcher can get it either by sending observers to the event or from the available photos of the crowd. The authors also suggest that one person per 2 square feet is the maximum occupation density. Finally, valid estimates of the number of participants require observations from several vantage points. That requirement comes from the fact that mass protests are usually more densely packed at the center and the front than at the back or sides. Thus, scholars need to approximate the geographic distribution of occupation density.

Cell phone data. Pierskalla and Hollenbach (2013) study the impact of cell phone technology on the mass political violence in Africa. They hypothesize that the availability of cell phone coverage substantially increases the chance of violent collective actions. They use cell phone data to calculate cell phone coverage as well as to estimate the number of participants of these actions. To do this, they match GSM data with the geographic locations of recorded episodes of collective violence. For each event, they approximate the number of participants by summing up unique cell phone identifiers in the area. While this approach has distinct advantages, several concerns about the validity of the estimates remain. First, the number of protest participants can be systematically overestimated. Indeed, the authors assume that at the location, every owner of a cell phone joins the collective action. Second, the number of protesters could also be underestimated since the number of citizens without mobile phones is not observable. Although, systematic overestimation is more plausible as approximately 91 percent of South African adults owned mobile phones in 2016 (Asongu and Asongu, 2019).

The next section provides a review of approaches that use machine learning techniques to measure the size of the protest.

2.2 Machine learning methods of measuring protest participation

Several recent studies use alternatives to newspaper-based approaches to generate protest events data (see Steinert-Threlkeld, 2019, for a review). These studies try to make use of the increased availability of data generated on the Internet. That data often include social media posts and photos of the protest. It is reasonable to expect that these new methods will deliver more accurate estimates of the size of the protest and its collective actions. In many cases, the producers of that content are representative of the population of protesters (e.g., users of digital platforms such as Facebook and Twitter). In contrast to newspaper-based approaches, the new ones provide explicit and clear methodology behind the estimates. They also try to take into account potential biases in the data. Next, I'll discuss in detail two of the recent papers.

Won, Steinert-Threlkeld and Joo (2017) develop an automated system that analyzes images shared during contentious collective actions. They use a visual segment recognition approach to automatically assess the attributes of mass events. They start the project by generating these three types of labels for each image in the dataset: event status (protest or nonprotest); the size of the crowd; and the level of perceived violence. This task is implemented by workers hired at Mechanical Turk (AMT). The authors assign two workers to each image to improve the reliability of the results. Then the authors use 50-layer ResNet architecture to build Convolutional Neural Network that uses labeled images for training. The trained classifier can distinguish a protest crowd from other gatherings (e.g., sporting events and concerts). Most importantly, the model analyzes segments on images to identify and count the total number of participants in the event in the photo.

Zhang and Pan (2019) introduce a system that generates protest data by using texts and images from social media as inputs. To identify mass protests, they develop a two-stage classifier comprised of several convolutional neural networks and recurrent neural networks with long, short-term memory. As in the previous study, the authors obtain initial data from social media posts. Recall that reports or posts that cover collective action events are rare compared to the total number of documents in the corpus. Indeed, less than 0.01 percent of sampled posts from the Chinese digital platform Weibo contain information about actual offline collective action. To address this problem, the authors replace random samples of posts with a collection of posts obtained by using a dictionary prepared by experts. The authors' final dataset consists of almost 10 million posts with corresponding texts, images, and metadata (e.g., geolocation, number of reposts, time of posting, etc.). In the next step, two classifiers are trained separately. The Convolutional Neural Network is used for image classification. A mixture of the Recurrent Neural Network with Long Short-Term Memory and the Convolutional Neural Network is used to analyze textual components. Finally, the authors augment trained classifiers by another neural network to jointly model data representation and perform the task of classification. This sophisticated machine learning architecture allows distinguishing social media posts that describe offline protests from posts that discuss the same topics without references to collective action.

CHAPTER 3

Machine learning model of protest detection

3.1 General approach to protest detection

Most of the existing machine-learning studies of protests rely on images or texts as the primary source of data. In contrast, my research builds on the ideas of Pierskalla and Hollenbach (2013) and develops a classification algorithm to estimate the size of a protest from the GSM data. To the best of my knowledge, that is the first study that applies statistical learning to GSM data in order to detect a protest event and estimate the number of participants.

In this section, I discuss the general idea of the algorithm. The next section describes its actual implementation. Here, I consistently refer to Figure 3.1 to guide the reader. The model described below attempts to guess the number of participants conditional on the fact that the protest event, indeed, occurred.

Step 0: Setup. Consider citizens of town X. A group of them collectively protest at time Y. While citizens can protest for a long time, Figure 3.1a represents a snapshot of the start of the political rally. Assume that a researcher knows the location of each citizen and how that location changes two hours after the start of the protest.

Step 1: Location-based classification of citizens. First of all, the research needs to identify locations with the highest density of citizens. This locations are the candidates for the place of the protest event. I start by dividing the map of the location into squares of the same size (Figure 3.1b). This size can be almost arbitrary and is required to initiate the detection algorithm. I then calculate the number of citizens in each square. One of the squares contains more citizens than the average number of citizens per square. I suggest that all the citizens in that square in fact participate in the protest. I call them core protesters. Citizens A, B, and C are far away from the core protesters. They do not present physically at the place of protest. One can thus suggest with confidence that A, B, and C do not participate in the protest.

I refer to A, B, and C as nonprotesters. Apart from core protesters and nonprotesters, there are two other citizens in the town (labeled #1 and #2). Both of them are relatively close to the core protesters. Thus, one cannot say for sure that #1 and #2 are nonprotesters since they physically present near the center of the protest event. To correctly estimate the total number of protesters, one needs to decide if #1 and #2 are, in fact, part of the protest.

What evidence can a researcher use to infer whether #1 and #2 participate in the collective action? A naive approach would suggest training classifiers using the geographic coordinates of core protesters and nonprotesters and applying the resulting model to observations with unobserved labels (#1 and #2). Effectively, the trained model would establish a border around the core protesters. It then would classify any citizens within this border as protesters and otherwise as nonprotesters. By construction, anyone who happens to be around during the protest would be counted in the total number of the collective action participants. Does it sound reasonable? Consider the following. Most of the protest events take place in the centers of urban areas. For instance, from 20,000 to 100,000 tourists visit the National Mall in Washington, D.C., daily (Benton-Short, 2007). At the same time, many political rallies also take place in that same area. The naive machine learning classifier would treat these tourists as protesters. The resulting estimates would thus be systematically biased upward. An alternative strategy is described below.

Step 2: From locations to trajectories. Recall that the researcher knows how the location of each citizen changes. In other words, he knows the path each citizen made during the first two hours of the protest event. This is known for everyone disregarding his or her actual type (protester or non-protester). For each citizen, it is possible to depict his or her trajectory. I define a trajectory as a path made by a citizen during the first two hours of the protest event in the absence of information about his actual location. While the path made by the citizen has specific coordinates on the map, his or her trajectory is the same path with initial coordinates moved to the point (0,0). Formally, I define a trajectory as a spatial sequence that is ordered in time but not location-specific.

Figure 3.1c depicts hypothetical trajectories of core protesters and nonprotesters. Here, I introduce a vital assumption: because protesters participate in the same collective action, their paths should look alike. Indeed, consider the Women's March, a large-scale protest in Washington, D.C., in January 2017. Event organizers proposed to start the march on Independence Avenue at the southwest corner of the Capitol and continue along the National Mall. Existing photos and videos suggest that protesters indeed followed that route (Weber, Dejmanee and Rhode, 2018).

In contrast, nonprotesters do not engage in the same activity. They can be located in places remote from each other; their trajectories should not systematically correlate. Combined, they should look like white noise. Figures 3.1d and 3.1f illustrate the idea of the differences in trajectories of protesters and nonprotesters.

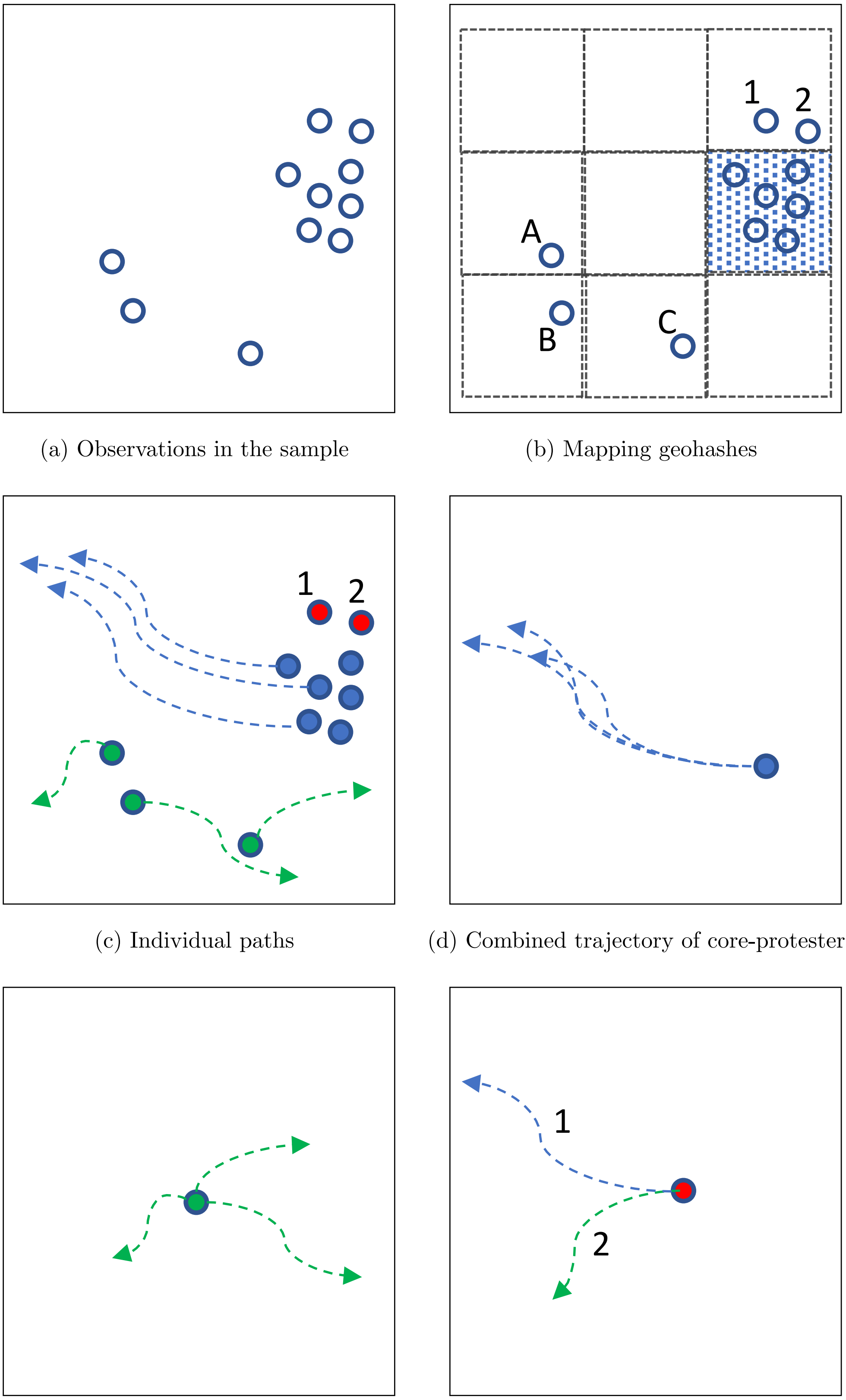
Step 3: Trajectory-based classification of citizens. A trajectory-based classifier can have precisely the same architecture as one that uses locations as its primary input. Although one nuance is essential here, the first classifier knows the geographic coordinates of the observations, and the second one does not. Indeed, in the latter case, all trajectories start from the same point (0,0). Thus, the only informative feature that this model can exploit is the similarities in the trajectories of core protesters.

Step 4: Making inferencing about citizens in "grey zone". In the last step, a researcher applies the trained model to the trajectories of citizens whose status is unknown. In my example, the trajectory of citizen #1 looks similar to the average trajectory of the core protesters. But the trajectory of the second one does not overlap the latter. Thus, the classifier assigns the former to the group of protesters and the latter to the nonprotesters.

Remarks. One objection to the above approach comes from the group of protesters; the trajectories can vary a lot. For instance, citizen #1 could be further from the core group of protesters at the start of the protest. In this case, her trajectory will deviate from the average path of the core group. Figure 3.6 illustrates that example. Such deviations can significantly reduce the predictive capacity of the classifier. To deal with this issue, I suggest using CNN as a classifier of choice in order to solve this task. That network trains visual filters and applies them to each segment of the picture. Such a model should thus not consider complete paths. Instead, it should look for overlaps in trajectories of protesters. I provide further details in Section 3.4.

3.2 Data

I implement and test the performance of the algorithm developed in Section 3.1 using the Women's March in Washington, D.C., on January 21, 2017. The campaign was organized the day after the inauguration of President Donald Trump. The march was comprised of more than 300 street demonstrations. Participants of the events took to the streets to express opposition to the administration and its policies (Fisher, Dow and Ray, 2017). To train the algorithm, I match proprietary GSM data (cell phone location data) and metadata of protest events from the Crowd Counting Consortium (Chenoweth and Pressman, 2017). The latter is an open-source project established to collect data on political crowds reported in the United States. Specifically, I use this open-source data to determine the



(a) Observations in the sample

(b) Mapping geohashes

(c) Individual paths

(d) Combined trajectory of core-protesters

(e) Combined trajectory of non-protesters

(f) Trajectory-based classification of observations with with unknown type

Figure 3.1: Illustration of the general approach to protest detection

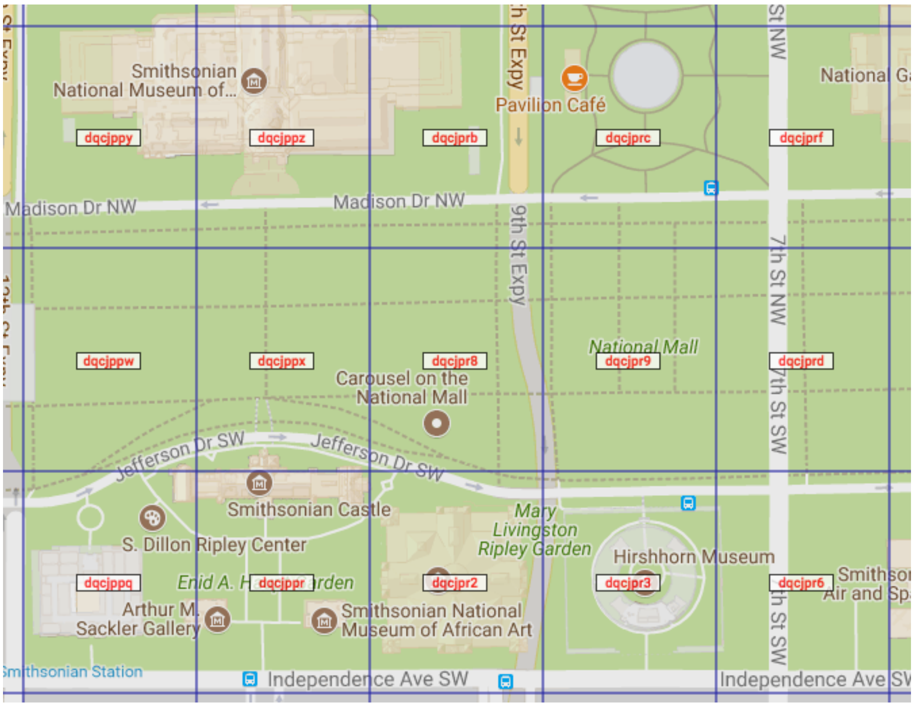


Figure 3.2: Example of geohash distrbution

time and location of protest events across the country. I focus on a sample of 341 protest events organized on the same day.

I start by identifying the geographic location of each area that experienced a protest event on January 21, 2017. For each location, I then determine the unique GSM identifiers to approximate the population of interest. It is noteworthy that, each GSM identifier reports its location each 15 minute periods regardless of the actual activity. Following step #2 of the previous section, I divide the map of the location into cells with sides of 38 meters by using eight-digit geohashes (i.e., alphanumeric codes that correspond to a rectangle with sides of 38 meters) (see An et al., 2013, for details). Next, I calculate the density of observations for each geohash at the start time of the protest event. Finally, I select geohashes in the top one percentile with respect to the density (a similar method is used in Sobolev et al., 2019). An example of the geohash distribution is provided in Figure 3.2.

Following the theoretical model, I assign the observations within selected geo-

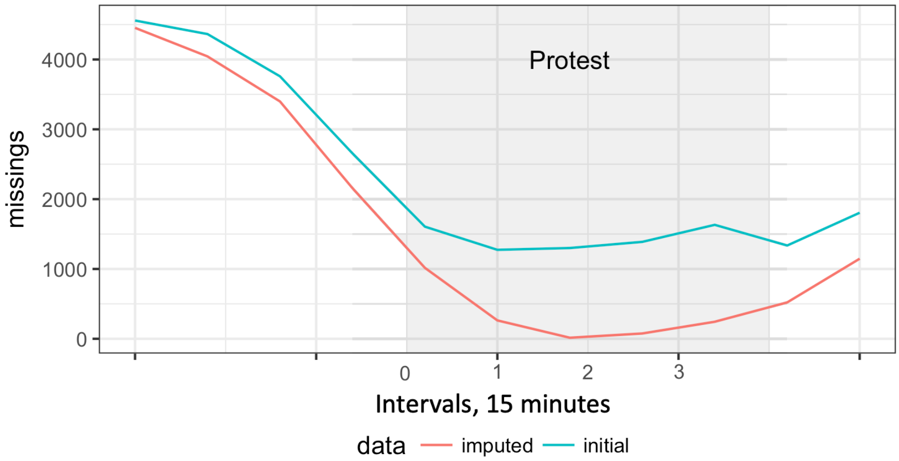


Figure 3.3: Missing data and linear imputation

hashes to the core protesters group. I consider geohashes located at least 1.5 miles away from the center of the core protesters group as those with nonprotesters.

3.3 Feature engineering

To construct the trajectories, I consider 15-minute intervals to code an average location of the GSM identifier. I then transform locations into paths by subtracting the initial position of each observation from its location in every 15-minute interval. Thus, the eight points represent the trajectory of each observation. It is noteworthy that not all of the observations reported their coordinates for each 15-minute interval. To address this issue, I conduct linear interpolation (i.e., if an observation is missing in one of the intervals, I impute its location by calculating the mean of locations of the previous interval and the following one).

This solution is not ideal and might be a pitfall at the training step of the algorithm. Indeed, Figure 3.2 shows that 20 percent of observations at the interval levels are missing. Supposedly, the classifier can identify interpolated parts of the trajectories as overlapping features of one of the group. Thus, the classification task could be corrupt. I have two considerations concerning this issue. First, as one can see, the number of missing goes down at the start of the protest event.

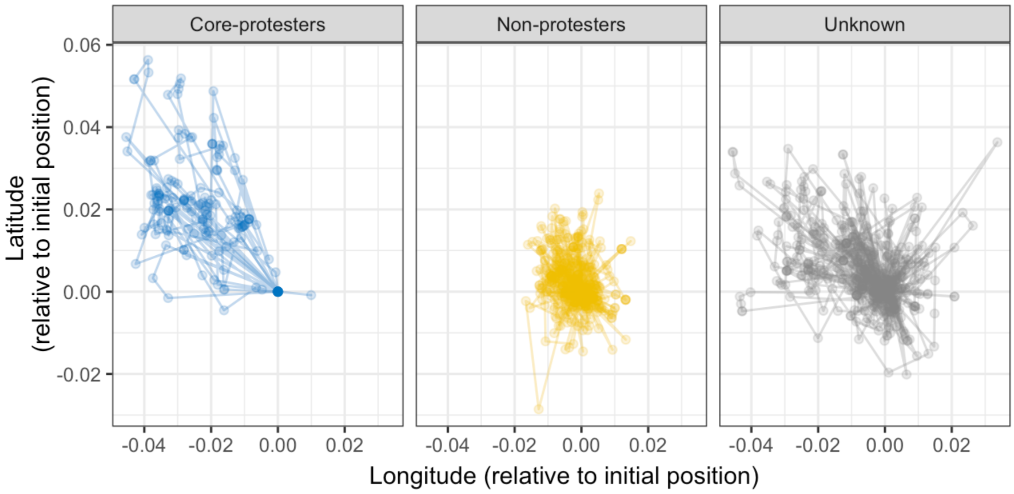


Figure 3.4: Groups of observations

The drop in missing observations coincides with the time the maximum number of people appear in the open area. I also check that most of the missing data come from the nonprotesters group. Finally, I suggest using auxiliary data, a binary-coded table that indicates if an observation is missing in a given interval of time. This auxiliary input helps the classifier distinguish between observed and imputed parts of trajectories and puts lower weights on the latter. In other words, this auxiliary input works as a regulator that induces a penalty if a classifier attempts to use imputed data.

Are the resulting features consistent with the theoretical model from Section 3.1? Figure 3.4 shows the aggregated trajectories from a randomly sampled protest event. Indeed, the paths of those in the core protest geohashes share a significant overlap. Trajectories of nonprotesters look like white noise. Finally, the group of observations in the “grey zone” resembles a mixture of the first two groups. Further, I use processed data as input for the training and test steps of the classification.

3.4 Implementation

CNN Architecture. Typically, Artificial Neural Network (ANN) is made up of many highly interconnected nodes, which process information by their dynamic state response to external inputs (Appeltant et al., 2011). Nodes are organized in layers of three types. The first type, the input layer, contains the raw data and communicates to layers of the second type (also called hidden layers). The latter processes data via a system of weighted connections and activation functions. The last hidden layer links to the output layer. This layer contains scores for each label in the data.

When ANN is initially presented with a pattern, it makes a random guess about the class it belongs to (e.g., protesters or nonprotesters). Next, ANN calculates the difference between the prediction and the actual class of the observation, adjusts its connection weights. This process repeats itself each time the network is presented with a new input pattern and is called statistical learning via the backward error propagation of weight adjustment.

In contrast to ANN, Convolutional Neural Network performs pixel-wise feature extraction. The idea of neural networks with convolutional layers is introduced in LeCun et al. (1998). The authors develop a classifier to implement computer vision task (e.g., image recognition). They argue that in order to successfully perform the task of computer vision, the model should process elements of the images independently from the location of these elements on the image. The latter means that features should be identified locally. To do that, the model analyzes each part of the image with a window of the pre-specified size. The results of the convolution operation can be passed to the next layer. A sequence of convolutional layers produces a higher-order feature map. In other words, after processing the raw data, the model identifies complex patterns of the image (e.g., a person's eye color).

Two other essential elements of CNN include the pooling layer and the fully

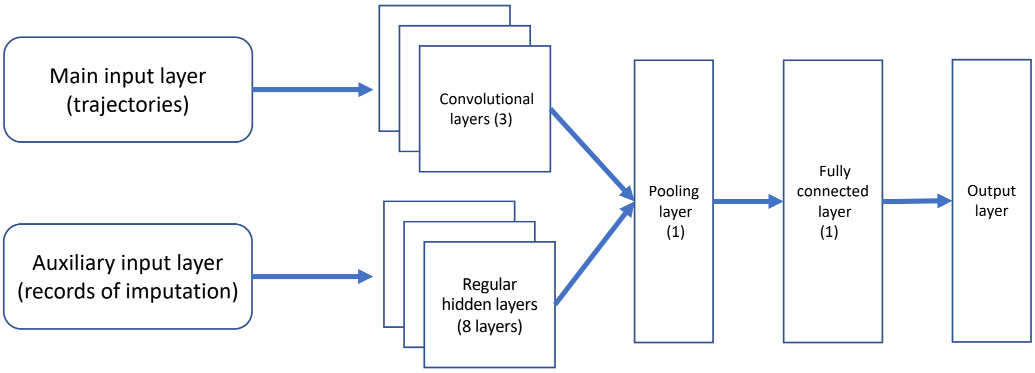
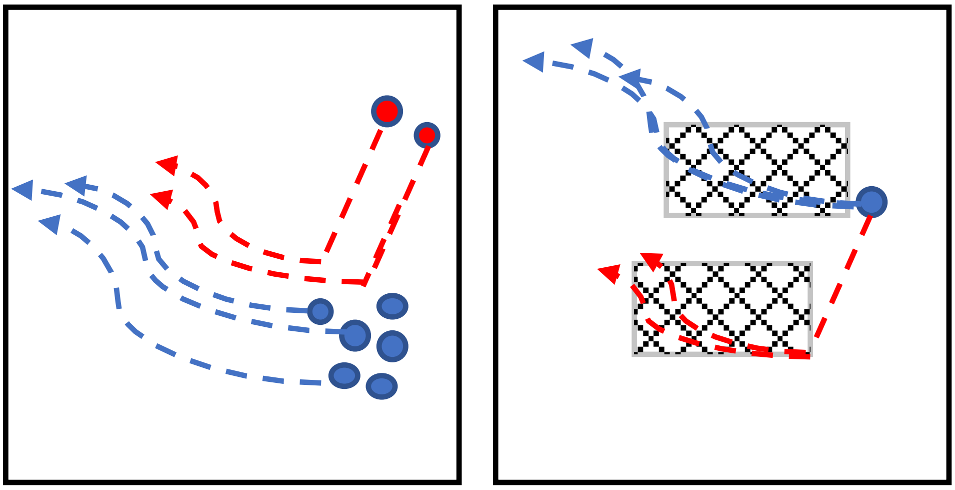

Main input layer (trajectories)
Auxiliary input layer (records of imputation)
Convolutional layers (3)
Regular hidden layers (8 layers)
Pooling layer (1)
Fully connected layer (1)
Output layer

Figure 3.5: Architecture of proposed CNN

connected layer. The former concatenates the output of separated clusters of the model. The latter connects every component of the previous layer with each element of the fully connected layer. Fully connected layers allow the model to learn nonlinear combinations of the features generated by convolutional or regular hidden layers of the neural network.

The CNN feature extraction step aids the model in recognizing that two individuals had similar trajectories in the presence of such shifts. The algorithm is able to perform classification (protesters or nonprotesters) by looking for low level features of the trajectories such as edges and curves. It then builds up to more complex features through a series of convolutional layers. CNN outperforms others, I argue, because it applies feature-recognition filters to all possible segments of the fixed-size of the trajectory. Thus the model is able to identify the presence of the class feature (e.g., a specific zig-zag sequence) disregarding its distance from the origin (0,0).

Setup. In this section, I develop the Convolutional Neural Network to estimate the size of the protest. I argue that CNN, with two separate input layers, should have the best performance across the potential alternatives. I include Logistic Regression, Random Forest, and simple Artificial Neural Network in the list of other options. I also consider two specifications for each class of classifiers. The first specification only takes trajectory data as its input, while the second



(a) Paths of core-protesters (in blue) and observations in the "grey zone" (in red)

(b) A convolutional filter

Figure 3.6: Filters in CNNs

one also uses an auxiliary input layer with records of missing observations.

Figure 3.6 illustrates the idea discussed above. Recall the example from the previous section. Consider that both citizens #1 and #2 are, in fact, protesters. At the same time, they join the protest later than core protesters. It is noteworthy that the researcher does not observe a person during all the time he or she is protesting. Instead, the researcher only keeps the trajectories generated in the first two hours after the official start of the protest. Thus, if #1 and #2 join the collective action 30 minutes after the start, their trajectories should differ from the average path of the core protesters in two respects. First, their trajectories have a tail at the beginning that depicts the route they took to catch up with the other protesters. Second, if the protest event is a march or a rally, #1 and #2 will follow other protesters with a delay. Thus, their trajectories will not have the path that core protesters crossed at the end of the second hour. Now consider red and blue trajectories from 3.6b. All three models on the list of alternatives (Logistic Regression, Random Forest, and simple Artificial Neural Network) should consider red and blue trajectories as very distinct from each other. At the same time, CNN can tune one of its filters to recognize just that part of the trajectory that all observations share.

What CNN developed in this study has the following architecture (see Figure 3.5). The model has two independent input layers. The first layer contains the trajectory data. Three convolutional layers analyze these data. The auxiliary input provides the records of imputation. The latter is sequentially analyzed with eight hidden layers (each layer has eight neurons). The pooling layer concatenates the outcomes of the convolutional layers with the last hidden layer. The latter passes the results of concatenation to the fully connected layer to explore nonlinear combinations from the feature map. Finally, the output layer produces predictions about the class of the observation. This layer uses one-hot encoded matrix with two classes and "softmax" activation function for calculating resulting scores.

Performance metrics. To assess the performance of the trained models, I use four conventional metrics: accuracy, precision, recall, and F1 score. I define them below.

Accuracy - the ratio of correctly predicted observation to the total observations.

Precision - the ratio of correctly predicted positive observations to the total predicted positive observations.

Recall - the ratio of correctly predicted positive observations to all observations in the actual class.

F1 score - the weighted average of precision and recall. Since I would like to avoid both upward and downward biases in the estimates, it is equally important to minimize errors in predicting both classes in the data. Thus, accuracy and F1 score are the most critical metrics for the task of this study.

Training. Figure 3.7 shows the training of the model. This figure helps us make two observations. First, the model has smoothing convergence. Second, the calculated total loss of the training set becomes lower than the complete loss of

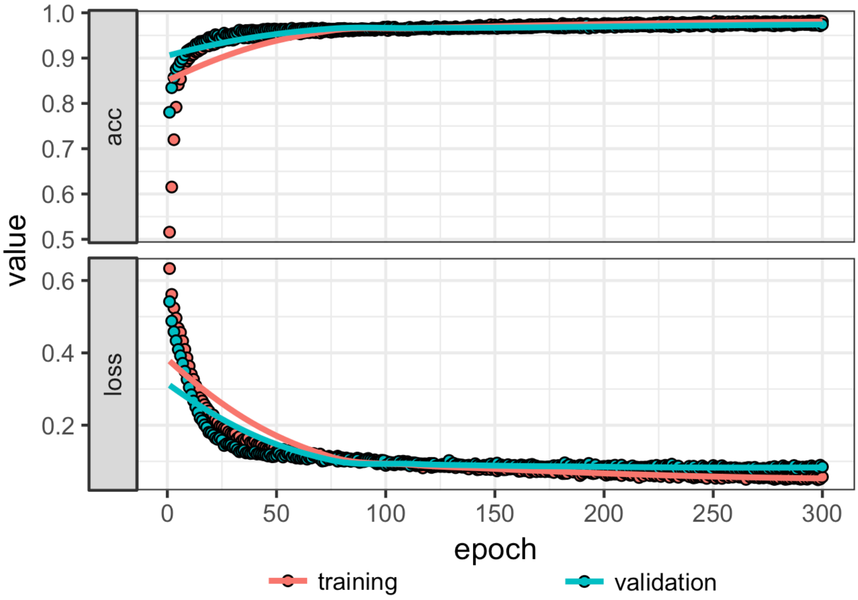


Figure 3.7: Model training

the validation set after the 100 epochs. The latter indicates potential overfitting. That is why I stop the training step after the first 100 epochs.

I train the model separately for each protest event. For each protest event, core-protesters (observations in the geohashes in the top one percentile with respect to the density) and nonprotesters (observations in geohashes located at least 1.5 miles away from the geohashes with core-protesters) are divided into a training and a test set. Observations in the "grey zone" are excluded. The latter contains 1/3 of the observations. The final scores represent the average of the corresponding scores weighted according to the number of observations in each location. All scores are calculated calculated on the test set.

CHAPTER 4

Results

Table 4.1 reports the main results of the study. First, the convolutional neural network indeed shows the highest scores on all metrics. A specification of CNN with additional auxiliary input provides the best performance across all the trained models with recall, F1 score, and accuracy of 94 percent and precision of 95 percent. Two of three referent models-Logistic Regression, Random Forest, and ANN-show very similar performances. CNN outperforms its specifications with and without auxiliary input by 10 and 15 percentage points, respectively.

Why does the auxiliary input layer only improve the performance of the CNN model? As I mentioned before, the usefulness of auxiliary data comes from the ability of the model to distinguish between observed and imputed parts of the trajectories. However, apart from CNN, models do not treat the path as an image. In the absence of convolutional filters, none of them can efficiently identify protest trajectories and consequently make use of auxiliary data. There is also an additional rationale for the inability of Logistic Regression and Random Forest to use the auxiliary input. In the ANN and CNN models, the primary and auxiliary inputs are trained separately in the first layers of the network. They are concatenated at the stage when both auxiliary and central parts of the network operate higher-order features. At the same time, Random Forest and Logistic Regression do not allow for the separate training of inputs. Instead, they pool both inputs, making it harder for classifiers to figure out the essential features of the classes.

I use both CNN specifications to predict the classes of observations in the grey

Model	Recall	Precision	F1-score	Accuracy
Logistic Regression	0.68	0.62	0.65	0.63
Logistic Regression (aux. layer)	0.72	0.70	0.72	0.72
Random Forest	0.78	0.83	0.80	0.79
Random Forest (aux. layer)	0.76	0.81	0.79	0.79
ANN	0.77	0.83	0.79	0.80
ANN (aux. layer)	0.78	0.83	0.80	0.80
CNN	0.88	0.84	0.86	0.86
CNN (aux. layer)	0.94	0.95	0.94	0.94

Table 4.1: Performance of trained models on test data

Specification	Dependent variable:	
	Estimates from news	
CNN	0.888***	
	(0.055)	
CNN with auxiliary input		0.974***
		(0.050)
Observations	341	341
R2	0.439	0.533
Adjusted R2	0.437	0.531
Residual Std. Error	1.397	1.275
F Statistic	264.895	386.286

Note: $^{***}p < 0.01$

Table 4.2: Validation of models using estimates from newspapers

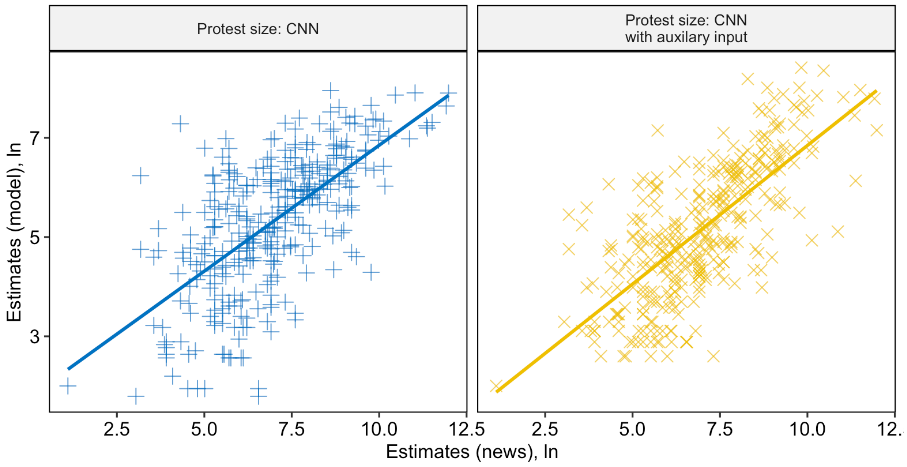


Figure 4.1: Correlation with protest data from newspapers

zone. To calculate the final estimates, I sum up the number of core protesters with the number of predicted protesters for each location. Finally, I attempt to conduct an indirect validation of the main results. To do it, I compare resulting numbers from both CNN specifications with estimates of protest size from Crowd Counting Consortium project (Chenoweth and Pressman, 2017). Figure 4.1 and Table 4.2 show the results of the comparison (values are log-transformed). Estimates from both models are linearly associated with estimates from Crowd Counting Consortium. Regressing, resulting estimates on numbers from Crowd Counting Consortium project return the coefficients of almost 1. At the same time, estimates from CNN with auxiliary input slightly better explain the variation in numbers from Crowd Counting Consortium project.

CHAPTER 5

Conclusion

The goal of this thesis is to develop a model that estimates the number of participants of the events using the GSM data. The model is based on the assumption that participants of the same collective action should have similar trajectories during the event. At the same time, paths of those who are away should not express the same patterns in trajectories. I develop a Convolutional Neural Network model with two input layers. The first input layer contains data on the paths of easily identified subgroups of protesters and easily identified subgroups of non-protesters. The second input layer indicates if a specific part of each trajectory is not observed directly but rather imputed.

The resulting model identifies those who share the same path during the protest with an accuracy of 94 percent and a precision of 95 percent. The proposed algorithm significantly outperforms the alternatives: Logistic Regression, Random Forest, and Artificial Neural Network. The superior performance of CNN is explained by its capacity to train and apply filters to various parts of the trajectories. Thus, in contrast to other models, CNN is robust to potential shifts in trajectories explained by the fact that protesters do not appear at the same time and the same place. I validate the results using estimates of the protest events from the national press.

REFERENCES

Acemoglu, Daron and James Robinson. 2005. Economic origins of dictatorship and democracy. Cambridge University Press.

An, Jin, Cheng Cheng-qi, Song Shu-hua and Chen Bo. 2013. ”Regional query of area data based on Geohash.” Geography and Geo-Information Science 29(5):31–35.

Benton-Short, Lisa. 2007. ”Bollards, bunkers, and barriers: securing the National Mall in Washington, DC.” Environment and Planning D: Society and Space 25(3):424–446.

Budros, Art. 2011. ”Explaining the first emancipation: social movements and abolition in the US north, 1776-1804.” Mobilization: An International Quarterly 16(4):439–454.

Chenoweth, Erica and Jeremy Pressman. 2017. ”This is what we learned by counting the women's marches.” The Washington Post 7.

Fisher, Dana, Dawn Dow and Rashawn Ray. 2017. ”Intersectionality takes it to the streets: Mobilizing across diverse interests for the Women's March.” Science Advances 3(9):eaao1390.

Gandhi, Jennifer and Adam Przeworski. 2006. ”Cooperation, cooptation, and rebellion under dictatorships.” Economics Politics 18(1):1–26.

Grimmer, Justin and Brandon M Stewart. 2013. ”Text as data: The promise and pitfalls of automatic content analysis methods for political texts.” Political analysis 21(3):267–297.

Hanna, Alex. 2017. ”MPEDS: Automating the generation of protest event data.” Working paper.

Jenkins, J Craig and Craig M Eckert. 1986. ”Channeling black insurgency: Elite patronage and professional social movement organizations in the development of the black movement.” American Sociological Review 51(6): 812–829.

Koopmans, Ruud and Dieter Rucht. 2002. ”Protest event analysis.” Methods of Social Movement Research 16:231–259.

Kuran, Timur. 1989. ”Sparks and prairie fires: A theory of unanticipated political revolution.” Public Choice 61(1):41–74.

LeCun, Yann, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86(11): 2278–2324.

Lohmann, Susanne. 1994. ”The dynamics of informational cascades.” World Politics 47(1):42–101.

Maney, Gregory M and Pamela E Oliver. 2001. ”Finding collective events: Sources, searches, timing.” Sociological Methods Research 30(2):131–169.

McPhail, Clark and John McCarthy. 2004. ”Who counts and how: estimating the size of protests.” Contexts 3(3):12–18.

Molnar, Csaba, Frederic Kaplan, Pierre Roy, Francois Pachet, Peter Pongracz, Antal Doka and Adam Miklosi. 2008. ”Classification of dog barks: a machine learning approach.” Animal Cognition 11(3):389–400.

Myers, Daniel J and Beth Schaefer Caniglia. 2004. ”All the rioting that’s fit to print: Selection effects in national newspaper coverage of civil disorders, 1968-1969.” American Sociological Review 69(4):519–543.

Olzak, Susan. 2004. ”Ethnic and nationalist social movements.” The Blackwell companion to social movements pp. 666–693.

Pierskalla, Jan and Florian Hollenbach. 2013. ”Technology and collective action: The effect of cell phone coverage on political violence in Africa.” American Political Science Review 107(2):207–224.

Saraf, Parang and Naren Ramakrishnan. 2016. EMBERS autogsr: Automated coding of civil unrest events. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM pp. 599–608.

Schrodt, Philip. 2012. ”Precedents, progress, and prospects in political event data.” International Interactions 38(4):546–569.

Sobolev, Anton, Keith Chen, Jungseock Joo and Zachary C. Steinert-Threlkeld. 2019. ”Newspapers and social media accurately measure protest size.” Working paper.

Steinert-Threlkeld, Zachary C. 2019. ”The future of event data is images.” Sociological Methodology 49(1):68–75.

Tarrow, Sidney G. 1989. Democracy and disorder: protest and politics in Italy, 1965-1975. Oxford University Press.

Tilly, Charles, Louise Tilly and Richard Tilly. 1975. The rebellious century: 1830-1930. Harvard University Press.

Walton, John and Charles Ragin. 1990. ”Global and national sources of political protest: Third world responses to the debt crisis.” American Sociological Review 55(6): 876–890.

Watkins, S Craig. 2001. ”Framing protest: News media frames of the Million Man March.” Critical Studies in Media Communication 18(1):83–101.

Weber, Kirsten M, Tisha Dejmanee and Flemming Rhode. 2018. ”The 2017 Women’s March on Washington: An analysis of protest-sign messages.” International Journal of Communication 12:2289–2313.

Weyland, Kurt. 2010. ”The diffusion of regime contention in European democratization, 1830-1940.” Comparative Political Studies 43(8-9):1148–1176.

Won, Donghyeon, Zachary C Steinert-Threlkeld and Jungseock Joo. 2017. ”Protest activity detection and perceived violence estimation from social media images.” In Proceedings of the 25th ACM international conference on Multimedia. pp. 786–794.

Zhang, Han and Jennifer Pan. 2019. ”Casm: A deep-learning approach for identifying collective action events with text and image data from social media.” Sociological Methodology 49(1):1–57.