Online Appendix for The Effect of Drone Strikes on Civilian Communication: Evidence from Yemen

This PDF file includes:
Supplementary Text
Figs. S1 to S3
Tables S1 to S6

Relevant Literature

Scholars have used cellphone data to study locational trajectories and patterns of mobility among subscribers (Eagle et al., 2009; Gonzalez et al., 2008; Calabrese et al., 2011; Blumenstock, 2012b; Deville et al., 2014), as well as to infer wealth and other socioeconomic properties (Blumenstock, 2012a; Smith et al., 2013; Blumenstock, 2016). These data have also been extended to studies on patterns of interpersonal contact (Aharony et al., 2011; Eagle et al., 2009; Jiang et al., 2013), evolving behavioral trends (Altshuler et al., 2012; Palla et al., 2007), transmission of diseases (Finger et al., 2016), and security related matters (Altshuler et al., 2013).[1]

More closely related to this work, CDRs have also been used to examine the change in patterns of communication after rapidly evolving and transient events, and in particular emergencies. Similarly to this work, Candia et al. (2008), Bagrow et al. (2011) and Gao et al. (2014) detect anomalous events based on the aggregate volume of communications. In particular, Bagrow et al. (2011) identify a behavioral change in social networks of communication after exogenous shocks such as bombings, earthquakes or blackouts. They find that there are spatially and temporally localized spikes in communication following an emergency, but that the information actually gets broadly transmitted resulting in what they term as communication avalanches.[2]

While our work focuses on quantifying the effect of anomalous events on aggregate cell-

[1] Scholars are increasingly turning to big data from social media to measure varied social contexts such as government censoring (King and Roberts, 2013), the structure of social networks, and how information spreads (Monroe et al., 2015). Data from Twitter or Facebook has also been used for measuring political mobilization (Bond et al., 2012), for monitoring anti-American views (Jamal et al., 2015), and support for ISIS (Mitts, 2019). However, social media data can be limited to the extent that such users are representative of the larger population - a concern allayed by the more universal cellphone network (Nagler and Tucker, 2015). In addition, big data needs to be used cautiously, and is best suited for data-poor contexts (Lazer et al., 2014).

[2] Gao et al. (2014) focus on characterizing the change in the patterns of information flow during emergencies as compared to non-emergency events. Relatedly, Lin and Lazer (2011) study how emergencies affect people?s effective social networks, and find that during emergencies people's strong ties are reasserted (see also Kenett and Portugali, 2012).

phone activity, several papers use other related techniques and measures. For instance, Altshuler et al. (2013) detect anomalies in patterns of communication not based on the aggregate volume of communications per se, but rather by making use of structural properties of the network of calls and texts. Blumenstock et al. (2016) use mobile airtime transfers to show how people engage in more altruistic behavior during a prominent earthquake in Rwanda by transferring mobile airtime to people affected by the earthquake. Blumenstock et al. (2014) use mobile money transactions to study the relationship between violence and people's financial decisions in Afghanistan.

Finally, modeling and prediction of conflict has benefited from the availability of event data on violence. Shapiro and Weidmann (2015) combine call detail records with spatio-temporal data on violence events, and find that increased mobile communications reduced insurgent violence, both at the district level and for specific local coverage areas. Modeling of conflict has also been studied without use of call detail records; as an example, Zammit-Mangion et al. (2012) provide a predictive framework of high accuracy as applied on the Afghan War Diary.[3]

Data Appendix

Our dataset combines data on cellular activity and drone strikes in Yemen from January 2010 until October 2012, based on availability of the relevant call detail records (CDRs).

[3] Another relevant literature for our paper is work that studies the relationship between conflict and information and communication technology (ICT). While most of the literature has studied the impact of improved communications on conflict, there is less work on how conflict affects communication. Early work suggested that technological innovations such as cellphones and internet would lead to an increase in conflict and political violence, since better communication technology helps rebel groups overcome collective action and coordination problems (Pierskalla et al., 2013). However, recent work has strongly questioned this link between ICT improvement and conflict (Dafoe and Lyall, 2015). For instance, improvements in ICT may lead to less violence, since better communication enables the government to engage in more effective intelligence gathering and for civilians to pass on information about insurgents to governments (Weidmann, 2015).

Call Detail Records Data

Our information on CDRs is a country-wide anonymized dataset of individual level cellphone communications from subscribers of a major cellphone service provider in Yemen. These include all calls, indicating time and location.

User: An anonymized identifier of the user who initiated the communication.

Recipient: An anonymized identifier of the recipient who received the communication.

Time: Time stamps of the call start and end of the communication. Given in Unix Epoch time (the number of seconds that have elapsed since 00:00:00 UTC Thursday, 1 January 1970, minus the number of leap seconds that have taken place since then).

Location of User: The geo-referenced location (latitude and longitude) of the cell tower that serviced the call on the side of the initiator.

Location of Recipient: The geo-referenced location (latitude and longitude) of the cell tower that serviced the call on the side of the recipient.

Drone Strikes Data

Our information on drone strikes in Yemen comes from merging the New America Foundation. and the Bureau of Investigative Journalism datasets.[4] This data includes the universe of drone strikes as recorded in sources in the public domain. Note that we supplemented this data, when possible, with information provided in publicly available news reports.

Date: Drone strike date.

Location: Drone strike location, including its geolocation (latitude and longitude).

Militants Killed: Numbers of militants killed in the strike. Low and high estimates are provided.

Civilians Killed: Numbers of civilians killed in the strike. Low and high estimates are pro-

[4] For the New America Foundation data, see `http://securitydata.newamerica.net/drones/yemen-analysis.html`. The Bureau of Investigative Journalism dataset is available at `https://www.thebureauinvestigates.com/stories/2017-01-01/drone-wars-the-full-data`.

vided.

Unknown Killed: Numbers of individuals killed in the strike where it is not known if they were militants or civilians. Low and high estimates are provided.

Total Killed: Total number of individuals killed in the strike, including militants, civilians, and unknown. Low and high estimates are provided.

Rank of Militant Killed or Targeted: Based on news reports. We coded as "high-level" militants provincial al-Qaeda in the Arabian Peninsula commanders or higher commanders. We coded as "not-high level" militants local commanders, unranked/unspecified militants, or cases where there was no mention of a high-level target in any of the available reports. In a few cases, any militant who was known to be involved in the planning for attacks on American targets was coded as "high-level".

Type of Target: Some reports indicate what type of target (militant facility, house/building, and/or vehicle, or unknown) was hit in the drone strike.

Militant facility: Some reports specifically outline that certain targeted buildings were training camps or known ammunition depots. When this level of fidelity is available, the target was coded as a militant facility.

House/Building: The number of residential houses or unspecified buildings targeted.

Vehicle: Number of cars, trucks, or motorcycles targeted.

Number of Missiles: Indicates the number of missiles fired in the drone strike.

Summary Statistics

Table S1 shows the summary statistics for our drone strikes data. The Table indicates that the average drone strike leads to seven to nine militant deaths. The data also implies that most strikes happen in the late evening or early morning.

Measurement of Variables for Panel Models

Our main dependent variable is call volume (incoming and outgoing). To capture the fact that people's call patterns differ by time of day, we consider three eight-hour intervals: morning (12:00 am - 8:00 am), midday (8:00 am - 4:00 pm) and evening (4:00 pm - 12:00 am). Subsequently, we aggregate our spatial data by eight-hour intervals for each unique tower location.[5] To account for the fact that call volume increases over time between 2010 and 2012, we normalize call volume using a z-score normalization. We find that call volumes tend to be periodic by day of the week, so we compare the number of calls during an interval against the number of calls made the same day of the week and during the same time interval for the ten preceding and ten following weeks.[6]

Our main treatment variable is the occurrence of a drone strike within a certain geographic proximity. Specifically, we consider towers within a radius of fifteen miles of a drone strike as being impacted. For robustness, we also consider radii up to 100 miles using five-mile intervals. For the time of drone strikes we rely on news accounts. When news accounts do not mention the estimated time of the strike, we impute the time based on our detection methods. For strikes which we do not detect and for which we are unable to impute the time, we include a time uncertainty measure.[7]

[5] Note that we treat all geolocated antennas in the same location (typically four or five) as one tower location. We do this in order to avoid redundancy.

[6] In other words, calls made Monday morning on October 10, 2011 are compared to the ten preceding Monday mornings (between June 27, 2011 - October 3, 2011) and the ten following Monday mornings (between October 17 - December 19, 2011). This choice ensures a rich enough set of 20 baseline call volumes to normalize against, so that normalization is meaningful. In light of the increasing trend of daily call volume in the data, a longer normalization window would be undesirable. As alternative measures, we also normalize call volume by tower and month (instead of by tower and 20-week window) and we also use 20-week windows that skip periods with drone strikes. Note also that if the 20-week window falls outside of the January 2010-October 2012 period spanned by the call records, we use an asymmetric 20-week window around the week of the strike. We skip dates during Ramadan. Ramadan is treated as a distinct unit since, as due to fasting associated with the Islamic holy month, call patterns are "flipped", and people are more likely to make phone calls at night than during the day. To normalize the call volume during Ramadan, we use 20 days during Ramadan of the same year.

[7] One could argue that this could introduce bias if drone strikes for which we are able to impute the time, or for which we know the approximate time from media sources, had more of an impact on call volume relative to strikes where we do not know the time. In response to this concern, we show that our results are robust to using a 24-hour period with *all* drone strikes (see Table S5).

Measurement of Variables for Anomaly Detection Methods

Given a drone strike, we measure the number of incoming and outgoing calls that are initiated per five-minute interval on the day of the strike.[8] Since the effect of a drone strike on cellphone activity is spatially localized to the area around the strike, we consider cell towers within a fifteen-mile radius of the drone strike.[9] Similar to our fixed effects models, we detect an anomaly by comparing the call volumes the day of the strike to the baseline call volumes of other dates at the location of the strike. Given a strike and a five-minute interval, to create these baseline samples of "normal behavior", we look at the number of calls that are initiated during that five-minute interval of the day, the same day of the week as the day of the strike, over the ten weeks preceding and the ten weeks following the week of the strike, for a total of 20 baseline samples.[10] To declare an anomalous effect in a drone strike, we require that the anomaly persist for at least five consecutive five-minute intervals.[11]

[8] Our data includes the drone strike date, and possibly an estimate of the time of day. The date and time labelling however could be imprecise. For example, if a drone strike is labelled in the dataset as having happened on "May 10," we want to allow for the possibility that (i) it happened very late on May 9, and it was mis-labeled as having happened on May 10, or (ii) it happened during the first few hours on May 11, and it was mis-labeled as having happened on May 10. We note that some strikes are known to have happened during the night. We thus expand our "day" of calls to include 11 p.m. the day before until 3 a.m. the day after. This is a 28-hour period, or 336 five-minute intervals.

[9] We ran tests with all of the following values for the radius: 5, 10, 15, 20, 25 miles. A radius of fifteen miles gave the best detection performance. However, the anomaly detection methods can detect strikes for all of these values for the radius.

[10] This choice ensures a rich enough set of 20 baseline call volumes to compare against, so that the comparison is meaningful. In light of the increasing trend of daily call volume in the data, a longer window of baseline samples would be undesirable. Similarly to the fixed effects analysis, we skip dates during Ramadan, and continue in one week increments/decrements until we get ten weeks in both directions. Since our dataset does not cover calls outside of the window January 2010 - October 2012, if we do not have data for a baseline date in one direction, then we add more baseline dates in the other direction, so we will always total 20 baseline dates of 336 five-minute intervals each.

[11] We ran tests for all integer values between one and ten for the number of consecutive five-minute intervals required to declare an anomaly. Although the anomaly detection methods can detect strikes across this range of values, using five and six consecutive five-minute intervals gave the best detection performance.

Robustness Checks

In this section, we provide more details on our rich set of robustness checks that assess our measurement and specification choices.

Empirical Specification for Panel Models

To estimate the effect that drone strikes have on local call volume, we use a panel dataset with two-way fixed effects that exploits the temporal and spatial variation in drone strikes. We propose the following model:

$$\text{Call\_volume}_{\text{j,t}} = \alpha_{j,t} + \beta_1 \text{ Strike}_{\text{j,t}} + \beta_2 \text{ Strike\_uncertain\_time}_{\text{j,t}}$$

$$+\mathbf{\Phi} \cdot \mathbf{Z_{j,t}} + \gamma_j + \delta_t + \varepsilon_{j,t},$$

where the outcome variable $\text{Call\_volume}_{\text{j,t}}$ is the normalized number of calls made at tower j and time interval t, and the main explanatory variable $\text{Strike}_{\text{j,t}}$ is a dichotomous measure that indicates a drone strike proximate to tower j at time interval t. The model controls for $\text{Strike\_uncertain\_time}_{\text{j,t}}$, a dichotomous measure which captures drone strikes that happened around tower j, on that day, but we are unsure what time they occurred. We include $\mathbf{Z_{j,t}}$, a (column) vector of important covariates including urban density, day of the week, and total casualties. We use $\mathbf{\Phi}$ to denote the corresponding (row) vector of coefficients.

The model also includes γ_j, a fixed effect at the tower level that controls for unobserved characteristics over time that may correlate with drone strikes, such as the demographic composition of an area, and δ_t, a month-year fixed effect, to control for common factors that change over the period, such as a large-scale military offensive. Note that the quantity of interest is β_1, which identifies the average effect of a drone strike on call volume as captured by proximate towers. $\epsilon_{j,t}$ is the error term. We cluster standard errors by tower, as call volume by tower is correlated over time.

Different Model Specifications

Table S2 shows our results for our main specification using an OLS panel two-way fixed effects regression. The results in Model 1 indicate that a drone strike increases the call volume of the surrounding towers by over one fifth of a standard deviation on average. This implies that drone strikes increase call volume by about 850 calls during an eight-hour span for each of the surrounding towers, on average.[12] Thus, the effect is both statistically significant and substantively large, more than three times the size of the average difference in call volume between Sundays (high-calling day) and Fridays (low-calling day).

The results remain significant after we introduce month fixed effects and a covariate for uncertainty of the time of the strike in Model 2, as well as covariates for the day of the week and the number of towers around the strike zone in Model 3.[13]

We also estimate our results using a zero-inflated negative binomial regression model. This is an appropriate model for our call volume data, which exhibits both overdispersion (variance is higher than the mean) and an excess of zeros. The results in Table S3 suggest two key findings. First, drone strikes are associated with higher call volume, which is consistent with our main results. Call volume is about 1.13 times greater (exponent of 0.126) during drone strikes for towers that report a call volume more than zero. Second, drone strikes are associated with higher instances of towers having zero call volume. Indeed, the predicted mean for zero call volume rises from 0.11 for towers not affected by drone strikes to 0.19 for towers impacted by drone strikes.[14]

[12] This calculation is derived by multiplying the coefficient (0.227) by the standard deviation of call volume (3,772), which results in 856.244.

[13] To account for temporal and spatial correlation in drone strikes, we also ran a fixed effects specification with Conley standard errors. Our results were robust to this specification, remaining precisely estimated and statistically significant.

[14] The Vuong test also suggests that the zero-inflated negative binomial model is a significant improvement over a standard Poisson model (z=218.88), implying that drone strikes are causing an unusual amount of dropped towers.

Different Call Volume Measures

In this section, we show that our results are robust to our measurement choices. In Table S4 we use alternative measures of call volume using our preferred specification of Model 2 from Table S2. The results are robust when we use a regular call volume measure without any standardization (Model 1), and when we standardize call volume by tower and month (Model 2), as opposed to by tower and 20-week window. Similarly, our results are robust when we use a 20-week window that omits days with drone strikes (Models 3 and 4), in order to make sure that the baseline samples we compare call volume against are not corrupted by other strikes.

In Model 5 and Model 6, we assess the impact of drone strikes on incoming and outgoing call volume separately. The results indicate that drone strikes have a stronger impact on incoming call volume. This implies that our results are partly driven by people calling to areas affected by drone strikes to check on events rather than people experiencing the strike and calling to inform others.

Table S5 shows that our results are robust to using daily call volume (24-hour periods), instead of using call volume in eight-hour periods, with all drone strikes, including strikes where we are uncertain about the time.

Tower Shutdowns

One concern with our results is that our call volume data contains a large amount of zeros for call volume (almost 17% of the entire sample). It is probable that some of these zeros are due to the fact that there were no phone calls made during the period. However, based on comparable call volume records, it seems likely that the majority of zeros result from towers not operating for hours, days, weeks, or even months at a time.

These downed towers may threaten our results if towers were not shutdown at random, particularly if drone strikes were affecting the proper functioning of cellular towers. For

instance, towers could be shut down prior to drone strikes in order to prevent militants from communicating with one another. Alternatively, cellular towers may have been damaged during drone strikes and not repaired in militant-held areas. While zero call volume in most cases would bias our results on the effect of drone strikes on calls downwards, due to the reduction in call volume, there are some cases where it may bias our results upwards. For example, if towers are shut down prior to drone strikes or afterwards, this would reduce the average level of call volume from our baseline samples, biasing upwards our estimation of the effects of drone strike on call volume.

We handle this concern in a variety of ways. First, we show our results are robust to using a zero-inflated negative binomial regression model (see Table S3). Second, we look for instances of zero call volume in an eight-hour period for all towers within a fifteen-mile radius of a drone strike.[15] We consider the eight-hour drone strike interval, 48 hours prior to the strike, and 48 hours after the strike. We find that for 39 strikes in our data (out of 50 strikes that happened within fifteen miles from a tower and for which we have information on the time of the strike), there were 65 towers which reported at least one period of zero call volume during this period (out of 203 towers within fifteen miles of a drone strike, for which we know the approximate time). While this sounds very high, in some of these cases (eight towers affected by fifteen strikes), the tower was shut down for the entire four-day period. In these instances, it is probable that the tower was down due to circumstances not connected to drone strikes.[16] For the remaining strikes where we see intervals of zero call volume, we find thirteen strikes (nineteen towers) where a tower "dropped" right before or after a strike, implying it was connected to the drone strike.

We rerun our main specification (Model 2 in Table S2), to account for these "suspicious"

[15] Note that this excludes instances where we either do not know the approximate time of the drone strike or where the drone strike happened more than fifteen miles away from any tower.

[16] It is possible that authorities shut down the towers two days or earlier before the strike to prevent communication between militants, and then keep the towers down for another two days, or longer. While we cannot rule out this possibility, the patterns of call volume in our data suggest that towers experience outages for long periods of time that are not connected to drone strikes or other ongoing conflict.

strikes. The results in column "Dropped Towers" in Table S6 indicate that our findings are robust, even after we drop these nineteen locations from our analysis. Overall, these results provide an assurance that our main results are not driven by downed towers.

Drone Strikes During Conflict

Another possible concern is that our estimates are biased when drone strikes occur during conflict. In this case, our model may not identify the impact of drone strikes on call volume, but may instead be affected by the broader impact of violence on call volume (Papadogeorgou et al., 2020). To deal with this concern, we removed all 26 strikes that occurred during the time of the Abyan offensive (May 12, 2012 - June 15, 2012). The results in the "Abyan Offensive" column of Table S6 indicate that our findings are robust, even after we drop these strikes from our analysis.

Drone Strikes During Ramadan

Another potential concern is that drone strikes during Ramadan have more of an impact than other drone strikes. This could be the case if during Ramadan people are more likely to communicate about drone strikes since they are at home fasting during the day or if they are making more phone calls at night, when drone strikes are more likely to occur. Alternatively, our results could be partially driven by the fact that our normalized measure of call volume treats Ramadan as a distinct unit. The "Ramadan" column of Table S6 indicates that our results are robust, even after we exclude the Ramadan periods (as well as the eight strikes that occurred during Ramadan) from our analysis.

Multiple Drone Strikes on the Same Day

Another possible concern is that we have several instances where drone strikes happened on the same day in close proximity. This could bias our results upwards since a binary measure

would be providing an estimate of the cumulative impact of several strikes. To account for this, we rerun our main specification using an aggregate measure of drone strikes: for this variation, variable $\text{Strike}_{j,t}$ in Equation (1) takes integer values, to count the number of distinct drone strikes. There are 20 instances where two or more drone strikes affected the same location, covering a total of 45 strikes. As reported in column "Multiple Strikes" of Table S6, our findings are robust to this effect.

Accounting for the Impact of Bots on Call Volume

One potential concern with our analysis is that the results could be driven by non-human phone numbers ("bots").[17] For instance, an increase in call volume after drone strikes could be the result of an automated phone system that responds to drone strikes, developed by militants or civilians.[18] Including bots in our dataset could skew the results of our analysis and their interpretation. The "Bots" column of Table S6 indicates that our main results are robust to the removal of bots.[19] This implies that the increase in call volume from drone strikes can be attributed to human behavior.

Detecting Bots

In order to remove bots from our dataset, we identify several features that distinguish bots from humans, such as the aggregate number of phone calls, the number of texts, and the duration of phone calls. For each day of the dataset in 2012 and for each phone number, we aggregate the number of outgoing and incoming calls, the number of outgoing and incoming texts, and the total duration of calls.

[17] As noted by several observers, bots have played a large role in social media such as Twitter and Facebook.
[18] For the analysis, we included phone cards as bots. While phone cards are used by humans, one number generates the calls of many individuals, and therefore phone card behavior is very different from human behavior.
[19] In the Detecting Bots section, we describe the process by which we identify and remove bots.

We use k-means clustering, a method of unsupervised learning, on unique phone numbers' call and text features in order to segment phone numbers into categories. For each unique phone number, we calculate the following averages across the days of 2012: duration of phone calls , incoming and outgoing call count , incoming and outgoing text count . We also calculate the number of days in 2012 that the phone number exceeded either of the two thresholds: 10,000 seconds of calls a day, 1,000 texts a day.

Clustering yielded three main clusters, which were robust across different input values for the desired number of clusters. One large cluster contained phone numbers with call durations of around 50,000 seconds per day and extreme incoming count to outgoing count ratios. Another cluster contained phone numbers with call durations of around 10,000 seconds per day. The third cluster contained phone numbers with few or no phone calls.

Each of the three main clusters captures a portion of the bots, but also many seemingly real phone numbers. In order to minimize false classifications of real numbers as bots in each cluster, we further used the following filters to classify numbers as bots:

1. has an average of more than 50,000 seconds of calls per day, has exceeded 50,000 seconds of calls per day for 100 days in 2012, and has a ratio of incoming to outgoing calls (or vice versa) of less than 0.1,

2. has an average of more than 10,000 seconds of calls per day, has an average of more than 50 total phone calls per day, has a ratio of incoming to outgoing calls (or vice versa) of less than 0.1, and has a ratio of incoming to outgoing texts (or vice versa) of less than 0.1 (or has not used texts at all), or

3. has over 500 total calls on some day, or

4. has more than 86,400 seconds of phone calls on some day, or

5. has over 2,000 texts on some day and[20] has degree in the network higher than 1,000,

[20] Filter "has over 2,000 texts on some day", without the condition on the degree being higher than 1,000,

6. has an average of less than ten seconds of calls per day and has a ratio of incoming to outgoing calls (or vice versa) of less than 0.1.

These decision rules allow us to filter several specific types of bots. The total amount of phone numbers removed during our bot removal process is 4044 (responsible for over 120 million calls). Although there are likely additional bots in the network, we have removed the worst offenders.

Description of our Anomaly Detection Methods

In this section, we discuss in more detail the tests we run with the proposed anomaly detection methods. We test three anomaly detection methods on the classification task of deciding between strikes and non-strikes, based only on the volume of calls, and without knowledge of whether a strike happened or not. All the proposed anomaly detection methods have area under the ROC curve (AUC) higher than 0.70, establishing a good classification performance (Figure S1, Figure S2).

Description of Anomaly Detection Methods

We propose an anomaly detection method that is based on comparing the instances under examination against the empirical distribution of observations. For robustness, we also employ two additional anomaly detection methods that have been used in the literature: anomaly detection based on density estimation using kernel functions; and a technique that declares anomalies based on a simple threshold rule on the deviation from normal activity, which has been used specifically for identifying events from volume of cellphone activity (Bagrow et al., 2011; Gao et al., 2014).

would filter out a high number of real numbers. Similarly, filter "has degree higher than 1000", without the condition of over 2,000 texts on some day, would also filter out a high number of real numbers. Combining the two conditions removed many bots, while not removing too many real numbers.

Method 1 — Anomaly Detection Based on the Empirical Distribution

For each five-minute interval, we have 20 baseline samples of volume of communications; these form the empirical distribution for the volume of communications under normal behavior. For each strike and for each five-minute interval, we compare the observed value against the formed empirical distribution of the 20 samples. We call the observation an anomaly if it lies in the $\alpha \cdot 100\%$ right tail of the empirical distribution, that is, if

$$\frac{\text{number of baseline samples with value} \geq \text{observation}}{20} \leq \alpha,$$

where α is an input parameter that we fine-tune.

Method 2 — Anomaly Detection Based on Density Estimation

Another approach to anomaly detection is to first estimate a probability density from the data under normal behavior, and then employ any parametric or non-parametric method for anomaly detection on the estimated density. The density is approximated as the average of evaluations of a kernel function using the observed samples: if $(x_1, \ldots, x_n)$ is the realization of a sample drawn i.i.d. from the unknown distribution with density f, then the kernel density estimator of f is

$$\hat{f}_h = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i),$$

where $K(\cdot)$ is the kernel and $h > 0$ a parameter called the bandwidth.[21] We make use of a kernel density estimation implementation Botev et al. (2010), which uses the Gaussian kernel $K_\gamma(x - x') = e^{-\gamma(x-x')^2}$ and optimizes for the bandwidth parameter γ.

We first estimate a density from the 20 baseline samples using kernel density estimation. We then call the observation an anomaly if it lies in the $\alpha \cdot 100\%$ right tail of the estimated

[21] This is known as Parzen–Rosenblatt window estimation (Rosenblatt, 1956; Parzen, 1962).

distribution, that is, if

$$1 - \hat{F}(\text{observation}) \leq \alpha,$$

where $\hat{F}$ is the cumulative distribution function of the estimated density $\hat{f}$, and where α is an input parameter that we fine-tune.

Method 3 — Anomaly Detection Based on a Deviation Threshold

A simpler approach to anomaly detection is to declare an anomaly based on a threshold rule on the deviation of the observation under examination from normal activity. We employ the following rule, used by others (Bagrow et al., 2011; Gao et al., 2014) to identify emergency (such as bombings and earthquakes) and non-emergency events (such as concerts and sporting events) through aggregate cellphone activity.

We declare the observation an anomaly if the z-value of the observation satisfies

$$\frac{\text{observation - average(baseline samples)}}{\text{St.Deviation(baseline samples)}} > z_{thr},$$

where z_{thr} is an input parameter that we fine-tune.

Testing the Performance of our Anomaly Detection Methods

We start with a dataset of 108 U.S. drone strikes and other covert actions between January 2010 and October 2012, documented by either of two sources: the New America Foundation and the Bureau of Investigative Journalism. In order to assess the performance of the proposed anomaly detection techniques, we run a series of tests on subsets of the 108 strikes.

We can run different tests for our anomaly detection methods by changing the radius from a strike within which we are looking for evidence of anomalies in the calls as captured by surrounding cell towers. We run our tests for all of the following values of the radius:

$r = 5, 10, 15, 20, 25$ miles.[22]

To run meaningful tests, given a selection of mile radius r, we first remove the following strikes from our set of strikes:

- strikes that have zero call volume (within radius of r miles) for the strike and for all the baseline samples;
- strikes that have zero call volume (within radius of r miles) for the strike, but have non-zero call volume for all the baseline samples.

We also take special care to treat pairs of duplicate strikes as a single strike. We treat two strikes that happen on the same day as duplicate if there are towers with non-zero call volume that are within a radius of r miles from both strikes. We run tests for all of $r = 5, 10, 15, 20, 25$ miles. If two strikes are duplicate, then we treat them as one.[23] Else, we keep them as distinct strikes. When we take multiple strikes as one, we report detection if at least one of them is detected. Else, we report no detection. This way of counting is conservative and ensures that we are not counting multiple detected strikes, when only one is really causing the anomaly in the calls.

After applying the aforementioned initial filters on the set of strikes, we run our anomaly detection methods on a series of different inputs, for robustness. We run different tests by varying the set of strikes that we run the anomaly detection methods on. We vary the set of strikes across the following dimensions:

- we run tests with and without strikes with dropped towers within radius of r miles from the strike.
- we run tests with and without strikes with call volume (within radius of r miles) lower than ten calls across all five-minute intervals of the strike day;

[22] We choose 25 miles as our cutoff since according to our panel analysis the effects of drone strikes on call volume weaken after 25 miles.

[23] Notice that for this to happen, we need the distance of the strikes to be less than $2r$ miles. Therefore, if the distance between the two strikes is greater than $2r$ miles, we keep them distinct.

- we run tests with and without strikes that happened during Ramadan;
- we run tests including all strikes that happened during the Abyan offensive (May 12, 2012 - June 15, 2012);[24] we also run tests excluding the strikes that happened between May 12, 2012 and June 15, 2012 in Jaar, Zinjibar, and Shuqrah, which were the most heavily affected regions in the Abyan offensive; and we also run tests excluding all strikes that happened between May 12, 2012 and June 15, 2012, across Yemen;
- we run tests with and without the strikes for which we do not know the time of the day when they happened.

Given a set of strikes, we can vary the composition of the set of baseline samples that we input into the anomaly detection methods. The baseline samples capture call volume on the same day of the week as the strike day, in the weeks preceding and following the strike. We want the baseline samples to be uncorrupted by other strikes in the dataset. We thus run tests with a moderate "collision" check: when testing for detection of a strike s, we skip a baseline day if another strike t happened on that day, as long as strike t happened within r miles from a tower which is within r miles from strike s. We also run tests with a conservative "collision" check: when testing for detection of a strike s, skip a baseline day if another strike t happened on that day, or the previous day, or on the following day, as long as strike t happened within r miles from a tower which is within r miles from strike s. In both cases, when we skip a baseline day, we replace it with the same day of the week from another week. We also run tests with no "collision" check.

Finally, we run tests varying the duration threshold above which we declare an anomaly: we test all thresholds of $1, 2, 3, 4, 5, 6, 7, 8, 9$, and 10 consecutive five-minute intervals.

We note that we run tests with all possible combinations of the above dimensions, both to identify the settings that result in the best detection performance, and to test the robustness

[24] During this period, there was a large military operation by the Yemeni military against Islamist militant forces that included many drone strikes.

of the detection methods.

In order to evaluate our detection methods under the different variations proposed above, we test them on the classification task of detecting anomalies on a balanced pool of dates that includes strike dates and non-strike dates. In particular, for each strike, we test our detection methods both on the strike day (using the location where the strike happened), as well as on the day four days prior to the strike day (again using the location where the strike happened), assuming no strike happened on that day in that location.

To evaluate the performance of the anomaly detection methods on telling apart strike days from non-strike days, we use ROC (Receiver Operating Characteristic) curves, which are formed by plotting the true positive rate against the false positive rate, as the detection parameter (α for Methods 1 and 2, z_{thr} for Method 3) varies.[25] The area under the ROC curve (AUC) is a measure of the classification performance, with higher area corresponding to more accurate classification. A comparison of the ROC curve of a classifier (and the corresponding AUC) against the 45-degree line showcases the advantage of the classifier over a classifier which guesses at random. Figure S1 shows the ROC curves and the AUC for the setting that results in the highest average AUC across the three detection methods.

This setting is the following: radius of $r = 15$ miles, removing strikes with dropped towers, removing strikes with call volume lower than ten calls across all five-minute intervals of the strike day, removing strikes that happened during Ramadan, not removing strikes during the Abyan offensive, removing strikes for which we do not know the time when they happened, using a conservative collision check, and using a duration threshold of six consecutive five-minute intervals to declare an anomaly.

We also show in Figure S2 the ROC curves for the setting that results in the highest average AUC across the three detection methods, subject to having more than 35 strikes

[25] The values for the parameters used are as follows. Method 1: α starts at 0.05 and moves in increments of 0.05 until 0.95. Method 2: α takes the values 0.00001, 0.0001, 0.001, 0.01 and then moves in increments of 0.01 until 0.4. Method 3: z_{thr} takes the values 0.00001, 0.0001, 0.001, 0.01, 0.1, and then moves in increments of 0.1 until 5.0.

that remain after applying all the filters. This setting is the following: radius of $r = 15$ miles, not removing strikes with dropped towers, removing strikes with call volume lower than ten calls across all five-minute intervals of the strike day, removing strikes that happened during Ramadan, removing strikes that happened during the Abyan offensive in Jaar, Zinjibar, and Shuqrah, removing strikes for which we do not know the time when they happened, using a moderate collision check, and using a duration threshold of five consecutive five-minute intervals to declare an anomaly.

Using the ROC curves, we can also decide on the detection threshold parameters that balance sensitivity and specificity. We report the detection parameters that yield the best true positive rate subject to maintaining a false positive rate of at most 40% (30%). This results in the following choices: $\alpha = 0.15$ (0.1) for Method 1, $\alpha = 0.27$ (0.22) for Method 2, $z_{thr} = 0.8$ (1.3) for Method 3. When restricting to a number of strikes larger than 35, then the detection parameters that yield the best true positive rate subject to maintaining a false positive rate of at most 40% (30%) are $\alpha = 0.15$ (0.05) for Method 1, $\alpha = 0.18$ (0.14) for Method 2, $z_{thr} = 0.9$ (1.5) for Method 3.

Contextual Impact of Drone Strikes

In this section, we provide more details on the contextual impact of drone strikes by considering how unique the effects of drone strikes are on call volume as compared to other related shocks in Yemen, and by studying the heterogeneous impact of drone strikes on call volume, depending on the number and type of targets.

Comparison to Non-Drone Strike Events

We compare drone strikes to four other types of events. Two of the types we study are violent: al-Qaeda attacks (2011-2012) and the bombing of the Presidential Palace during

the Arab Spring events in 2011 (June 3, 2011). These events help highlight how similar or different drone strikes are compared to other violent events. The other two types of events are non-violent: the religious holiday of Eid al-Fitr (2010-2012), marking the end of the holy month of Ramadan[26], and the 2010 FIFA World Cup final (July 11, 2010). These instances help highlight how drone strikes compare to important religious and cultural events, which produce large shocks to the call volume network. Drone strikes are violent, localized, and instantaneous events. The presidential bombing is an event with the same attributes as drone strikes, while the al-Qaeda attacks spread over several days. The non-violent events we consider are not localized, and the World Cup final is instantaneous while Eid al-Fitr is not, as it is celebrated over the whole day. We next provide some background on these events.

Al-Qaeda attacks refer to actions taken by the group when they took over many rural towns like Lawdar, Zinjibar, and Jaar between 2011-2012. We have the geolocation and date of ten such strikes throughout Yemen. These attacks range from minor conflicts at security checkpoints with few casualties, to large scale attacks over territorial control with significant casualties. These al-Qaeda attacks are violent and localized, just like drone strikes. However, they are part of militant actions that unfolded over the course of several days.

The Presidential Palace bombing refers to a bombing at the Presidential Palace on June 3, 2011, wounding President Saleh, who reportedly had a collapsed lung and burns on 40 percent of his body.[27] This event occurred during the height of the Yemeni Revolution (Arab Spring), and in its aftermath Saleh escaped the country for three months. The Presidential Palace is located in the heart of the capital city, Sana'a, with a population of around two million people. This bombing caused chaos within the city, and was a violent, instantaneous, and localized shock to the patterns of communication.

[26] We have three such instances in our dataset. Eid al-Fitr 2010 began in the evening of Thursday, September 9 and ended in the evening of Friday, September 10. Eid al-Fitr 2011 began in the evening of Monday, August 29 and ended in the evening of Tuesday, August 30. Eid al-Fitr 2012 began in the evening of Saturday, August 18 and ended in the evening of Sunday, August 19.

[27] See `http://www.nytimes.com/2011/06/04/world/middleeast/04yemen.html`.

Eid al-Fitr, which translates to "feast of breaking the fast", is a religious holiday celebrated by Muslims that marks the end of the fasting period of Ramadan. This event is a non-violent, non-instantaneous shock with widespread effect all over Yemen.

Soccer is highly popular in Yemen. The final of the 2010 FIFA World Cup was played on July 11, 2010 with Spain defeating the Netherlands, 1-0 in extra time. The game, including regular time, extra time, and time allowed for intervals and additional time, spanned about two and a half hours, and there are various shocks to the volume of calls triggered by events during the game such as opening kickoff, end of regulation, and the goal scored during the extra time. These shocks were non-violent, instantaneous, and widespread throughout Yemen.

To examine these instances of events, we use our anomaly detection methods, which are well suited for studying the impact of rare and uncommon events on the patterns of communication. For each of these events, the date of the event is known, and is used by the anomaly detection methods. For the al-Qaeda attacks and the Presidential Palace bombing, for which we also have data on the location, we employ our detection methods using the known geo-coordinates. For Eid al-Fitr and the 2010 FIFA World Cup final, occasions which have no precise geographic location in Yemen, we employ our detection methods using the locations of the drone strikes in our dataset. By focusing on the locations of drone strikes, we are able to compare the effect of the non-violent events against the effect of a drone strike, controlling for location.

Out of the ten al-Qaeda attacks, we only detect anomalies for three of them — a lower detection rate compared to drone strikes. The two attacks with the largest anomalies were when al-Qaeda took over Zinjibar (May 27, 2011) over two days of heavy fighting, leaving sixteen dead (see Figure 5), and when the Yemeni army took back control of Zinjibar a few months later (September 10, 2011). The other detected attack was when Yemen's security authorities struck al-Qaeda killing fifteen and arresting others in Abyan and Marib

governorates (March 24, 2011).

This implies that we can detect attacks with high impact, but it is hard to detect lower impact attacks, as there would not be enough people nearby to have been directly affected.[28]

The Presidential Palace bombing resulted in a clear anomaly in the volume of calls (see Figure 5), which all three proposed anomaly detection methods detect, and which lasted for around eight hours. In contrast, the drone strike peaks last for at most only a few hours. The palace bombing was in Sana'a, the capital city, and resulted in about 130 casualties. Simultaneously, the bombing caused a surge in mobility near and around the Presidential Palace.

The long-lasting commotion in a city of a large population contributed to the longer effect in the volume of calls as compared to the effect from a typical drone strike: both a drone strike and the Presidential Palace bombing are instantaneous shocks to communications, yet the repercussions of the palace bombing on communications lasted quite longer.

We next discuss the non-violent events. We compare the call volume on Eid al-Fitr against the call volume on 20 weekdays (i.e., excluding Fridays) during Ramadan of the same year. Across the three years that we have call data on (2010, 2011, 2012), people make a lot more calls on Eid al-Fitr, and during the entire day, compared to a typical Ramadan day, as shown in Figure 5. Importantly, the effect of this important religious holiday is widespread throughout Yemen.

We also compare the call volume on the 2010 FIFA World Cup final day against the call volume on 20 baseline days, at the locations where the drone strikes happened. We test the performance of the anomaly detection methods on the classification task of detecting anomalies on a balanced pool of "game" and "non-game" instances. Anomaly detection Method 3, when employed with a threshold of $z_{thr} = 1.3$, correctly detects anomalies in 82.9% of the "game" instances, while the false positive rate on "non-game" instances remains

[28] We also estimate the impact of al-Qaeda attacks on call volume using panel fixed effects models. The estimates indicate small and insignificant effects (possibly because we only have data on ten attacks).

at 36.2%. The match started at or shortly after 21:30 local time and ended after 90 minutes of regular time, 30 minutes of extra time, as well as some time allowed for intervals and additional time. The only goal was scored in minute 116 of the game. In the plot in Figure 5 we see peaks in the volume of calls that correspond to the start of the game, the end of regular time, and the end of the game. Similarly to drone strikes, this highly popular event has a temporally sharp effect on the volume of calls. Differently from drone strikes, the effect, as detected by our anomaly detection methods at the various strike locations, is widespread throughout Yemen.

Effect of Drone Strikes and Impact Heterogeneity

It seems likely that the effects of drone strikes on call volume will depend partly on the nature and number of targets. To account for the possible heterogeneous effects of drone strikes on call volume, we rerun our main specification with a series of interaction models in Figure 3. Specifically, we measure the impact of the strike using factors such as the total number of casualties, militant casualties, civilian casualties, and militant rank of the target.

Our results already reported in the main text suggest that one possible factor driving call volume after drone strikes is the overall impact which the strike had on militants in the area. We also considered other strike attributes such as time of day and urban density, and found that strikes earlier in the day are more likely to have a larger impact.

References

Aharony, N., W. Pan, C. Ip, I. Khayal, and A. Pentland (2011, dec). Social fMRI: Investigating and Shaping Social Mechanisms in the Real World. *Pervasive Mobile Computing 7*(6), 643–659.

Altshuler, Y., Y. Elovici, A. B. Cremers, N. Aharony, A. Pentland, and M. Cebrian (2013). *Stealing Reality: When Criminals Become Data Scientists (or Vice Versa)*, pp. 133–151. Springer New York.

Altshuler, Y., M. Fire, E. Shmueli, Y. Elovici, A. Bruckstein, A. S. Pentland, and D. Lazer (2013). Detecting Anomalous Behaviors Using Structural Properties of Social Networks. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, Berlin, Heidelberg, pp. 433–440. Springer-Verlag.

Altshuler, Y., W. Pan, and A. S. Pentland (2012). Trends Prediction Using Social Diffusion Models. In *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'12, Berlin, Heidelberg, pp. 97–104. Springer-Verlag.

Bagrow, J. P., D. Wang, and A.-L. Barabási (2011). Collective Response of Human Populations to Large-Scale Emergencies. *PLoS ONE 6*(3).

Blumenstock, J., M. Callen, and T. Ghani (2014). Violence and Financial Decisions: Evidence from Mobile Money in Afghanistan. *Mimeo*.

Blumenstock, J. E. (2012a). *Essays on the Economic Impacts of Mobile Phones in Sub-Saharan Africa*. {PhD} dissertation, Berkeley University.

Blumenstock, J. E. (2012b). Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development 18*(2), 107–125.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science 353*(6301), 753–754.

Blumenstock, J. E., N. Eagle, and M. Fafchamps (2016). Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters. *Journal of Development Economics*.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 489*(7415), 295–298.

Botev, Z. I., J. F. Grotowski, and D. P. Kroese (2010). Kernel density estimation via diffusion. *Annals of Statistics 38*(5), 2916–2957.

Calabrese, F., Z. Smoreda, V. D. Blondel, and C. Ratti (2011). Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. *PLoS ONE 6*(7), e20814.

Candia, J., M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical 41*(22), 224015.

Dafoe, A. and J. Lyall (2015). From cell phones to conflict? Reflections on the emerging ICT-political conflict research agenda. *Journal of Peace Research 52*(3), 401–413.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem (2014, nov). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*(45), 15888–15893.

Eagle, N., Y. de Montjoye, and L. M. A. Bettencourt (2009, aug). Community Computing: Comparisons between Rural and Urban Societies Using Mobile Phone Data. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, Volume 4, pp. 144–150.

Eagle, N., A. S. S. Pentland, and D. Lazer (2009, sep). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences 106*(36), 15274–15278.

Finger, F., T. Genolet, L. Mari, G. C. de Magny, N. M. Manga, A. Rinaldo, and E. Bertuzzo (2016, jun). Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. *Proceedings of the National Academy of Sciences of the United States of America 113*(23), 6421–6.

Gao, L., C. Song, Z. Gao, A.-L. Barabási, J. P. Bagrow, and D. Wang (2014). Quantifying Information Flow During Emergencies. *Scientific Reports 4*.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi (2008). Understanding individual human mobility patterns. *Nature 453*(7196), 779–782.

Jamal, A. A., R. O. Keohane, D. Romney, and D. Tingley (2015). Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses. *Perspectives on Politics 13*(01), 55–73.

Jiang, Z.-Q., W.-J. Xie, M.-X. Li, B. Podobnik, W.-X. Zhou, and H. E. Stanley (2013, jan). Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences of the United States of America 110*(5), 1600–5.

Kenett, D. Y. and J. Portugali (2012, jul). Population movement under extreme events. *Proceedings of the National Academy of Sciences of the United States of America 109*(29), 11472–3.

King, G. and M. E. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review 107*(02), 326–343.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science 343*(6176), 1203–1205.

Lin, Y.-R. and D. Lazer (2011). The effect of social contexts on network response to emergencies. Technical report.

Mitts, T. (2019). From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West. *American Political Science Review 113*(1), 173–194.

Monroe, B. L., J. Pan, M. E. Roberts, M. Sen, and B. Sinclair (2015). No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science. *PS: Political Science & Politics 48*(01), 71–74.

Nagler, J. and J. A. Tucker (2015). Drawing Inferences and Testing Theories with Big Data. *PS: Political Science & Politics 48*(01), 84–88.

Palla, G., A.-L. Barabasi, and T. Vicsek (2007). Quantifying social group evolution. *Nature 446*(7136), 664–667.

Papadogeorgou, G., K. Imai, J. Lyall, and F. Li (2020). Causal inference with spatio-temporal data: Estimating the effects of airstrikes on insurgent violence in iraq. *arXiv preprint arXiv:2003.13555*.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics 33*(3), 1065–1076.

Pierskalla, J., N. H. F. Hollenbach, and F. M. HOLLENBACH (2013). Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa. *American Political Science Review 107*(02), 207–224.

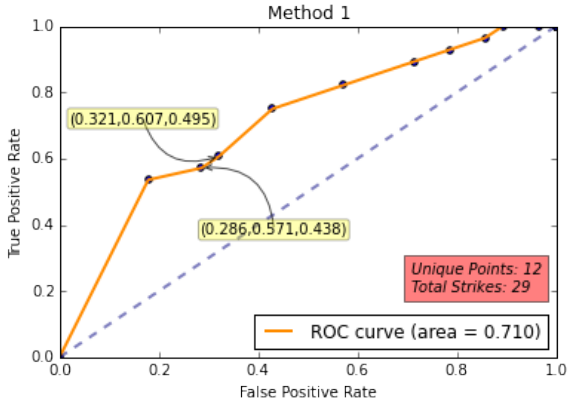
Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics 27*(3), 832–837.

Shapiro, J. N. and N. B. Weidmann (2015). Is the Phone Mightier Than the Sword? Cellphones and Insurgent Violence in Iraq. *International Organization 69*(02), 247–274.

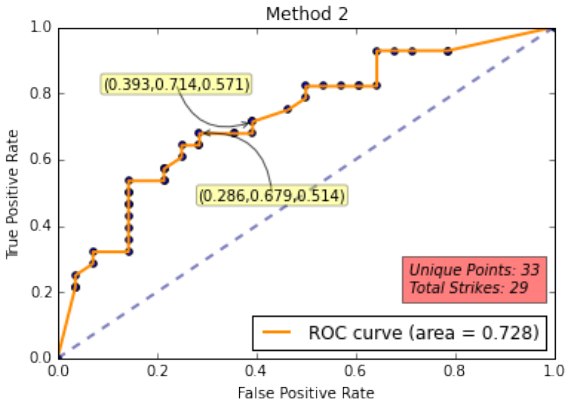
Smith, C., A. Mashhadi, and L. Capra (2013). Ubiquitous Sensing for Mapping Poverty in Developing Countries.

Weidmann, N. B. (2015). Communication, technology, and political conflict. *Journal of Peace Research 52*(3), 263–268.

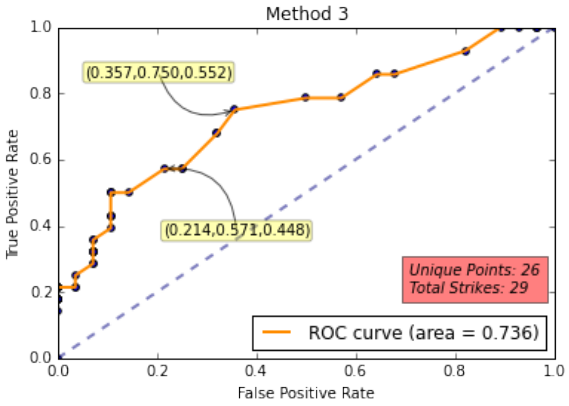
Zammit-Mangion, A., M. Dewar, V. Kadirkamanathan, and G. Sanguinetti (2012, jul). Point process modelling of the Afghan War Diary. *Proceedings of the National Academy of Science 109*, 12414–12419.



FPR ≤ 0.3: $\alpha = 0.1$. FPR ≤ 0.4: $\alpha = 0.15$

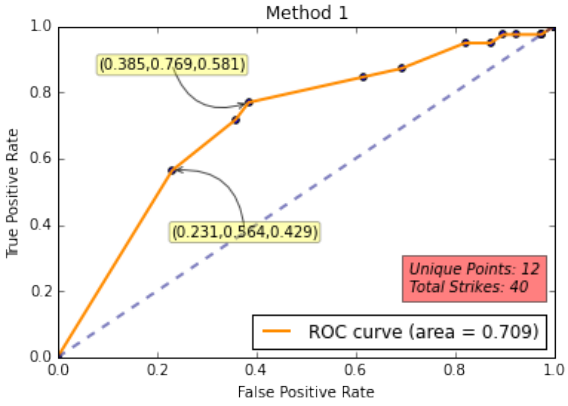


FPR ≤ 0.3: $\alpha = 0.22$. FPR ≤ 0.4: $\alpha = 0.27$

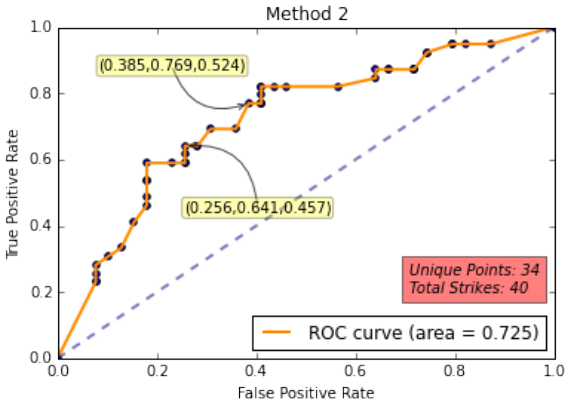


FPR ≤ 0.3: $z_{thr} = 1.3$. FPR ≤ 0.4: $z_{thr} = 0.8$

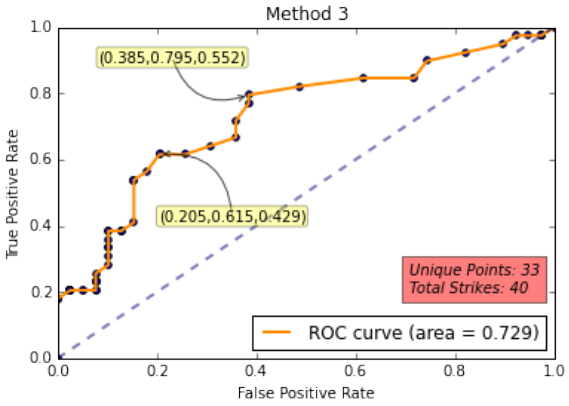
Figure S1: ROC Curves and Selection of Detection Thresholds. We show this for all methods, for the setting with the highest average area under the curve (AUC) across the three methods. The labeled points on the ROC curves correspond to the parameter settings that attain the best true positive rate (the second number in the triplet) subject to false positive rate (the first number in the triplet) of at most 30% or 40%. The third number in the triplet is the detection rate on all strikes from the dataset that remain after we remove (i) strikes that have zero call volume (within radius of fifteen miles) for the strike and for all the baseline samples; and (ii) strikes that have zero call volume (within radius of fifteen miles) for the strike, but have non-zero call volume for all the baseline samples.



FPR ≤ 0.3: $\alpha = 0.05$. FPR ≤ 0.4: $\alpha = 0.15$



FPR ≤ 0.3: $\alpha = 0.14$. FPR ≤ 0.4: $\alpha = 0.18$



FPR ≤ 0.3: $z_{thr} = 1.5$. FPR ≤ 0.4: $z_{thr} = 0.9$

Figure S2: ROC Curves and Selection of Detection Thresholds - Over 35 Strikes. We show this for all methods, for the setting with the highest average area under the curve (AUC) across the three methods, subject to having more than 35 strikes that remain after applying all the filters.The labeled points on the ROC curves correspond to the parameter settings that attain the best true positive rate (the second number in the triplet) subject to false positive rate (the first number in the triplet) of at most 30% or 40%. The third number in the triplet is the detection rate on all strikes from the dataset that remain after we remove (i) strikes that have zero call volume (within radius of fifteen miles) for the strike and for all the baseline samples; and (ii) strikes that have zero call volume (within radius of fifteen miles) for the strike, but have non-zero call volume for all the baseline samples.

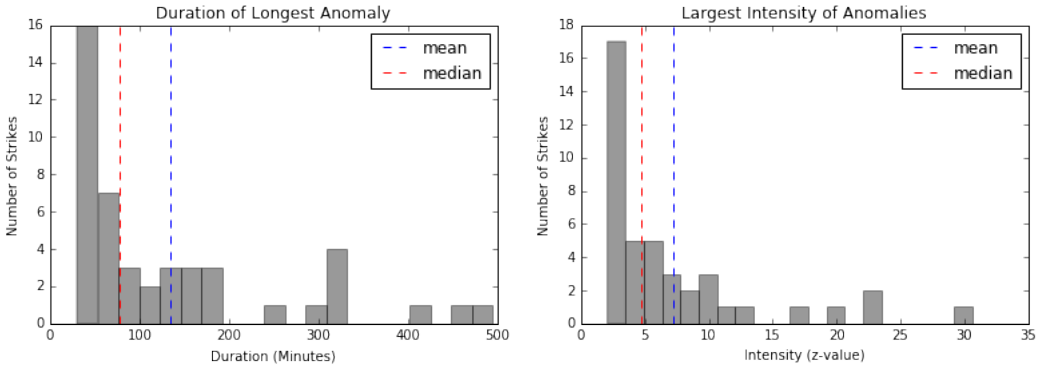


Figure S3: Histograms for the Duration and Intensity of Call Volume Anomalies due to Drone Strikes.

Duration of the longest anomaly in the volume of calls; and intensity of the greatest anomaly, measured as the largest call volume deviation (z-value) in the streaks of anomalous five-minute intervals that are at least six five-minute intervals long. We use all strikes that are detected using anomaly detection Method 3 and parameter $z_{thr} = 1.3$. Method 3 is the method that yields the largest area under the ROC curve (AUC). $z_{thr} = 1.3$ is the threshold that attains the best true positive rate subject to a false positive rate of at most 30%.

Table S1: Descriptive Statistics for Measures of Drone Strikes

Statistic	N	Mean	St. Dev.	Min	Max
Two Data Sources	108	0.52	0.50	0	1
Morning Strike	108	0.18	0.38	0	1
Midday Strike	108	0.08	0.28	0	1
Evening Strike	108	0.21	0.41	0	1
Time of Strike Uncertain	108	0.53	0.50	0	1
Latitude	108	14.15	0.98	12.81	17.18
Longitude	108	45.99	1.05	43.69	49.17
Distance to Nearest Tower (miles)	108	3.89	7.01	0.04	54.25
Total Killed (high estimate)	108	9.47	9.27	0	43
Total Killed (low estimate)	108	6.95	7.16	0	38
Civilians Killed (high estimate)	108	0.88	3.56	0	26
Civilians Killed (low estimate)	108	0.46	1.92	0	14
Militants Killed (high estimate)	63	9.37	9.26	0	43
Militants Killed (low estimate)	63	7.16	7.17	0	38
High-Level Target	55	0.40	0.49	0	1
Militant Facility	49	0.16	0.37	0	1

Notes: The data on drone strikes in Yemen are from two sources: the New America Foundation and the Bureau of Investigative Journalism. 'Two Data Sources' indicates that the strike was mentioned in each of these data sources.

Table S2: Impact of Drone Strikes on Call Volume

	Dependent Variable: Call Volume		
	Model 1	**Model 2**	**Model 3**
Drone Strike	0.227^{***}	0.199^{***}	0.154^{**}
	(0.063)	(0.064)	(0.064)
Drone Strike Uncertain Time		0.024^{***}	-0.032^{***}
		(0.009)	(0.010)
Urban Density (Number of Towers)			0.00001
			(0.00002)
Total Killed			0.005^{***}
			(0.001)
Sunday			0.016^{***}
			(0.001)
Monday			0.010^{***}
			(0.001)
Tuesday			0.018^{***}
			(0.001)
Wednesday			0.005^{***}
			(0.002)
Thursday			0.014^{***}
			(0.001)
Saturday			0.022^{***}
			(0.001)
Tower FEs	Yes	Yes	Yes
Month FEs	No	Yes	Yes
Observations	3,045,758	3,045,758	3,045,758
Adj. R-squared	0.001	0.029	0.029

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$

Notes: Standard errors are clustered by tower. Call volume is standardized by tower using a 20-week window for a baseline where we compare the number of calls during an eight-hour interval against the number of calls made the same day of the week and during the same time interval for the ten preceding and ten following weeks.

Table S3: Impact of Drone Strikes on Call Volume, Zero-inflated Negative Binomial Model

	Dependent Variable:	
	Call Volume	**Zero Call Volume**
Drone Strike	0.126***	0.619**
	(0.032)	(0.315)
Tower FEs	Yes	Yes
Month FEs	Yes	Yes
Observations	631,154	631,154
Non-zero observations	557,011	557,011
Zero observations	74,143	74,143

***p < .01; **p < .05; *p < .1

Notes: Standard errors are clustered by tower. Results generated using the zero-inflated negative binomial regression command (**zinb**) in Stata. For more details on this command, see `https://stats.idre.ucla.edu/stata/dae/zero-inflated-negative-binomial-regression/`

Table S4: Impact of Drone Strikes on Alternative Measures of Call Volume

	Dependent Variable: Call Volume					
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Drone Strike	519.604*	0.322***	0.207***	0.220***	0.229***	0.114*
	(284.024)	(0.073)	(0.066)	(0.066)	(0.064)	(0.062)
Uncertain Time	−48.554	0.173***	0.025***	0.053***	0.022**	0.023***
	(43.099)	(0.008)	(0.009)	(0.011)	(0.009)	(0.009)
Tower FEs	Yes	Yes	Yes	Yes	Yes	Yes
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,497,957	2,974,209	3,052,005	3,065,652	3,045,299	3,045,296
Adj. R-squared	0.630	−0.0003	0.033	0.040	0.028	0.028

***p < .01; **p < .05; *p < .1

Notes: Standard errors are clustered by tower. Unnormalized call volume (incoming and outgoing) is used in Model 1. Call volume is standardized by tower and month (as opposed to by tower and 20-week window) in Model 2. Call volume is standardized using a 20-week window that skips days when a drone strike happened in Model 3. Model 4 uses a 20-week window that skips days when a drone strike happened, or a drone strike happened the previous day, or a drone strike happened the following day. Model 5 uses standardized incoming calls (20-week window). Model 6 uses standardized outgoing calls (20-week window).

Table S5: Impact of Drone Strikes on Daily Call Volume (24-Hour Periods)

	Dependent Variable: Daily Call Volume	
	Strikes with Time	**All Strikes**
Drone Strike	0.140**	0.154**
	(0.058)	(0.061)
Tower FEs	Yes	Yes
Month FEs	Yes	Yes
Observations	1,018,718	1,018,718
Adj. R-squared	0.038	0.038

***p < .01; **p < .05; *p < .1

Notes: See Table S2. Standard errors are clustered by tower. Call volume is based on 24-hour intervals.

Table S6: Additional Identification Concerns

	Dependent Variable: Call Volume				
	Dropped Towers	**Abyan Offensive**	**Ramadan**	**Multiple Strikes**	**Bots**
Drone Strike	0.205***	0.207***	0.188***	0.262***	0.208***
	(0.073)	(0.062)	(0.068)	(0.077)	(0.064)
Uncertain Time	0.025***	0.016	−0.021**	0.024***	0.023**
	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)
Tower FEs	Yes	Yes	Yes	Yes	Yes
Period FEs	Yes	Yes	Yes	Yes	Yes
Observations	3,002,048	2,932,659	2,777,191	3,045,758	3,045,737
Adj. R-squared	0.029	0.029	0.035	0.029	0.029

***p < .01; **p < .05; *p < .1

Notes: See Table S2. Standard errors are clustered by tower.