Violence against civilians in the Syrian civil war: Evidence from new micro-level data

Andrew Halterman (MIT)

May 1, 2020

Abstract

What explains armed groups' extensive violence against civilians in the Syrian civil war? Existing theories of violence against civilians in civil war offer indeterminate predictions about why it occurs. I compile the largest available micro-dataset on civilian death in civil war, consisting of data on the dates, locations, and causes of over 100,000 civilian deaths in the Syrian war, along with fine-grained data on armed groups' territorial control, the locations of arrests during protests in 2011, and a novel measure of regime threat. Using this data, I systematically evaluate existing theories' abilities to explain violence in Syria. I find little support for prominent theories of violence against civilians that emphasize territorial control, regime threat, or differences between types of violence. Instead, strategic logics of deliberate civilian violence and mass violence in areas of anti-regime mobilization better explain casualties in Syria.

Word count: 8,773

1 Introduction

Despite extensive research on violence against civilians in civil war, the literature presents mixed predictions about when, why, and to what extent armed groups in civil war will target civilians.[1] Many scholars and practitioners of civil war, especially those studying insurgency and counterinsurgency, see violence against civilians as irrational and counterproductive (in addition to being morally repugnant): Indiscriminate violence against civilians reduces their willingness to cooperate and encourages them to defect to the other side (Nagl 2002; Nagl et al. 2008; Kocher, Pepinsky, and Kalyvas 2011; Lyall, Blair, and Imai 2013; Dell and Querubin 2016, 2017). Violence creates cycles of revenge and hatred-driven opposition to the perpetrating actors (Petersen 2002; Balcells 2017). Other research finds that collective targeting of civilians can be effective in reducing their support for rebels and weakening their ability to continue fighting (Valentino 2000; Valentino, Huth, and Balch-Lindsay 2004; Eck and Hultman 2007; Downes 2007; Fjelde and Hultman 2014). Targeting civilians can also reduce post-war opposition to the government (Balcells 2017).

The ongoing Syrian civil war has seen widespread targeting of civilians during the conflict with at least half a million killed and a majority of civilians displaced from their homes.[2] Despite the scale of civilian causalities in Syria and extensive scholarly attention to other features of the conflict, little work has directly studied the causes of civilian victimization in Syria. Most research on civilian causalities in the conflict consists of summary statistics on the number of fatalities and weapons used (e.g. Guha-Sapir et al. 2015, 2018) or documents the human rights abuses that are occurring (e.g. work by Amnesty International, Human Rights Watch, and Syrian groups such as the Syrian Observatory for Human Rights). This work has been limited by the availability of data and has not systematically engaged with social science theories of violence against civilians.

This paper introduces the most detailed micro-level quantitative dataset of civilian casualties in a single civil conflict. Using new measurement techniques and a range of raw datasets, I produce a dataset that provides information on the dates, geographic coordinates, and

weapons used in over 100,000 civilian deaths, daily measures of the degree of territorial control, the number and locations of arrests in the early phases of the uprising in Syria, and a novel measure of threat to the regime's survival. The data can be analyzed as a standalone dataset of civilian casualties or as a locality (town or neighborhood) panel dataset and is available for scholars of civil conflict to analyze.

This dataset allows for a systematic comparison of existing theories in explaining wartime violence against civilians in Syria. I enumerate and systematically test a set of theories of violence against civilians in conventional civil wars. The theories I test consist of a collateral damage theory that civilians are harmed as a byproduct of conventional fighting, a theory that violence is used strategically to shape civilian behavior, that violence is a political tool for shaping the post-war order, that violence results from threats to the regime, and that violence is used selectively to consolidate territorial control.

Applying a set of existing theories to a single conflict, and specifically the civil war in Syria, allows us to systematically compare the performance of existing theories on a new case. Given this news data, which theories are able to explain violence in Syria? Are they sufficient to understand the causes of killing in Syria?

I find support for theories of violence against civilians that emphasize strategic factors in violence against civilians. Violence is much more likely in front line areas and in areas with pre-war political opposition to the regime. The importance of pre-war mobilization indicates that pre-war politics remain important throughout the war. Several other theories make predictions that are not borne out in the data. Little variation in casualties exists across different levels of territorial control and threat to the regime. The degree of selectivity in violence, important in several theories, does not seem to vary across several variables in Syria.

More broadly, this new micro-level dataset will help researchers address a range of questions. Scholars are interested in cycles of violence and retribution (Shellman 2006; Balcells 2017), which are best studied using data at the village or neighborhood level and measured

daily. Both scholars of civil conflict and methodologists are increasingly interested in understanding and estimating spatial diffusion spillover effects (e.g. Zhukov 2012; Egami 2018). This data, with its detailed geographic information, will help enable work on diffusion and spillover effects. Finally, civilian casualties in civil war result from a number of processes. Future research could focus on specific aspects of the violence: what predicts highly targeted violence, such as abductions or torture? What forms of violence are employed in recently captured areas? By answering questions like these, researchers can improve our understanding of conflict processes and hopefully find ways to limit the loss of life in future civil wars.

2 The civil war in Syria

The civil war in Syria began as a series of protests in mid-March 2011, as part of a broader wave of anti-regime protests in the Middle East. In April, the protests spread and the government response quickly became violent, with security forces firing on protesters. By July 2011, a nascent Free Syrian Army began armed opposition to the state. Fighting intensified through mid-2012 and 2013, with rebel forces, including Jabhat al-Nusra, making important territorial gains, including in Aleppo. On August 21, 2013, a chemical attack in Ghouta caused several hundred deaths and drew greater international attention to the conflict. In 2014, fighting began between ISIS and other rebel groups, and US airstrikes began in September 2014.

The years 2015 and 2016 are especially relevant for this study, given the data availability I describe later. In the first half of 2015, Jabhat al-Nusra and other mostly Islamic forces were advancing in Idlib and northwest Syria. Meanwhile, ISIS was advancing in Palmyra, Kobane, and al-Haskakeh. Russian airstrikes began in October 2015. Syrian government forces took Latakia in January 2016 and parts of Daraa Governorate. A ceasefire was in effect from February to July 2016, but territorial changes continued. Turkish military forces and Kurdish PYD and YPG forces each took territory in northern and northeastern Syria

in the fall, and in December 2016, government forces took the last areas of rebel-controlled Aleppo.

The civil war in Syria shares important similarities with previous civil wars. First, as a "conventional" civil war characterized by high rebel military capacity and an an absence of guerrilla insurgent tactics, the war is different from many of the wars studied in the civil war literature (Iraq, Afghanistan, Colombia, El Salvador, Chechnya, etc.) but is part of a growing population of conventional civil wars (Kalyvas and Balcells 2010). It has involved a great deal of foreign involvement, an increasingly common feature of civil wars (Anderson 2019). Syria was a middle income country with high literacy rates and connection to the rest of the Middle East. Unlike civil wars fought in poorer countries, the war in Syria takes place in a context of ubiquitous phones and internet, making detailed local information on violence and armed actors much more available than in other civil wars.

Existing work on the Syrian civil war has focused on the rise of ISIS (Warrick 2015; Giglio 2019), international involvement in the conflict (e.g. Hughes 2014), and the recruitment and role of foreign fighters in the conflict (Hegghammer 2013; Rosenblatt 2016). Civilian victimization, and especially the extent to which existing theory explains it, has been less well studied.

3 Theories of violence against civilians in conventional civil war

The data collection process was shaped by the need to produce information on the variables that existing theories see as important for understanding violence against civilians. To guide data collection and provide results on why civilians are killed in Syria, I lay out five existing theories of why government forces would target civilians and their observable implications in the Syrian civil war. First, I describe a "collateral damage" account of violence, that civilians are killed near areas of active fighting without being targeted deliberately. A

second set of theories concerns violence against civilians as coercive or denial strategies, explaining violence as an attempt to reduce the will or ability of rebel forces to continue fighting. A third theory posits that government forces will resort to widespread violence against civilians when regime survival is threatened. A fourth theory holds that violence during war is guided by the government's desire to eliminate political opponents that they identified before the war. Finally, I describe a theory of intelligence and territorial control that applies to irregular wars and extend its logic to conventional civil wars. The observable implications of these theories are summarized in Table 1. As a conventional civil war rather than the guerrilla war or insurgencies often studied in the literature, Syria is outside the scope conditions of many prominent theories of violence against civilians.

Conventional civil wars are differentiated from guerrilla wars by the the existence of defined front lines, areas of control, and conventional military operations (Kalyvas and Balcells 2010). Scholars of guerrilla war have emphasized the critical role of information: when combatants are not identified by uniform or location in reference to a defined front line, identifying enemies becomes the primary obstacle to targeting them. Success in guerrilla war or counterinsurgency thus relies on gaining the cooperation of civilians in identifying enemy combatants or collaborators (Kalyvas 2006; Nagl et al. 2008; Kilcullen 2010; Berman, Shapiro, and Felter 2011; Berman and Matanock 2015). In conventional civil wars, including the one being fought in Syria, the existence of defined front lines and areas of control make the identification of enemy fighters much easier, reducing the need to rely on civilian-provided intelligence. These differences mean that the Syrian civil war is outside the scope conditions of some of the best-known theories of violence against civilians.

3.1 Violence against civilians is collateral damage from military operations

A simple, non-political theory of violence against civilians would posit that violence is not deliberately targeted at civilians, but rather civilians are killed by accident or neglect in

	Distance to front line	Direct/indirect violence	Pre-war mobilization	Regime threat	Degree of control
Collateral Damage	√				
Coercive targeting	√		√		
Denial targeting	√	√			
Pre-/post-war politics		√	√		
Regime Survival				√	
Intelligence/control		√			√

Table 1: Several theories of violence against civilians in conventional civil wars (rows) and the type of variation predicted by each (columns). If a theory makes a prediction about a type of variation, it is indicated with a check mark. The intelligence and territorial control theory is included for comparison despite it applying to irregular, rather than conventional civil wars.

front line areas or areas of active fighting. This scenario includes both civilian casualties that result from "battlefield necessity" and those produced by carelessness, neglect, or malice in the context of ongoing military operations. This distinction is highly consequential morally (Walzer 1977) but less so politically. In either case, civilian are killed because they happen to be in an area of active fighting between armed belligerents. The observable implications of this theory are clear: civilians should not be killed away from frontline areas and rarely targeted directly. Pre-war political factors, such as protests or arrests, should have no relationship to wartime violence against civilians.

3.2 Violence is used to prevent civilian support to combatants

A second set of theories argue that armed groups can use violence to shape the behavior of civilians, coercing them into withdrawing support from their side or inflicting enough damage that civilians are unable to provide support to their side. Combatants in conventional interstate wars may have reasons to rationally target civilians in wartime as a method of coercing them (e.g. Valentino, Huth, and Croco 2006) and militaries have adopted strategies of bombing populations, most notably during the strategic bombing campaigns of the Second World War. These theories assume that civilians can be targeted collectively (Sullivan 2012): combatants do not need specific, individualized intelligence on which side a civilian supports, but can infer it to some extent from the areas where a civilian resides or a group they are a member of.

Later work argues that a "denial" strategy could be at work to deny access to resources, rather than coercing civilians to stop providing resources in both guerrilla wars (Eck and Hultman 2007; Downes 2007; Fjelde and Hultman 2014; Koren and Bagozzi 2017) and conventional civil wars (Krcmaric 2018). Pape (1996) introduces the canonical distinction between "denial" strategies that target civilians to reduce their ability to produce arms, recruits, food, and supplies used by the combatants, and "punishment" strategies to impose costs on civilian populations and thereby coerce combatants to give up. A "denial and punishment" theory would predict greatest levels of violence in areas of Syria completely held

by one side, especially areas with high perceived support for opponents or areas where opponents generate a large portion of their matérial. We would also expect to see greater mass- or collective violence than we would expect if armed groups required specific, individualized intelligence.

3.3 Violence is used to shape the post-war order

Balcells (2017) suggests that governments use more than wartime factors when targeting civilians, making broader political calculations that take into account the postwar political order. "Indirect" violence against civilians, such as airstrikes on cities, are rational in areas with high pre-war anti-regime political mobilization. Civilians in these areas are seen as threats to the regime, both during the war, when they provide support for the rebels, and after the war, when they will pose political problems by supporting opposition parities. Balcells, in studying the Spanish Civil War, used pre-war opposition vote share, finding greatest violence near a (rational) 50% tipping point. In Syria, the model would need to rely on alternative measures of political opposition, and the authoritarian logic of repression might result in a different tipping point than the 50% produced by democratic systems. Thus, the theory would predict greater violence, specifically indirect violence, in areas of high anti-regime opposition before the war.

3.4 Violence results from threats to regime survival

A "desperation" theory (Valentino 2000; Hultman 2007; Downes 2008) suggests that governments or rebels target civilians during protracted conflict or when they are performing poorly on the battlefield, in order to pressure the opposing side into conceding. As the war drags on, armed groups become more willing to try strategies they would not have considered at the beginning of the war and may be increasingly willing to trade long term costs (in civilian opposition) for short term benefits (in reduced will or ability). In Syria, this theory would predict that as the threat to the regime increases or the duration of the war increases, the number of civilians killed by the armed groups (especially the government)

would increase. Actors would prefer selective targeting of civilians but resort to mass killing under situations of greater threat.

3.5 Violence and intelligence are used to consolidate territorial control

Finally, a prominent theory of civilian violence in irregular civil war (rather than conventional civil war) sees violence against civilians as counterproductive and therefore irrational. Both scholars (e.g. Kalyvas 2006; Berman and Matanock 2015) and practitioners (Nagl 2002; FM 3-24, Kilcullen 2010) have proposed a theory of armed group success in guerrilla civil war that relies on the cooperation of the local population. If the "fundamental problem" (Kalyvas 2006) of irregular conflict is identifying enemies and enemy collaborators, civilian cooperation is crucial for gaining the intelligence needed to identify enemies. Cooperation is provided to armed groups that successfully control territory (Kalyvas 2006) or provide governance and services (Sunderland 1964; Nagl et al. 2008). If winning the support of civilians becomes the *sine qua non* of counterinsurgency, violence against civilians, and especially indiscriminate violence, becomes counterproductive. Kalyvas (2006) makes more specific predictions about selective violence: it should peak in areas under partial control, where armed groups have enough presence to generate tips from civilians but still in the process of consolidating control.

This set of theories was developed to explain violence in guerrilla insurgencies, not conventional civil wars with defined front lines. In conventional civil wars, the identification problem is reduced and armed groups have consolidated rears, meaning that civilian intelligence and logistical support is less important. The role of civilian support in conventional civil wars should not be discounted completely, however. As in conventional interstate wars, civilians produce the arms, food, and supplies needed by conventional intrastate combatants to sustain their fighting (fjelde2014weakening; Koren and Bagozzi 2017). This reliance on civilians implies that the same dynamics as guerrilla war could be portable to conventional civil war.

If the "hearts and minds"/intelligence/territorial control theory does apply to Syria, violence against civilians should be fairly uncommon and when present, specifically directed at individual civilians. Indiscriminate violence against civilians is irrational and therefore rare. Where violence does occur, it should be in areas of partial control, rather than fully contested or fully controlled areas, as armed groups attempt to consolidate their control.

4 Constructing a Micro-Level Dataset of Civilian Casualties

To understand which of these theories explain the broad patterns of violence in Syria, I collect a new set of large-scale quantitative data on violence against civilians, along with information needed to test each theory. I collect several types of data: the locations and dates of civilian deaths, the cause of death, groups' control of territory, the degree of territorial control, distance to the front line, a measure of pre-war mobilization, and a measure of threat to the regime's survival. I collect daily data on each of these and geocode them, when applicable, to the neighborhood, village, or city. I compile this information from a number of sources, enhance it using methods described below, and combine it into a single consolidated panel dataset, covering around 6,000 unique locations observed daily in Syria between 2011 to 2016 with the number and cause of civilian deaths at each point, and with territorial control available after 2015. The dataset thus consists of day-resolution counts of deaths in Syria, augmented with the geographic coordinates of the death at the neighborhood or village level, the group occupying that location on that day, and its occupier and degree of control.

An overview of the data sources and their dates of coverage are presented in Table 2.[3]

4.1 Data on civilian deaths and their locations

I use data provided by the Syrian Shuhada site to provide raw data on civilian deaths in the Syrian civil war .[4] The Shuhada ("Martyrs") dataset records the name, date, combatant

Data	Source	Modifications	Dates Covered
Cause of Death	Shuhada	Balcells' direct/indirect coding	2011–2016
Cause of Death (alternative)	VDC	Balcells' direct/indirect coding	2011–2018
Location of Death	Shuhada	Geocoding algorithm	2011–2016
Distance to Frontline and Enemy Areas	Carter Center	Geocoding algorithm, nearest neighbors, distance calculation	2015–2016
Early Phase Arrests	Syrian Center for Statistics and Research	Geocoding algorithm, nearest neighbors, distance calculation	2011–2016 (only 2011 used)
Threat to Regime	Good Judgement Project forecasts	Forecast duration adjustment	Nov 2011–Feb 2012, June 2012–June 2013

Table 2: Data inputs to the final dataset, with sources, dates covered, and modifications made.

status, cause of death, place name of birth, and place name of death from 2011 through 2016. The majority of the data, 65.7% of civilian casualties, is drawn from the Center for Documentation of Violations in Syria (VDC), which compiles information on the death of civilians, rebels, and regime forces in Syria. Price, Gohdes, and Ball (2014) find the VDC dataset to be an accurate source with as good of coverage as any other single dataset. Another 23.8% of the reported deaths come from the Syrian Center for Human Rights, and the remaining 10.5% casualties are compiled from news sources, YouTube, opposition groups, Facebook posts, and Local Coordinating Councils. The Shuhada dataset consolidates data from across these sources, standardizes their formats, and augments them with greater detail on the locations of Syrians' deaths. Figure 1 shows an example of the raw data in English and Arabic. The requirement that casualties be documented with a name, date, and location of birth and death suggests that the deaths included in the Shuhada and VDC datasets are underreports of the total. Their cooperation with opposition groups could also induce an underreporting of civilians killed in government areas or by rebel combatants. I scrape and clean this dataset, merging information from separate English, Arabic, and combatant status pages into a consolidated dataset. Figure 2 reports the breakdown of causes of death in the Syrian civil war by cause over time from the beginning in 2011 to mid-2016.

The Shudada dataset includes fine-grained geographic information on the location of civilian causalities, reporting the governorate and city of death for 96% of casualties, and reports an additional neighborhood of death for 26.0% of the casualties. Although this information is present, it has not been used by researchers at resolutions finer than the governorate level (e.g. Guha-Sapir et al. 2018) because the locations are provided as free text, not as geographic coordinates. There are 4,543 unique Arabic-language place names reported by the Shuhada dataset, and 4,507 unique English transliterated place names. Looking up geographic coordinates for the 4,543 locations by hand would be a laborious undertaking.[5] Instead, I create an algorithm that can take in place names in either Arabic or English and return their geographic coordinates, using information about the governorate and type of

معلومات الشهيد (رقم 58409)

الاسم الكامل	بلال محمد خير الضيخ
الاسم الأول	بلال
الاسم الآخر	الضيخ
العمر بالسنة	
حي الإستشهاد	
مدينة الإستشهاد	تلبيسة
محافظة الإستشهاد	حمص
مدينة مسقط الرأس	تلبيسة
محافظة مسقط الرأس	حمص
تاريخ الإستشهاد	8/3/2013
الجنس	ذكر
ملاحظات	استشهد بالاشتباكات مع قوات النظام.
المصدر	مركز التوثيق /http://www.vdc-sy.org
طريقة الإستشهاد	بطلق ناري
الجنسية	سوريا
فيديو أول	https://www.youtube.com/watch?v=oNqbdXsRMJM

Martyr Information (Martyr ID: 58409)

Full Name	Blal mhmd khyr aldhykh
First Name	Blal
Last Name	Aldhykh
Age in Years	
Neighborhood of Death	
City of Death	Paneling
Province of Death	Homs
City of Birth	Paneling
Province of Birth	Homs
Date of Death	8/3/2013
Gender	Male
Comments	استشهد بالاشتباكات مع قوات النظام.
Data Source	مركز التوثيق http://www.vdc-sy.org/
Death Method	Gunshot wound
Nationality	Syria
First Video	https://www.youtube.com/watch?v=oNqbdXsRMJM

Figure 1: Raw data from Shuhada in Arabic and English. The data on location of death is available as a free-form text box, meaning that it cannot be merged with other geographic data until place names are resolved to their geographic coordinates. The Arabic and English datasets report the same information but English data uses an inconsistent transliteration scheme.

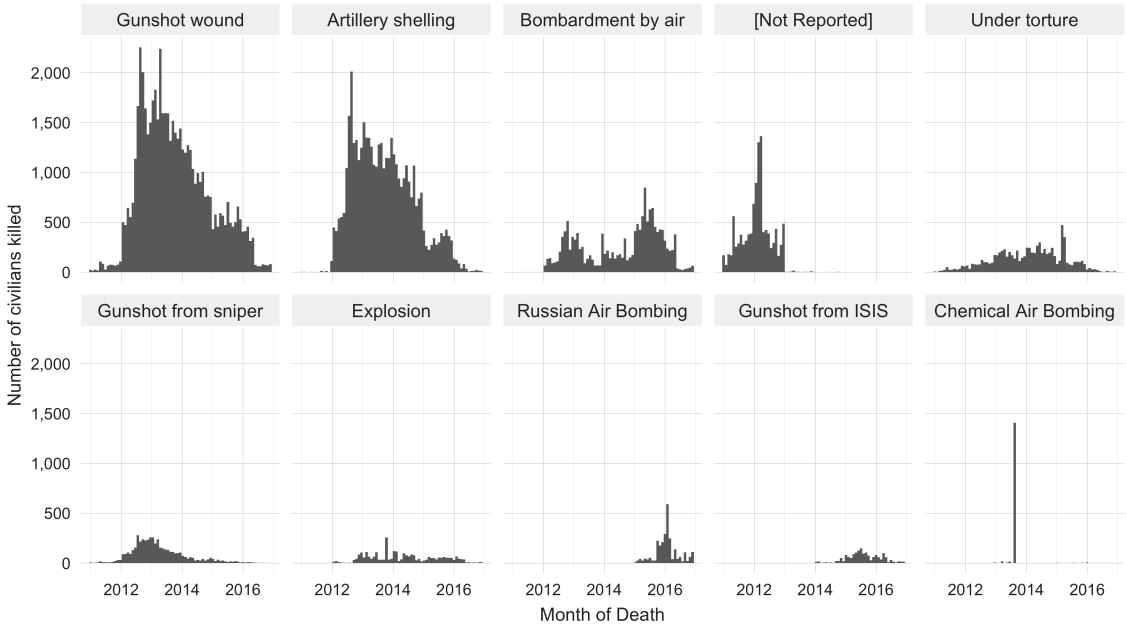


Figure 2: Causes of civilian death per month reported in the Shuhada dataset. Limited to the top 10 causes, comprising 94% of total civilian casualties. Two of the three greatest causes of death, aerial bombardment and artillery, are produced using heavy, indirect weapons.

place (city vs. neighborhood) to return a precise match.

I look up each place name from the Shuhada dataset in a search database populated with the Geonames gazetteer (Wick and Boutreux 2011) of place names, limiting the search to the given governorate in Syria and using a series of rules to determine the best match from the potentially multiple results. Searches are limited to the reported governorate of death, and towns and neighborhoods are prioritized over other geographic features such as street names or natural features. Exact matches are preferred but fuzzy matches based on edit distance are also allowed. I attempt the search first using the Arabic place names; if no results are found, I attempt again with the transliterated English place names.[6] After completing this process, I was able to obtain geographic coordinates for 90.0% of the civilian casualties: 104,134 out of a total of 116,026 civilian deaths recorded in the dataset. This language-agnostic algorithm can be used anywhere that researchers have structured geographic place name data and would like to produce geographic coordinates. Automating the task makes the process much faster, reproducible, easily extensible to other datasets, and allows the dataset to be rapidly expanded as more data is made available.

Figures 3 and 4 show an example from the cleaned and geolocated dataset and present a validity check on the data. On May 30th, 2014, a barrel bomb attack in the Aleppo neighborhood of Bustan al-Qasr was reported in the *Al-Arabiya* newspaper.[7] The attack in the northern neighborhood of Midan received less press attention but is still recorded in the dataset. Figure 4 shows an earlier period of conflict in Aleppo. Agence France-Presse reported a major Syrian military counter-offensive into the Salaheddin neighborhood of Aleppo on July 28.[8] Government forces assaulted the Salaheddin neighborhood with armor and helicopter gunships, and the rebels who controlled the neighborhood fought back with small arms and light weapons.

To answer more specific and theoretically motivated questions, though, the data needs to be enhanced with other variables from other data sources. Table 1 shows the degrees of variation that each theory has observable implications along. The following sections describe

Some barrel bomb deaths are coded as artillery deaths, but the note makes it clear that they are from air strikes: "ا‌ستشهد جراء القصف العشوائي لقوات الطيران السوري."
["Killed due to the indiscriminate shelling of the Syrian regime warplanes."]

Figure 3: Civilian causalities in Aleppo on May 30, 2014. A barrel bomb attack in the neighborhood of Bustan al-Qasr killed a reported 20 civilians. Not shown are five deaths geolocated to "Aleppo" without further neighborhood information.

Figure 4: Civilian casualties in Aleppo on July 28, 2012. Government forces and Syrian rebels fought in the neighborhood of Salaheddin. All causalities in Aleppo on this date had neighborhood geographic information reported.

other theoretically motivated variables I collect and add to the dataset.

4.2 Measuring front lines and territorial control

Understanding the role of territorial control in explaining civilian casualties requires information on which groups controlled which locations on each day. Territorial control maps in Syria are very popular and several organizations, including the Institute for the Study of War, Caerus Associates, and the New York Times have produced control maps since the beginning of the conflict. While very useful for getting a sense of the conflict from a macro level, these maps often do not have the right spatial and temporal resolution to be useful and are difficult to use in quantitative analysis.

I use a dataset compiled by the Carter Center, which tracks changes in territorial control at the village or neighborhood level from January 1, 2015 to present.[9] I scraped and formatted the data into a location–day panel dataset of territorial control. Each coding of a territorial control is accompanied by a citation to a news report or social media report describing the capture. The major drawback of the Carter Center data is that its coverage only extends back to January 1, 2015, while the civilian casualty death data from Shuhada extends back to the beginning of the conflict in 2011. After cleaning, the territorial control dataset contains 5,676 towns/neighborhoods, each of which has data on which group controls it, coded on a daily basis.

The Carter Center's data treats control as a hard indicator: every location is controlled by exactly one group in equal degree. To adapt this data to theories concerned with degree of control, I develop and compute two new continuous measures of territorial control. The first measure is the distance to the nearest locale controlled by a different group, to reflect how close an area is to the front line or nearest enemy. A second measure is the proportion of locales in a place's immediate neighborhood of 15 that are controlled by other groups, to better capture the "precariousness" of control.[10] Being surrounded by "enemy" areas is a conventional analogue to the degree of control measure that Kalyvas uses. Being surrounded

makes loss of control more likely and the presence of opposing forces more salient. If civilian cooperation is necessary to produce direct violence and civilians fear retaliation if control should flip, we should expect less direct violence in highly precarious areas.

I also compute and use as a control a measure of how urban or rural an area is, given the known issues with ignoring different urban and rural dynamics in conflict (Kalyvas 2004; Douglass and Harkness 2018). Settlements included in the Carter Center dataset are much denser in urban areas than rural. I calculate the median distance to these nearest 15 locales as a measure of of settlement density.

In an irregular war of the type Kalyvas mostly considers, front lines are permeable and civilians need to fear that groups have read into their opponents areas where they could punish civilian collaborators. The mechanism is different here. Front lines are relatively impermeable, but the lines can shift over time. These areas of partial control are insecure in that that are most at risk of flipping control, triggering the retaliatory killings that are a day-to-day fear in irregular civil wars.

I join the civilian casualty and territorial control datasets by making a panel dataset of all Carter Center locations and assigning civilian casualties to the nearest location.[11] After matching locations in the causality dataset and the territorial control dataset, most locations are within 1 kilometer of their match (see the supporting information). Only 232 deaths are more than 5 kilometers from a locale in the territorial control dataset.

4.3 Measuring direct and indirect causes of death

Some theories make specific predictions about the types of lethal violence directed against civilians, namely whether the violence is selective/direct or indiscriminate/indirect, with direct violence being violence committed with small arms, and indirect violence being the result of air strikes, artillery fire, and in the case of Syria, chemical weapons. Using Balcells' definitions of direct and indirect violence, I map each of the reported causes of death in the Shuhada dataset to those the two categories of direct and indirect. I also code regime

airstrikes separately, given their importance in Balcells' theory.

4.4 Measuring pre-war political mobilization with arrest data

In Balcells' (2017) main case of the Spanish civil war, she uses pre-war vote share as a measure of opposition to the regime. Syria's rigged elections make Syrian electoral data useless for this purpose.[12] As an alternative measure of the government's perception of anti-regime mobilization, I compile geolocated data on arrests of activists and protesters in 2011. Through July of 2011, when the Free Syrian Army began operating, the Syrian regime had free access to the entire country. Thus, during the initial protest phase of the conflict, arrests by government forces are a good proxy of government perception of opposition to the government in an area. The Syrian Center for Statistics and Research, an NGO, collected data on arrests of civilians by the government.[13] This data includes the name of the arrestee and the date and location of the arrest. Following a similar procedure to the Shuhada dataset, I scrape and geocode the arrests in the dataset to produce geographic coordinates for each arrest. After linking each location name to its geographic coordinates, I then merge it with the data on locations and civilian causalities during the rest of the conflict. Rather than assigning the arrests only the closest location in the panel data, I use a distance-decay function to assign partial credit to other nearby locales to account for spillover across locations, on the assumption that opposition is spatially correlated and arrests in one neighborhood provide some information about opposition in neighboring places.[14]

4.5 Measuring regime threat with forecasting data

As the situation for one side becomes more "desperate", it may be more willing to turn to indiscriminate violence against civilians to stave off defeat. Previously, though, fine-grained measures of threat to the regime had not existed. To measure threat to the regime, I employ contemporaneous forecasts about the probability of Assad leaving power made as part of the US intelligence community's Good Judgement Project, producing a daily value for the

estimated threat to the regime. The Good Judgement Project (GJP) was a project sponsored by the Intelligence Advanced Research Projects Activity to study whether accurate geopolitical forecasts could be produced from crowdsourced decisions. Since the conclusion of the project, all forecasts have been publicly released to researchers. While the precise performance of the project compared to classified analyses are not publicly available, media reporting on it suggests that its forecasts matched or exceeded the intelligence community's existing approaches.[15] The overall accuracy of the GJP forecasts suggest that their specific forecasts about the prospects of the Assad regime may be the best available assessment available to actors at the time of the probability of Assad remaining in power. As his probability of retaining power increases, a "desperation" theory would predict increasing violence against civilians. I take answers to questions asking about whether Assad will remain in power on a specific date in the future as a measure of threat to the regime. If forecasters believe that Assad is likely to remain in power, the regime is relatively secure. If they assess higher probabilities of him leaving power, the threat to the regime is higher. Because the forecasts are about survival in office on a particular date in the future, the probability of survival rises mechanically as as the date approaches. I therefore adjust the probabilities, setting the survival probability on the final day of the forecast window to 1 and removing a linear time trend from the forecasts.[16]

4.6 Panel dataset

I combine all of these datasets into a locality-week panel dataset. For each of the 5,676 unique locations, I construct a weekly panel with each variable. I also create an indicator variable for the occurrence of any direct or indirect deaths in a locale-week: $I\{\text{deaths}_{it}^{\text{direct}} > 0\}$, $I\{\text{deaths}_{it}^{\text{indirect}} > 0\}$, where i indexes localities and t indexes weeks and I is an indicator function. Dichotomizing the data is justified both theoretically and empirically. The differing casualties from single direct and indirect attacks are difficult to compare, as a single indirect attack will generally produce more casualties. Moreover, most theories of violence against civilians are concerned with whether, where, and how the attacks are occurring,

not with the deadliness of any particular attack. Finally, the modal weekly casualty count, given that a casualty occurs, is 1.

5 What causes civilian causalities in Syria?

To analyze the data, I use a set of logistic regressions of civilian causalities on a series of theoretically interesting variables. While the data is in the form of a panel, with localities observed over time, I opt to not use a fixed effects panel regression for both methodological and theoretical reasons. Methodologically, one of the key variables of interest, pre-war arrests, are constant over time and would be absorbed into unit fixed effects. Moreover, with rare binary outcomes, a large amount of data is wasted by using unit fixed effects (Beck and Katz 2001)[17]. The majority of localities do not experience casualties and would thus be dropped from the sample. However, where violence does *not* occur is a major component of some theories. Unit fixed effects are also theoretically suspect in this situation because they induce a bundling of treatments. A well known result from behavior economics is that people respond not only to levels of a variable, but also to changes in a variable (e.g. Kahneman and Tversky 1979). We would also expect that a locality whose distance *changed* from 10km to 1km from an enemy locale would experience both an effect of being closer to an enemy area and also an effect from having its status change, perhaps through a mechanism of people feeling more precarious than before.

Running these regressions produces several interesting findings. Two theories have major predictions borne out: violence is more likely near the front line, and pre-war political mobilization, measured using arrests, has a major effect. Several other theories do not have their hypothesized variation appear. The findings are summarized in 3.

Type of Variation	Observed?	Note
Degree of control	X	Degree of control is not related to casualties after accounting for distance to front line
Distance to front line	√	Localities closer to enemy localities experience more violence against civilians
Type of violence	X	Indirect and direct violence have similar patterns, in contrast to previous theory
Pre-war mobilization	√	Pre-war arrests are a good predictor of violence during the war
Regime threat	X	Periods when regime threat is high do not see greater casualties than periods when regime survival seems likely

Table 3: Predicted variation for each theory and whether it is observed in the Syrian civilian causality panel dataset.

Figure 5: Distance to the nearest "enemy"-held area and probability of a locale-week experiencing civilian casualties. The probability of both direct and indirect violence increases for areas that are closer to an "enemy"-controlled area, which is consistent with causalities being produced by frontline collateral damage or strategic mass violence theories. Shaded areas indicate 95% confidence intervals. See the supporting information for the regression table and a model without controls.

5.1 Proximity to enemy locations increases casualties

Proximity to the nearest enemy-controlled locality has a strong effect on the probability of a locality experiencing at least one casualty in a given week. In 2015–2016 period, 69.5% of 44,521 deaths for which territorial control data is available occur within 5 kilometers of an area held by a different group. Localities within a kilometer of an enemy-controlled locality have an estimated 1/2% probability of having at least one civilian causality resulting from direct or indirect violence (Figure 5). For localities that are 15 kilometers away, the predicted number of causalities is near 0 (although with a large confidence interval). This relationship accounts from the fraction of friendly controlled localities nearby and adjusts for the median distance to neighboring localities, to reduce any potential urban bias. It is somewhat surprising is that the relationship holds given the technology available for long range indirect fire, including airstrikes and artillery.

Figure 5 reports the relationship between the distance to an "enemy"-controlled area and the probability of experiencing a civilian casualty. It shows the predicted probabilities of a locale-week experiencing a civilian casualty by varying distance to the nearest enemy area using a logistic regression model. The model includes the distance adjusted by the average distance to nearby locales to account for urban/rural differences, and the fraction of nearby areas controlled by the same group, plus squared terms of each. The model is fit with n = 44,521 civilian casualties from 2015 and 2016 and standard errors are clustered by locality. The closer a locale is to the nearest place controlled by another group, the higher both direct and indirect violence is.

5.2 Arrests are good predictors of subsequent violence

Pre-war arrests are a good predictor of wartime violence. Localities that had more arrests in 2011 during the protest phase of the uprising were much more likely to have violence against civilians during the war, even after accounting for the degree of territorial control, distance to the nearest "enemy"-controlled locality, and the density of nearby settlements.

The correlation between arrests and subsequent violence is positive and significant for every type of violence: direct, indirect, and bombings specifically. Furthermore, the marginal effect of an arrest increases, as the upward curve of the slopes indicate.

The finding that pre-war arrests matter is consistent with what Balcells' (2017) "rivalry" dynamic : violence is used to kill people in areas that appear to have greater opposition to the regime. While violence in civil war can be increasingly driven by processes and cycles endogenous to the war itself (Kalyvas 2006), in this case, pre-war factors matter throughout the war.

5.3 No variation by degree of territorial control

The analysis also shows several interesting negative findings, where relationships predicted by existing theories do not appear. Specifically, no link exists in this analysis between civilian casualties and the degree of territorial control or threat to the regime. Direct and indirect causes of death also occur in similar patters, rather than being substitutes for each other, as many theories imagine.

The results do not show any link between the degree of territorial control, measured by the proportion of neighboring localities under the control of the same group and the probability of violence against civilians (Figure 7). After controlling for proximity to the nearest "enemy"-held locality, the number of neighboring localities held by an opponent does not change the probability of experiencing civilian causalities. The result is perhaps not surprising, though. The intelligence and territorial control theory that predicts it has a scope condition limited to guerrilla civil wars.

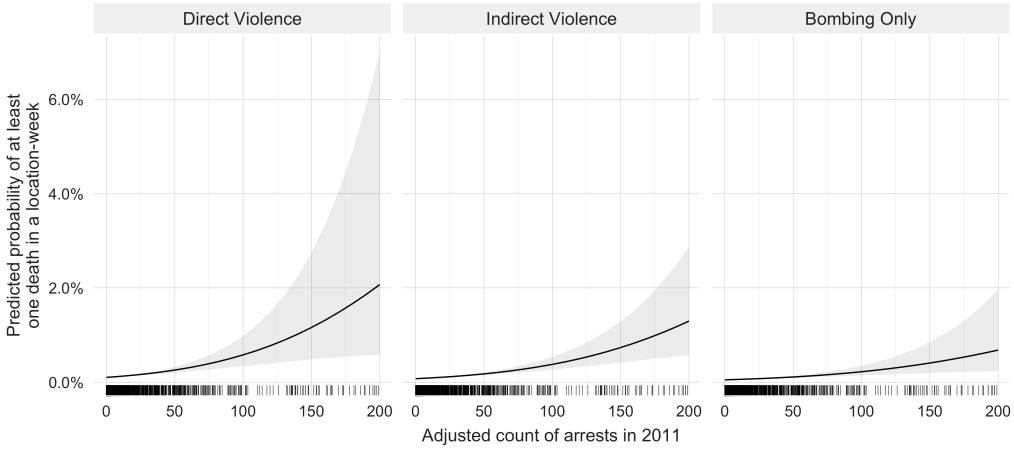


Figure 6: Number of 2011 arrests and predicted probabilities of a locale-week experiencing a civilian casualty. The arrests are adjusted to give localities partial credit for arrests in nearby localities. Models include controls for distance to the nearest "enemy" locale, the density of nearby localities, and the number of neighboring localities controlled by the same group. The ticks along the x-axis shows the distribution of localities by number of arrests. 95% confidence intervals accounting for clustering at the locale-level.

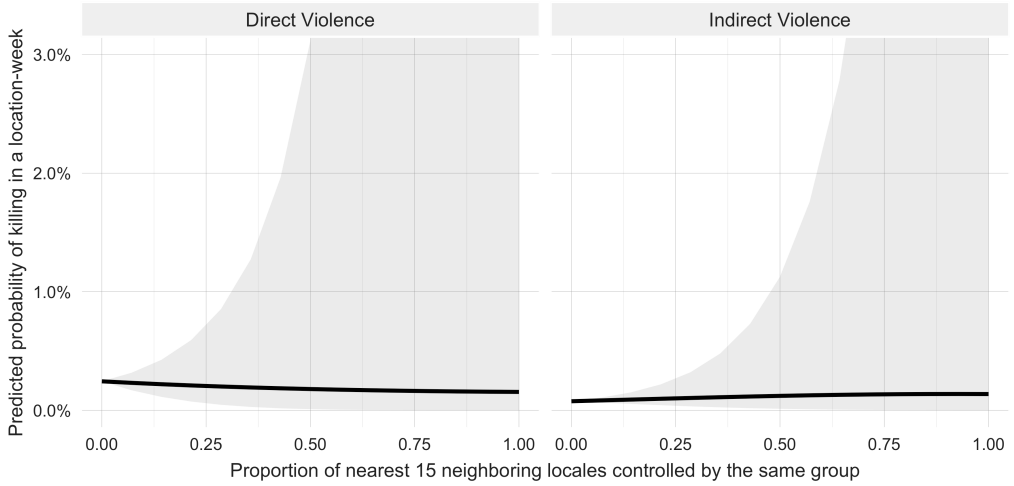


Figure 7: Degree of control and probability of a locale-week experiencing civilian casualties. Control measured with the proportion of "friendly" locales in the immediate neighborhood of 15 localities. Intelligence and territorial control theories would predict low levels of indirect violence throughout and the greatest amount of violence in mixed control areas. Instead, violence is not related to the proportion of nearby areas under the control of a different armed group. Shaded areas indicate 95% confidence intervals (truncated).

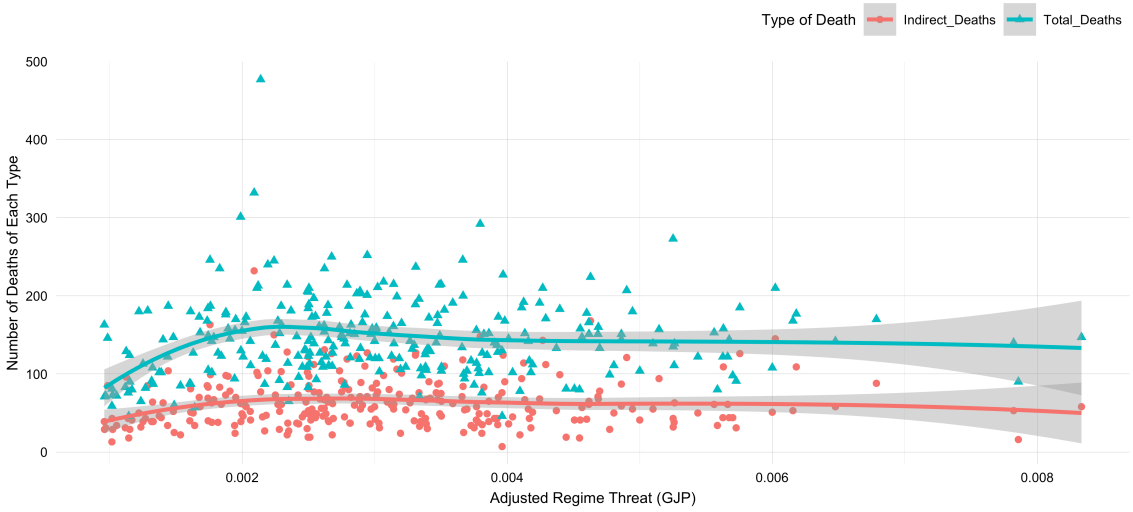


Figure 8: Duration-adjusted forecasts of Assad leaving power as a measure of regime threat, compared to the number of total and "indirect" (i.e. bombing and shelling) civilian causalities per day in Syria. The "desperation" or regime threat theory predicts a positive relationship between regime threat and violence, especially indiscriminate, against civilians. A loess line shows little relationship, however.

5.4 No variation by regime threat

The results also do not support the argument that civilian victimization increases as threats to the regime become greater and its survival seems more at stake (Figure 8). There is a positive relationship between threat and casualties, but only as the estimated threat to the regime increases from 0, with the predicted number of causalities per day increases, from around 100 to 200. After that initial increase, however, the number of daily causalities is constant with the estimated threat to the Assad regime. In fact, the relationship between daily regime threat and the number of daily indirect civilian casualties is nearly flat. It does not appear that the Assad regime resorts to violence more during periods when forecasters assess its survival probabilities as lower. The results are similar with a seven day lag for regime threat. This data supports the conclusion that violence is less about desperation than the existence of any threat at all.

5.5 No variation across direct and indirect causes of death

Finally, several theories of violence in civil war hypothesize that different causes of civilian death– "selective" or "indiscriminate" (Kalyvas 2006) or "direct and"indirect" (Balcells 2017)–should occur under different conditions. The intelligence–territorial control theory sees selective and indiscriminate violence as imperfect substitutes. Groups prefer selective violence, but sometimes resort to indiscriminate violence when attacking enemy areas. The data does not reveal major differences in relationships between different variables and direct and indirect violence. Proximity to the nearest enemy locality (Figure 5) shows the same relationship for both direct and indirect violence. There does not seem to be a substitution effect where forces employ direct violence when it is easier in proximate areas and resort to indirect violence only in more distant areas. Instead, violence of both kinds decreases as distance to enemy localities increases. Arrests have the same relationship with direct, indirect, and bombing. Balcells' theory suggests that bombing will be used particularly against areas with high pre-war opposition. Here, though, the relationship is the same for all forms of violence. Moreover, periods of higher regime threat do not show a marked

increase in indirect violence, as some theories would predict (Figure 8).

5.6 Implications for theory

These findings have several implications for existing theory. First, they indicate that pre-war politics have an important effect. The effect of arrests on subsequent violence clearly supports Balcell's "rivalry" argument, where forces during a war target areas of pre-war opposition. The violence is not purely endogenous to the conflict, but also depends on pre-war political mobilization. Wartime politics is perhaps less important than previous research suggested: the degree of territorial control does not have an effect separate from the distance to the front line and the level of regime threat also has little effect after an initial level of threat is reached.

Second, the results indicate that different types of violence, direct and indirect or selective and indiscriminate, may be more complementary than was previously assumed. Consequential components of several theories consist of distinctions between different modes of violence. Actors may prefer selective violence when possible, and may employ direct and indirect violence in different circumstances. The results from this analysis suggest that actors do less substitution than previous theories assumed. Direct and indirect violence in Syria is highly correlated across different values for pre-war arrests, distance to the front line, and the number of friendly localities in an location's neighborhood. All of this suggests that the distinction may not be as important as was previously thought.

Finally, the results point to possible regime-specific mechanisms. Balcell's emphasis on pre-war opposition was borne out here. Her more specific prediction about violence being especially common in areas of medium-level opposition were not, however. The probability of violence continues to increase with th number of arrests, with additional arrests having even larger effects. Balcells' theory was developed for explaining violence in Spain, a democracy at the time, which could result in a regime that has different incentives. Authoritarian regimes may respond differently to mobilization, with the greatest violence applied to areas

of greatest opposition. Authoritarian governments come to civil wars with a unique set of capabilities and incentives. As strong, security-oriented states, they often possess strong security forces and the intelligence and control to use them effectively.

Second, the data on regime threat may point to a threshold effect for regime threat. While the relationship between daily threat and civilian causalities is mostly flat, there is a non-linearity as threat increases above 0. Some regimes may treat any threat to survival as demanding a violent response, while others wait until they believe their situation is desparate. Both of these results suggest a need for greater attention to specific regime's decisionmaking.

6 Conclusion

These findings have important implications for scholarship on killing in civil war. First, it highlights the importance of meso theories of civil war killing that account for politics. Much of the existing work on civil war and killing in civil war has operated either at a high-level structural level (e.g., the large-n work of the early 2000s) or at the very micro-level, focusing on individual motivations and emotions. Second, it points to the need for better theories of regime type in civil war killing. Authoritarian states face different incentives than democracies or third-party governs in how they wage civil war. Regimes have incentives to use civil wars to reshape the post-war political order (Balcells 2017), but authoritarian states may operate under stronger incentives than democracies to eliminate their opponents.

Methodologically, the paper makes several contributions. The data provided here can inform future studies of Syria: the automated techniques for geocoding text and measuring territorial control make extensions of this work easy to implement. Neither method is specific to Syria or Arabic, meaning researchers should be able to take them to other conflicts. The panel dataset will be useful in other studies, including on spatial spillover, the uses of specific weapons, and deeper understanding of specific parts of the war.

The field on civilian victimization in civil war is expansive and growing. Testing existing

theories on new cases is a crucial component of advancing our understanding and reducing our overconfidence about the causes of violence against civilians in civil war. Understanding the causes of violence in Syria specifically helps us understand a conflict that has been especially deadly for civilians. In addition to the scholarly importance of studying Syria, future research may create the possibility of alleviating the suffering that results from civil war.

Notes

[1]For advice and comments on previous versions of this paper, I thank Fotini Christia, Stathis Kalyvas, In Song Kim, Chappell Lawson, Rich Nielsen, Sara Plana, Roger Peterson, and Rachel Tecott, and seminar participants at MIT and Oxford. Funding was provided by a National Science Foundation Graduate Research Fellowship. I also acknowledge the efforts of the Syrians who collect and provide data on the civil war, often at enormous personal cost. This paper would not be possible without their work.

[2]Phillip Connor and Jens Manuel Krogstad, "About six-in-ten Syrians are now displaced from their homes," Pew Research. http://www.pewresearch.org/fact-tank/2016/06/13/about-six-in-ten-syrians-are-now-displaced-from-their-homes/, 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

[3]Several of the datasets I draw on were collected by Syrian activists and NGOs during the war, often at great personal cost. This work would not have been possible without their efforts.

[4]Available at http://syrianshuhada.com/. A complete copy of the original dataset is available in the replication materials.

[5]Assuming a rapid 60 seconds per lookup, this would take approximately 150 hours of work.

[6]Further details are provided in the supporting information.

[7]"Barrel bomb kills Syrian boy nicknamed 'biscuit seller' ", 2 June 2014, *Al Arabiya*, https://english.alarabiya.net/en/webtv/2014/06/02/Barrel-bomb-kills-Syrian-boy-nicknamed-biscuit-seller--1364.html

[8]Syrian army launches Aleppo counter-offensive, activists say", July 28, 2012, *Agence France-Presse*, https://archive.ph/20130104052912/http://www.nowlebanon.com/NewsArticleDetails.aspx?ID=423149#selection-905.0-905.60

[9]The project is described at https://www.cartercenter.org/peace/conflict_resolution/syria-conflict-resolution.html, and the map is available at https://d3svb6mundity5.cloudfront.net/dashboard/index.html

[10]A neighborhood size of 15 was chosen qualitatively by examining average distances to neighbors and inspection of maps across urban and rural areas.

[11]See the supporting information for details on the distance algorithm used.

[12]In the 2007 presidential election, the last before the onset of war, only 0.18% of votes were against Bashar al-Assad, on 96% turnout, according to the International Foundation for Electoral Systems. http://www.electionguide.org/elections/id/110/

[13]More information and the original dataset are available at https://csr-sy.org/. A copy of the scraped dataset is available in the replication materials.

[14]I use the decay function $f(n, d) = n \cdot \frac{1}{1+\left(\frac{d}{2000}\right)^{2.8}}$, where n is the number of arrests and d is the distance in meters. An area within 500 meters gets full credit for each arrest, an area 2 kilometers away receives 0.5, and an area 10 km away receives about 1% of each arrest. More details on the function and parameters are provided in the supporting information.

[15]See, e.g., Spiegel, Alix. "So You Think You're Smarter Than A CIA Agent". NPR.org. https://www.npr.org/sections/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent, or Tetlock and Gardner (2016).

[16]The precise questions and aggregation strategy are available in the replication materials.

[17]Beck and Katz (2001) discuss dyad fixed effects, but the effect is the same with unit fixed effects in time series cross sectional data with binary outcomes.

7 References

Anderson, Noel. 2019. "Competitive Intervention, Protracted Conflict, and the Global Prevalence of Civil War." *International Studies Quarterly* 63 (3): 692–706.

Balcells, Laia. 2017. *Rivalry and Revenge: The Politics of Violence During Civil War*. Cambridge University Press.

Beck, Nathaniel, and Jonathan N Katz. 2001. "Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon." *International Organization* 55 (2): 487–95.

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

Berman, Eli, and Aila M. Matanock. 2015. "The Empiricists' Insurgency." *Annual Review of Political Science* 18 (1): 443–64.

Berman, Eli, Jacob N Shapiro, and Joseph H Felter. 2011. "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq." *Journal of Political Economy* 119 (4): 766–819.

Dell, Melissa, and Pablo Querubin. 2016. "Bombing the Way to State-Building? Lessons from the Vietnam War." *Working Paper*.

———. 2017. "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies." *The Quarterly Journal of Economics* 133 (2): 701–64.

Douglass, Rex W, and Kristen A Harkness. 2018. "Measuring the Landscape of Civil War: Evaluating Geographic Coding Decisions with Historic Data from the Mau Mau Rebellion." *Journal of Peace Research*.

Downes, Alexander B. 2007. "Draining the Sea by Filling the Graves: Investigating the Effectiveness of Indiscriminate Violence as a Counterinsurgency Strategy." *Civil Wars* 9 (4): 420–44.

———. 2008. *Targeting Civilians in War*. Cornell University Press.

Eck, Kristine, and Lisa Hultman. 2007. "One-Sided Violence Against Civilians in War: Insights from New Fatality Data." *Journal of Peace Research* 44 (2): 233–46.

Egami, Naoki. 2018. "Identification of Causal Diffusion Effects Using Stationary Causal Directed Acyclic Graphs." *arXiv Preprint arXiv:1810.07858*.

Fjelde, Hanne, and Lisa Hultman. 2014. "Weakening the Enemy: A Disaggregated Study of Violence Against Civilians in Africa." *Journal of Conflict Resolution* 58 (7): 1230–57.

Giglio, Mike. 2019. *Shatter the Nations: ISIS and the War for the Caliphate*. PublicAffairs.

Guha-Sapir, Debarati, Jose M Rodriguez-Llanes, Madelyn H Hicks, Anne-Françoise Donneau, Adam Coutts, Louis Lillywhite, and Fouad M Fouad. 2015. "Civilian Deaths from Weapons Used in the Syrian Conflict." *British Medical Journal* 351: h4736.

Guha-Sapir, Debarati, Benjamin Schlüter, Jose Manuel Rodriguez-Llanes, Louis Lillywhite, and Madelyn Hsiao-Rei Hicks. 2018. "Patterns of Civilian and Child Deaths Due to War-Related Violence in Syria: A Comparative Analysis from the Violation Documentation Center Dataset, 2011–16." *The Lancet Global Health* 6 (1): e103–e110.

Hegghammer, Thomas. 2013. "Syria's Foreign Fighters."

Hughes, Geraint Alun. 2014. "Syria and the Perils of Proxy Warfare." *Small Wars & Insurgencies* 25 (3): 522–38.

Hultman, Lisa. 2007. "Battle Losses and Rebel Violence: Raising the Costs for Fighting." *Terrorism and Political Violence* 19 (2): 205–22.

Kahle, David, and Hadley Wickham. 2013. "ggmap: Spatial Visualization with ggplot2." *The R Journal* 5 (1): 144–61. http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* 47 (2): 99–127.

Kalyvas, Stathis N. 2004. "The Urban Bias in Research on Civil Wars." *Security Studies* 13 (3): 160–90.

———. 2006. *The Logic of Violence in Civil War*. Cambridge University Press.

Kalyvas, Stathis N, and Laia Balcells. 2010. "International System and Technologies of Rebellion: How the End of the Cold War Shaped Internal Conflict." *American Political Science Review* 104 (03): 415–29.

Kilcullen, David. 2010. *Counterinsurgency*. Oxford University Press.

Kocher, Matthew Adam, Thomas B Pepinsky, and Stathis N Kalyvas. 2011. "Aerial Bombing and Counterinsurgency in the Vietnam War." *American Journal of Political Science* 55 (2): 201–18.

Koren, Ore, and Benjamin E Bagozzi. 2017. "Living Off the Land: The Connection Between Cropland, Food Security, and Violence Against Civilians." *Journal of Peace Research* 54 (3): 351–64.

Krcmaric, Daniel. 2018. "Varieties of Civil War and Mass Killing." *Journal of Peace Research* 55 (1): 18–31.

Lyall, Jason, Graeme Blair, and Kosuke Imai. 2013. "Explaining Support for Combatants During Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107 (04): 679–705.

Nagl, John A. 2002. *Learning to Eat Soup with a Knife: Counterinsurgency Lessons from Malaya and Vietnam*. Praeger Publishers.

Nagl, John A, James F Amos, Sarah Sewall, David H Petraeus, and others. 2008. *FM 3-24: The US Army/Marine Corps Counterinsurgency Field Manual*. FM 3-24. University of Chicago Press.

Pape, Robert A. 1996. *Bombing to Win: Air Power and Coercion in War*. Cornell University Press.

Petersen, Roger D. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. Cambridge University Press.

Price, Megan, Anita Gohdes, and Patrick Ball. 2014. "Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic." *Human Rights Data Analysis Group, Geneva. Available at https://hrdag.org/publications/full-updated-statistical-analysis-of-documentation-of-killings-in-the-syrian-arab-republic/.*

Rosenblatt, Nate. 2016. *All Jihad Is Local: What Isis' Files Tell Us About Its Fighters*. New America.

Shellman, Stephen M. 2006. "Process Matters: Conflict and Cooperation in Sequential Government-Dissident Interactions." *Security Studies* 15 (4): 563–99.

Sullivan, Christopher Michael. 2012. "Blood in the Village: A Local-Level Investigation of State Massacres." *Conflict Management and Peace Science* 29 (4): 373–96.

Sunderland, Riley. 1964. *Organizing Counterinsurgency in Malaya, 1947–1960*. RAND Corporation.

Tetlock, Philip E, and Dan Gardner. 2016. *Superforecasting: The Art and Science of Prediction*. Random House.

Valentino, Benjamin. 2000. "Final Solutions: The Causes of Mass Killing and Genocide." *Security Studies* 9 (3): 1–59.

Valentino, Benjamin, Paul Huth, and Dylan Balch-Lindsay. 2004. "'Draining the Sea': Mass Killing and Guerrilla Warfare." *International Organization* 58 (02): 375–407.

Valentino, Benjamin, Paul Huth, and Sarah Croco. 2006. "Covenants Without the Sword: International Law and the Protection of Civilians in Times of War." *World Politics* 58 (3): 339–77.

Walzer, Michael. 1977. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic books.

Warrick, Joby. 2015. *Black Flags: The Rise of Isis*. Anchor.

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

Zhukov, Yuri M. 2012. "Roads and the Diffusion of Insurgent Violence: The Logistics of Conflict in Russia's North Caucasus." *Political Geography* 31 (3): 144–56.

Supporting Information for "Violence against civilians in the Syrian civil war"

28 January 2020

Contents

1 Data on Civilian Causalities

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

[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

from March 2011 through , of which 151,873 are coded as civilians. The documentation 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]

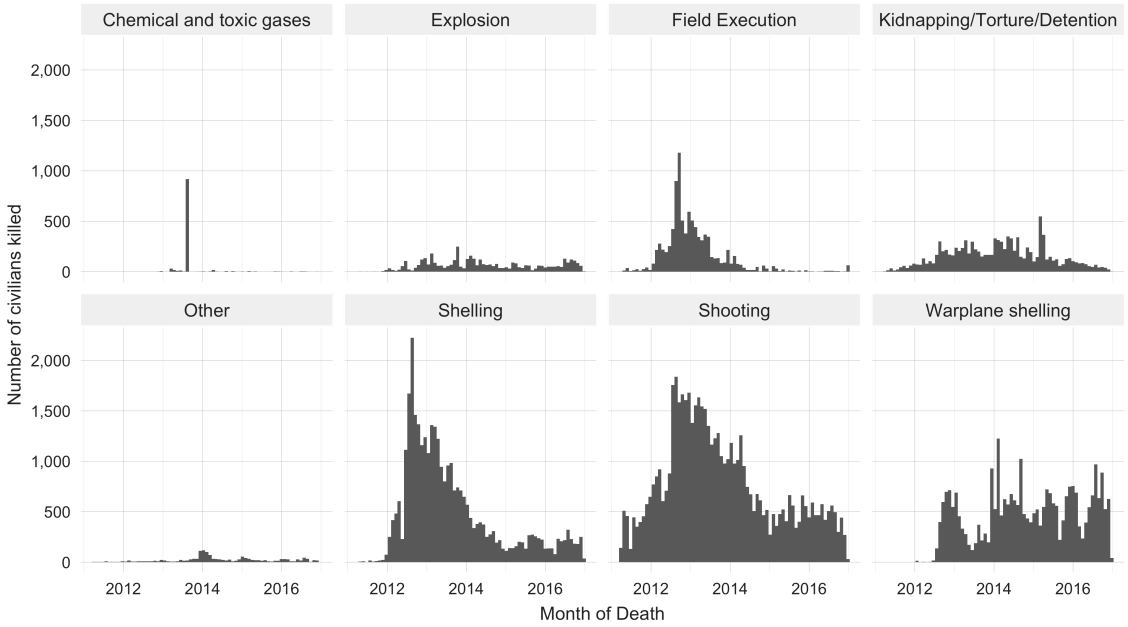


Figure 1: Causes of civilian death per month, VDC

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

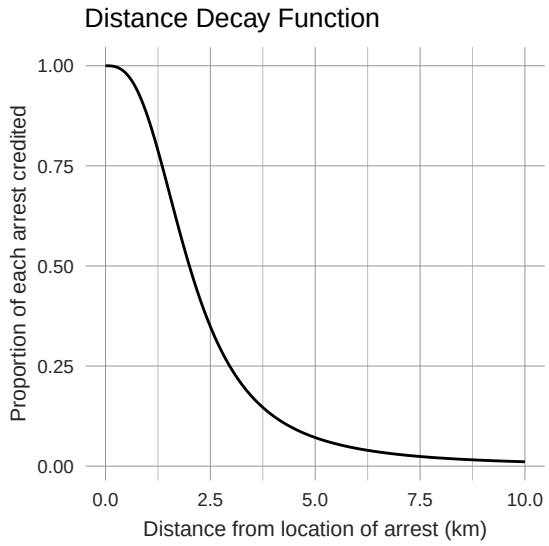
To provide a measure of where the Syrian government perceived the greatest challenge to its rule during the 2011 phase of the conflict, I obtained information on the locations of arrest in 2011. from the Syrian Center for Statistics and Research (CSR). This dataset includes a "town"/neighborhood field 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 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 geolocate the CSR arrests dataset using the same approach.

[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.

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.

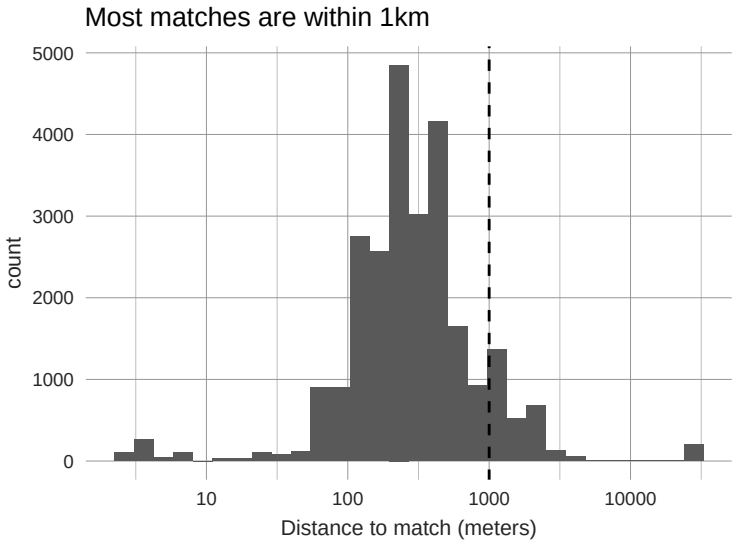
Second, I create an adjusted form of this measure that accounts for the median distance to point a's 15 nearest neighbors. This measure adjusts for the differing role of a kilometer in urban and rural areas: two kilometers is whole neighborhoods away in a city, but right next door in rural areas. The 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.

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

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, mean 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 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 sophisticated approach of setting the final day's probability to 0 and removing a linear time trend.

Table 1: Casualties geolocated 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

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).

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

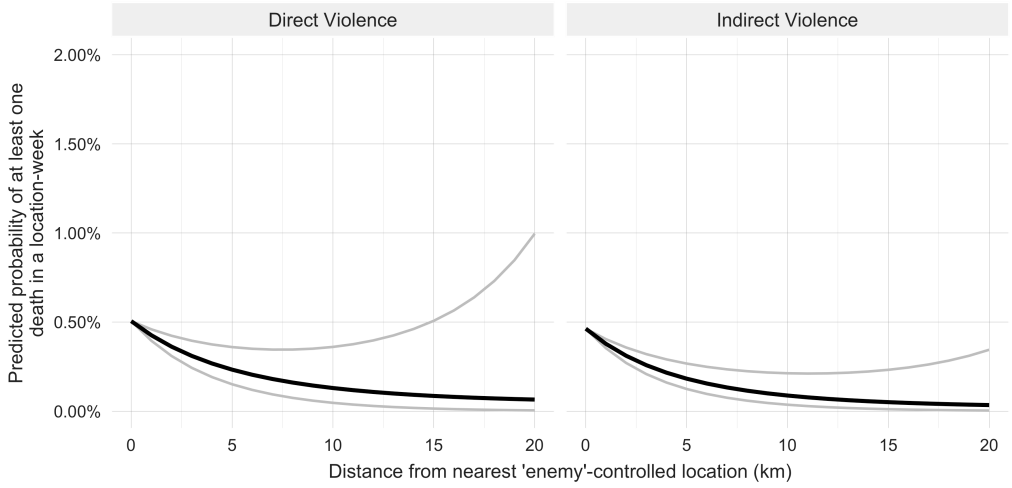


Figure 2: Relationship between civilian killing and distance to the nearest enemy-held area. Predicted probabilities of a locale-week experiencing a civilian casualty by varying distance to the nearest enemy area Computed with n = 44,521 civilian casualties from 2015 and 2016. 95% confidence intervals.

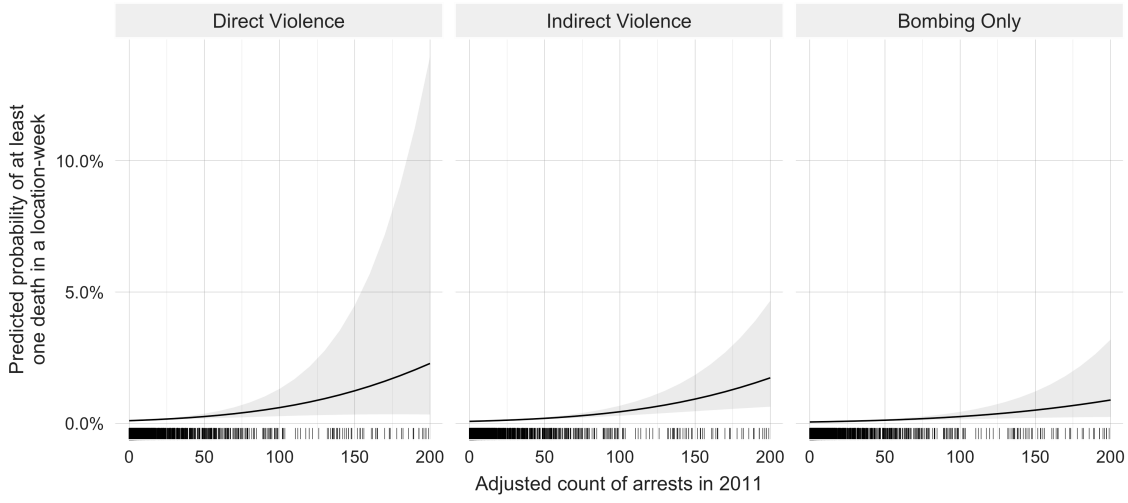


Figure 3: Predicted probabilites of a locale-week experiencing a civilian casualty with varying 2011 arrests. The "arrests only" model includes only 2011 distance-decayed arrests and its squared term.

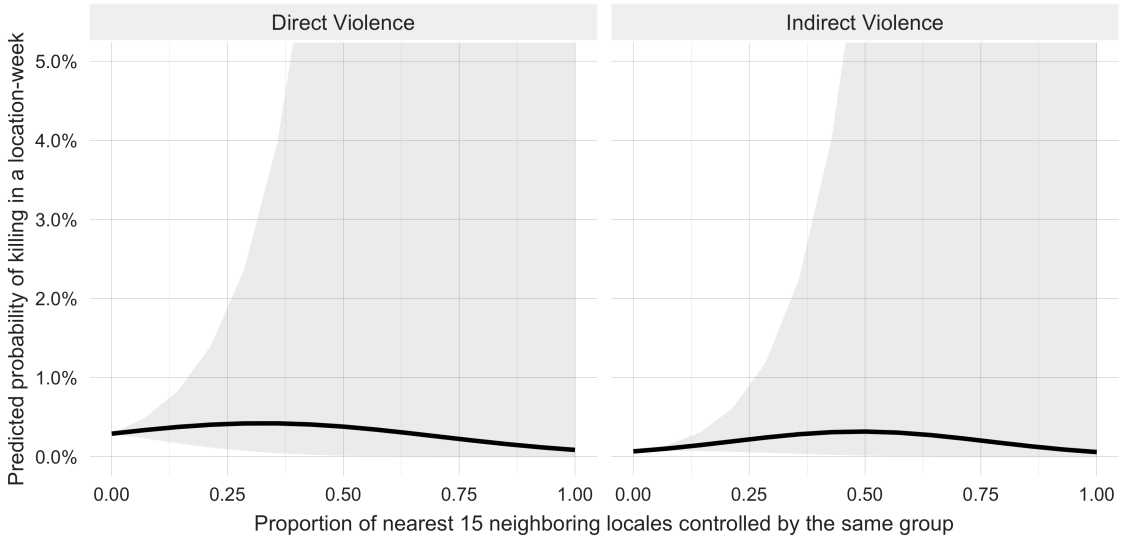


Figure 4: Relationship between civilian killing and proportion of "friendly" locales in the immediate neighborhood of 15 without controls. Computed with n = 44,521 civilian casualties from 2015 and 2016.

4 Regime threat alternative specifications

The weak relationship between regime threat and civilian casualities is robust to a lagged values of regime threat. Figure 5 shows the relationship between daily civilian casualities and estimated regime threat from seven days before and is mostly indistinguishable from the main figure in the paper.

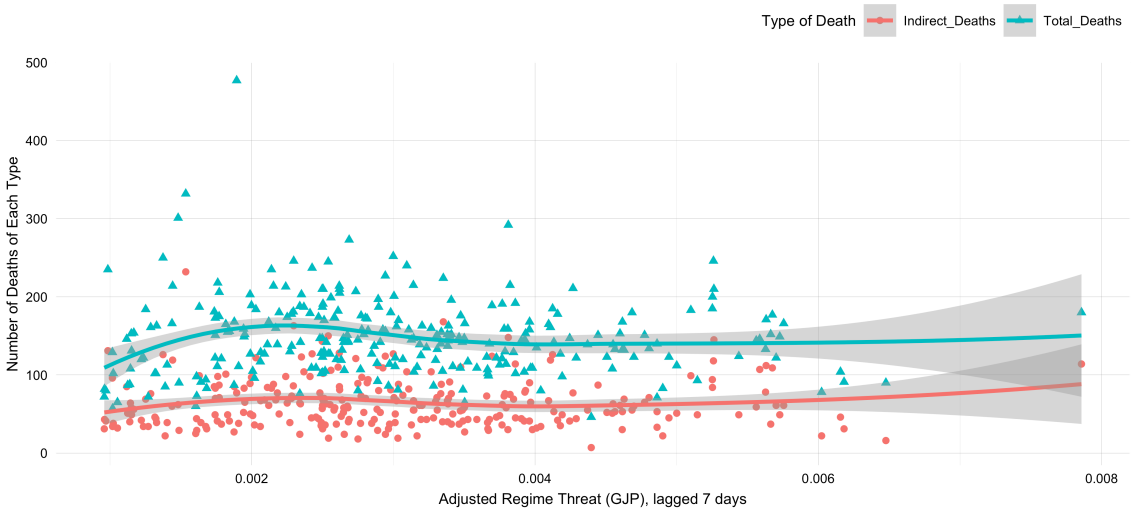


Figure 5: Relationship between civilian casualities and estimated regime threat from seven days before.

5 Note on Replication

The data analysis is performed in the provided `.Rmd` code using R version 3.6.0 running on macOS High Sierra 10.13.6. The code to transform and extend data is provided a Python file, which was 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: [REDACTED FOR ANONYMITY]

References

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

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