Supporting Information for "Violence against civilians in civil war: Evidence from Syria"

July 4, 2018

Contents

1 Data on Civilian Casualities

I compile data on Syrian civilian causalities from two sources. Both sources are collected by Syrian NGOs and made available online.

1.1 Shuhada

The Syrian Shuhada (Martyrs) dataset compiles data from several sources on casualties in Syria. The dataset, available in both English and Arabic, is available at http://syrianshuhada.com. I scraped the English and Arabic entries for each casualty in the dataset and compiled a dataset of around 150,000 casualties. In its original format, the dataset reports information on

- name
- age
- combatant status
- date of death
- province, city, and neighborhood of birth
- province, city, and neighborhood of death
- cause of death
- source of the report (e.g. other NGO report, Facebook, etc.)

These fields are often left unfilled, especially the age and neighborhood of birth or death fields. The information provided in the English and Arabic datasets is identical. Comments and source information are only provided in Arabic across the two versions. The names of casualties and place names in the English dataset are often simple transliterations of the Arabic forms, without vowels (e.g. "Mnbj" vs. "Manbij"). For that reason, working with Arabic place names is much easier.

1.2 Violations Documentation Centre data

As a second source of casualty data, I use the Violations Documentation Centre dataset[1]. The VDC dataset has a more transparently documented process for how casualties are recorded in the dataset. Unfortunately, the geographic information on the locations of deaths is not nearly as complete as the data available in the Shuhada dataset.

The 35 or so activists at the Centre who maintain the dataset gather information themselves or "reliable sources like field hospitals, cemeteries, casualties' families and some of the media centers."[2] The initial reports are then augmented with more details, videos, or photos, and added to the dataset. Finally, the reports are sent back to the field to be validated and updated by local activists.[3] The dataset reports deaths of non-Syrian army soldiers from March 2011 through , of which 151,873 are coded as civilians. The documentation

[1] Available at http://www.vdc-sy.info/

[2] http://vdc-sy.net/Website/?page_id=849

[3] http://vdc-sy.net/Website/?page_id=849

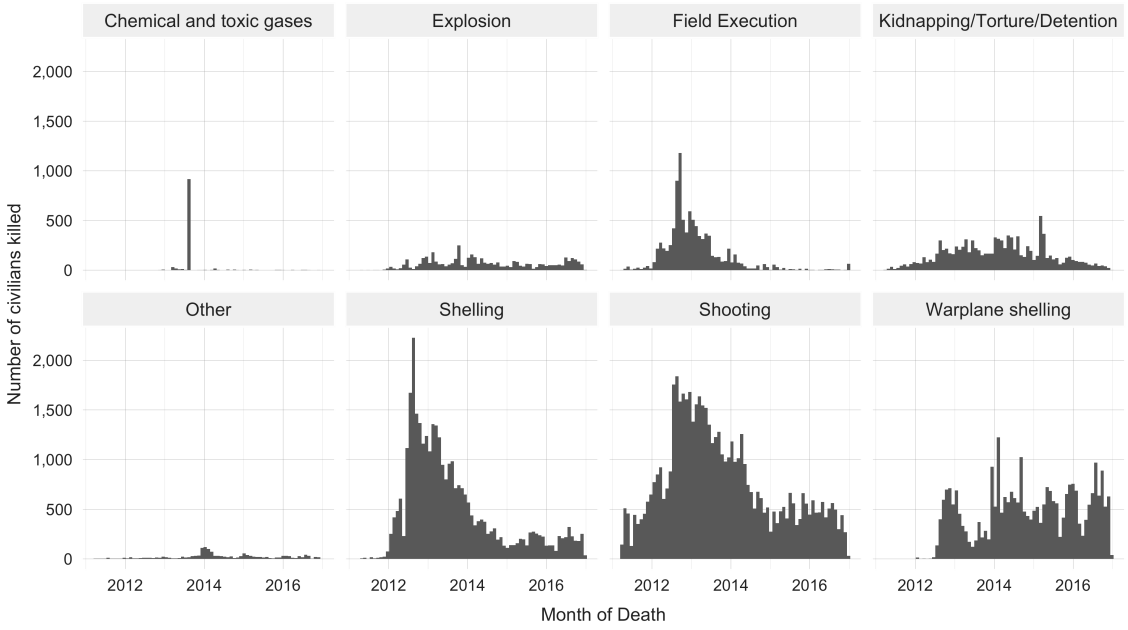


Figure 1: Causes of civilian death per month, VDC

requirements mean that many of the deaths in Syria are not recorded in the dataset. For context, the total death toll of the war in Syria, including rebel fighters, government soldiers, and civilians, is estimated at over 400,000.[4]

Figure 1 shows an alternative version of the causes of death figure in the main paper. This version, using VDC data with its slightly different categorization of causes of death, reveals an even higher level of indirect death in the dataset. The number of deaths from "warplane shelling" is much higher than the "aerial bombardment" deaths in the Shuhada dataset. Field executions, a direct and likely to be selective cause of death, decrease dramatically after 2013, meaning that theories that focus on selective deaths will be limited to explaining a shrinking share of causalities.

1.2.1 Syrian Center for Statistics and Research arrests data

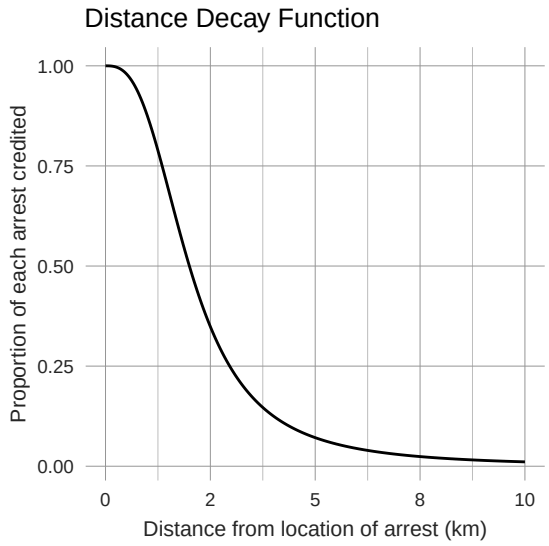
To provide a measure of where the Syrian government perceived the greatest challenge to its rule during the 2011 phase of the conflict, I obtained information on the locations of arrest in 2011. from the Syrian Center for Statistics and Research (CSR). This dataset includes a "town"/neighborhood field as the most fine-grained level of geolocation, with city information as well.

[4]Priyanka Boghani, "A Staggering New Death Toll for Syria's War – 470,000", *Frontline*, February 11, 2016, http://www.pbs.org/wgbh/frontline/article/a-staggering-new-death-toll-for-syrias-war-470000/; Al Jazeera, "Syria death toll: UN envoy estimates 400,000 killed", April 2016, http://www.aljazeera.com/news/2016/04/staffan-de-mistura-400000-killed-syria-civil-war-160423055735629.html

The distance decay function I use is a logistic decay function:

$$f(x) = \frac{1}{1 + \left(\frac{x}{2000}\right)^{2.8}}.$$

The two tuning parameters were set qualitatively based on the average distances between settlements with arrests and inspection of maps showing 2011 arrests. My preference is toward being conservative in the extend of spread. That is, I would prefer to treat the effects of arrests as more localized than they perhaps are. The value 2000 in the denominator represents the point at which the function is $0.5x$. The second parameter, set to 2.8, determines the steepness of the curve. The figure below shows the shape of the function.



2 Geolocation

Because territoriality is central to many of the prominent theories of violence against civilians, including the theories I test, I augment the datasets I use with automatically inferred geographic coordinates of civilian death. Having the data represented with coordinates makes it possible to measure the proximity or density of casualties and to merge the casualty dataset with other geographic datasets. To produce geographic coordinates from free form place names, I develop a geolocation algorithm based on all the available geographic information and queries to a database of place names. The algorithm is independent of language used: it attempts to use Arabic-language data and fails back to English data if Arabic data does not produce a match.

2.1 Algorithm

In its broadest outline, the geocoding process consists of making a structured query to a database of place names and geographic coordinates (a gazetteer) and selecting the most appropriate result from the results.

The gazetteer I use to look up place names' coordinates is the Geonames gazetteer (Wick and Boutreux 2011), the largest publicly available geographic gazetteer with around 11 million unique entries, each of which includes a place's name, alternative names, geographic feature type, country and province/governorate, and its coordinates. I downloaded the CSV dump of the gazetteer and loaded it into an Elasticsearch index. Once in an Elasticsearch index, the data can be queried by string match or matches on other structured field (e.g. searches can be restricted to a particular governorate in Syria). Using Elasticsearch specifically brings major benefits in query speed and allows for fuzzy string matches.

The Shuhada dataset is available in two versions: English and Arabic. I scrape both datasets and merge them on their shared unique casualty ID. The merged dataset then has information on the victims' reported locations of death in both transliterated Latin characters and in the original Arabic form. A single transliteration standard is not consistently applied in translating from Arabic to Latin characters, making lookups in the place name gazetteer very difficult.[5] Instead, I use the original Arabic forms whenever possible in geolocating the causalities.

I then construct a query for each place name. Each entry in the causality dataset has three fields of information about the location of the casualty's death: the governorate of death, reported for 96.9% of the casualties, the city of death, reported for 96.9% of deaths, and the neighborhood of death, reported for 25.1%. Because the same location information is reported in both the English and Arabic forms, these figures are identical across languages. I resolve each casualty's reported place name to coordinates using a constrained search of the gazetteer according to a rule-based system. First, I attempt to resolve the neighborhood, if given, to its geographic coordinates. I query Geonames/Elasticsearch, constrained to the specified Syrian governorate, and then prefer, in descending order, the codes for neighborhood, populated place, section of populated place, subdistrict, or other small geographic features such as markers, mosques, or squares.[6] In cases where multiple matches are found, the algorithm prefers results that have an exact name match. In further cases of ties, the algorithm returns the result that has the highest "relevance" score, as calculated by Elasticsearch's default relevance scoring. If the algorithm is unable to find coordinates for a neighborhood, either because no neighborhood information was reported (74.9%) or because neighborhood information is reported but cannot be geolocated (19.53%), I then attempt to geolocate the city. The algorithm for resolving city names is similar, with the algorithm preferring results with the codes for capitals, cities, or villages, and then as a fallback considering places with neighborhood, "locality," or "populated place" codes. See

[5] In this case, the authors of the dataset use a direct transliteration that omits vowels from the English forms. While this is faithful to the original Arabic form, it is not a common practice and the gazetteer entries rarely include this form as an alternative name.

[6] Inspection of the place names revealed that many of the neighborhood names reported in the Shuhada dataset appear only as names of mosques or squares in the gazetteer.

the replication materials for an efficient implementation and the precise codes used. After performing this process, I obtain neighborhood or city-level geographic coordinates for 89.2% if the civilian casualties in the Shuhada dataset.

I also also geolocate the CSR arrests dataset using the same approach.

I estimate 150 hours of human coding time for the Shuhada dataset by multiplying 9,050 unique place names with 60 seconds of average lookup time per place name.

2.2 Carter Center geolocation

The Carter Center dataset is a combination of two kinds of data: a dataset of which group controlled each location on January 1, 2015, and a dataset of territorial control change events. I first converted these datasets into a panel dataset of location–day control.

I consolidated several groups that the Carter Center kept separate, including consolidating "Kurdish forces" and "YPG", "opposition" and "anti-government", and "government" and "pro-government".

The territorial control data has higher resolution than the casualty data: 5,676 unique locations, as opposed to 1,841 unique locations in the casualty dataset after geolocation. Moreover, the precise place names and coordinates used by the control and casualty datasets are different. To merge the two datasets requires a technique for linking points across the datasets. I do this by associating each casualty with the nearest place the the Carter Center control dataset.

A naive nearest neighbor search is very costly: $\mathcal{O}(n^2)$. Needing to get the nearest neighbors for each location on each day d requires either caching or a d-fold increase in time. To accomplish both speedup of the nearest neighbor calculation and the storage of nearest neighbor results, I construct a k-d tree (Bentley 1975), which requires an average insert and search times of $\mathcal{O}(\log n)$. Once the kd-tree is constructed, it is very fast to query the nearest n closest points to a pair of coordinates.

While latitude and longitude are sufficient for finding the n nearest neighbors of a point, simply taking the euclidean distance of two points in latitude and longitude does not yield an accurate measure of their distance, due to the variable size of one unit of latitude and longitude as a function of the distance from the Earth's poles. To address this, I calculate the Haversine (great circle) distance between each of the nearest points to find the correct distance in meters.[7] The precise implementation can be seen in the replication materials.

Using the nearest neighbor query and accurate distance function, I then calculate several pieces of information for each point. First, I calculate the distance in meters from point a to the nearest locale that is controlled by a group different from the group controlling point a. This is done by querying the kd-tree for the nearest one neighbor of a and then calculating the Haversine distance to that point.

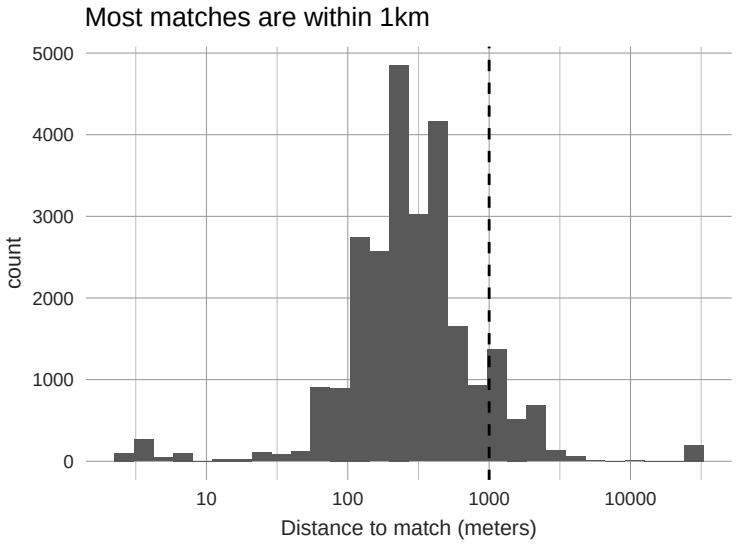
[7] The Haversine measure assumes the Earth is spherical, when it fact it bulges at the equator. This difference is negligible in this context.

Second, I create an adjusted form of this measure that accounts for the median distance to point a's 15 nearest neighbors. This measure adjusts for the differing role of a kilometer in urban and rural areas: two kilometers is whole neighborhoods away in a city, but right next door in rural areas. The Carter Center has much higher place density in urban areas than rural areas, so dividing by the median distance to neighbors serves the function of an urban/rural adjustment.

Third, I calculate the proportion of the 15 closest locales to each location that are controlled by the same group as that location. I include this measure as an alternative measure of the "precariousness" of control over an area. I do this to better reflect the possible effects of being surrounded by opposing areas. Even after accounting for the average distance between a place and its nearest neighbors, the distance to closest "enemy" location does still not reflect the degree of immediate contested control.

2.3 Assessing Match Quality

The matching procedure results in very close matches for most causalities. Most causalities are within 500 meters of their match (median distance = 267, median distance = 679.2). Only 232 deaths are more than than 5 kilometers from their match (see Table 1).



2.4 Good Judgement Project Data

The questions I use from the Good Judgement Project take the form, roughly, of "Will Assad remain in power on date X?". As the closing date of the question approaches, a rational forecaster would reduce the probability assigned to the event occurring, and this trend is clearly visible in the forecasts. To produce a measure of regime threat that is comparable across time, I divide each probability by the time remaining in each forecasting period, so, for example, a forecast that Assad has a 10% chance of vacating power before

Table 1: Casualities connected to places more than 5km away

geoname	count
Nā iyat Markaz al Mayādīn	201
Bīr Umm al Qabābīr	9
Arā ī ar Rābiyah al Gharbīyah	5
Tall al Jābir	5
Abū Dallah	2
Al afāyir	2
Ash Shūlā	2
Dibsī Faraj	1
Jāsim Wasmī	1
Khirbat as Suwaydīyah	1
Sahlāt Sahlāt Jubb as Sayl	1
Salmāsā	1
Şīrat Jubb ash Shā'ir	1

20 days from now produces a rate of $\frac{10\%}{20 \text{ days}} = 1\%$ per day. This simple process removed duration artifacts qualitatively better than a more sophisitcated approach of setting the final day's probability to 0 and removing a linear time trend.

3 Models

Table 2 and 3 report the full logit regression results from the models used to create the predicted probability figures in the main paper. The first set of regressions shows territorial control and distance to the front line effects. The second set (Table 3) shows results after incorporating arrests as a measure of pre-war mobilization and opposition to the regime. Both tables report cluster-robust standard errors for individual locales (n = 1761 unique locations).

4 Note on Replication

The data analysis is performed in the provided `.Rmd` code using R version 3.4.3 running on macOS High Sierra 10.13.6. The code to transform and extend data is provided in the `.py` file, run using Python 3.6. The `.py` file requires a running Elasticsearch server with a pre-built Geonames index. Code to create and run this database is available here: https://github.com/openeventdata/es-geonames/.

Table 2: Logistic regression of death on spatial variables

	Dependent variable:			
	Direct Death		Indirect Death	
	(1)	(2)	(3)	(4)
dist_to_enemy	-0.0002^{***}	-0.0002^{***}	-0.0002^{***}	-0.0002^{***}
	(0.00004)	(0.0001)	(0.00003)	(0.00004)
I(dist_to_enemy^2)	0.000^{***}	0.000^{**}	0.000^{***}	0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
I(dist_to_enemy^3)	-0.000^{**}	-0.000^{**}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
frac_friendly		−0.785		1.278
		(2.127)		(1.658)
I(frac_friendly^2)		0.329		−0.693
		(1.659)		(1.259)
median_dist		0.00003^{*}		−0.00001
		(0.00002)		(0.00004)
Constant	-5.283^{***}	-5.093^{***}	-5.371^{***}	-5.802^{***}
	(0.214)	(0.617)	(0.190)	(0.505)
Observations	3,627,468	3,627,468	3,627,468	3,627,468
Log Likelihood	−43,695.470	−43,638.470	−33,803.150	−33,780.850
Akaike Inf. Crit.	87,398.950	87,290.950	67,614.300	67,575.700

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors adjusted for clustering at the locale

Table 3: Logistic regression with arrests

	Dependent variable:					
	Direct Death		Indirect Death		Bombing	
	(1)	(2)	(3)	(4)	(5)	(6)
weighted_arrests	0.021***	0.020***	0.020***	0.018***	0.019***	0.017***
	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
I(weighted_arrests^2)	−0.00002**	−0.00002***	−0.00002***	−0.00002***	−0.00002***	−0.00002***
	(0.00001)	(0.00001)	(0.00000)	(0.00000)	(0.00001)	(0.00000)
dist_to_enemy		0.00004		−0.00003		−0.00002
		(0.00003)		(0.00003)		(0.00005)
I(dist_to_enemy^2)		−0.000		0.000**		0.000
		(0.000)		(0.000)		(0.000)
enemy_prox_adj		−0.348***		−0.294**		−0.141
		(0.119)		(0.150)		(0.306)
I(enemy_prox_adj^2)		0.004***		0.003**		−0.014
		(0.001)		(0.001)		(0.018)
frac_friendly		−1.124**		−0.230		0.065
		(0.488)		(0.443)		(0.436)
median_dist		−0.00002		−0.00002		−0.00004
		(0.00003)		(0.00005)		(0.0001)
Constant	−6.939***	−5.575***	−7.256***	−6.104***	−7.655***	−6.872***
	(0.097)	(0.330)	(0.090)	(0.375)	(0.095)	(0.374)
Observations	3,893,736	3,627,468	3,893,736	3,627,468	3,893,736	3,627,468
Log Likelihood	−40,628.060	−39,322.830	−31,553.980	−30,486.540	−21,014.850	−20,526.140
Akaike Inf. Crit.	81,262.110	78,663.660	63,113.950	60,991.070	42,035.700	41,070.280

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors adjusted for clustering at the locale

References

Bentley, Jon Louis. 1975. "Multidimensional Binary Search Trees Used for Associative Searching." *Communications of the ACM* 18 (9): 509–17.

Wick, Marc, and C Boutreux. 2011. "GeoNames." *GeoNames Geographical Database*.