

1 SIGACTS Overview

Our validation data are drawn from the US military’s “significant activities” (SIGACTS) database. Over the course of recent US and Coalition wars in Iraq and Afghanistan, the SIGACTS sought to record lethal and non-lethal insurgent and military events in a standardized fashion. To this end, discrete events were recorded by US and/or Coalition forces based upon the type of event that transpired, the unique date and time of that event, a written narrative describing the event, an event’s associated fatalities, and that event’s geolocation. The latter geolocations were recorded by US and Coalition forces via GPS, thus offering far more accurate geographic information than events geolocated by researchers from news(wire) reports. This feature—and the internal (US military) audience of SIGACTS during its implementation—ensures substantially more coverage of violence in Iraq and Afghanistan than what is commonly reported on in media sources, leading SIGACTS to serve as an unrivaled gold standard source for the validation of media-derived event data (Weidmann 2015; Weidmann 2016). That being said, it has been widely noted that the SIGACTS data are relatively sparser in their coverage of US- and/or Coalition-initiated events, in comparison to insurgent-initiated events. As such, our validation assessment follows past research in focusing only on insurgent-directed violence. We furthermore focus on SIGACTS’ Iraq data, as opposed to the SIGACTS Afghanistan data, given the breadth of extant research using SIGACTS and our event datasets in the Iraq context. Below we elaborate upon these items.

2 Past Usage of SIGACTS

Berman, Shapiro, and Felter (2011) provide an early analysis of Iraq’s SIGACTS data. Their study considers a SIGACTS-derived dependent variable measuring population-normalized district-month insurgent attacks against coalition and government forces. The authors’ temporal analysis window encompasses February, 2004 through December, 2008. Iraqi districts are chosen as spatial units based upon their being the smallest unit where reliable population data exist. The authors then analyze the effect of reconstruction project spending on this SIGACTS measure while controlling for district fixed effects, year fixed effects (interacted with Sunni population shares), Shia population shares, unemployment, income, and past SIGACTS incidents (lagged by a half year). They also cluster their standard errors by district to “allow errors to be correlated temporally” (fn 32, p. 796). Following mention of this, Berman, Shapiro, and Felter concede that such errors may also be spatially correlated across districts due to rebel and government strategies being coordinated at higher spatial scales—something that they do not account for, aside from a brief mention of a robustness model that alternatively clusters standard errors at the province level (fn 32, p. 796).

Condra and Shapiro use Iraq’s SIGACTS data to study the impacts of violence against civilians on subsequent insurgent violence at the district-week level. They operationalize the latter violence using population weighted counts of insurgent attacks derived from the MNF-I SIGACTS III data for the period February 4, 2004 through February 24, 2009. They then assess the effects of (lagged differences in) civilian killings¹ on their population-weighted insurgent attack dependent variable via linear models. These models include space-time fixed

¹Measured from Iraq Body Count (IBC) data.

effects, controls for population density and unemployment, and district-clustered standard errors. In their Supplemental Materials, the authors also reassess their core results when including a spatial lag of their dependent variable on the righthand side of their regressions, finding that some results become statistically weaker when doing so (p. 187). In a similar vein, Hagan et al. (2016) employ population-normalized measures of insurgent-initiated violence derived from SIGACTS in (i) a series of multi-level models of legal cynicism and (ii) spatial lag models of SIGACTS violence. The authors' SIGACTS data are aggregated to the district level for distinct targets, albeit with little temporal variation.

Several additional studies of Iraq more explicitly consider spatial(-temporal) conflict processes via analyses of SIGACTS. Johnson and Braithwaite (2009) and Braithwaite and Johnson (2012) utilize multiple SIGACTS-derived datasets on insurgent and counter-insurgent activities in Iraq for the period January 1st to June 30th 2005 to assess the efficacy of various COIN activities. In their analyses, they implement univariate and bivariate space-time analyses of conflict clustering patterns for SIGACTS data aggregated to 500-meter grids and one week intervals. Linke et al. (2012) conduct space-time Granger analyses of Coalition and insurgent violence dynamics for the Iraq War from 2004-2009. These analyses apply negative binomial count models to SIGACTS-aggregated conflict counts at the grid- three-day-level for both Iraq on the whole and Baghdad specifically. The authors use queen contiguity to calculate first- and second-order spatial lags, and include these in their models alongside multiple temporal lags. Weidmann and Salyehan (2016) similarly analyze SIGACTS-based event counts across 85 distinct Baghdad neighborhoods. To do so, they employ a computational model to specifically evaluate the competing effects of counterinsurgency and ethnic composition on violence in Baghdad.

Silverman (2020) studies the efficacy of Coalition-directed condolence payments² during the Iraq War. He analyzes such payments during the February, 2004 through December, 2008 period at the district-half year level of analysis, finding that such payments reliably reduce local rates of insurgent violence. For this analysis, Silverman uses SIGACTS to construct a "SIGACTS per 1,000 residents" dependent variable measuring insurgent attacks against Coalition forces, the Iraqi national government, Iraqi Security Forces, and civilians.³ Silverman divides his primary (condolence payments) independent variables by district population as well, and then controls for separate measures of coalition collateral damage, insurgent collateral damage, and several additional measures of Coalition spending and troop presence, half year fixed effects, district fixed effects, and Sunni \times half year fixed effects within a series of first differenced linear regressions. Each estimated model clusters standard errors by district, and robustness models then report results when all data are aggregated to the weekly and monthly (as opposed to half-year) levels.

Shaver and Shapiro (2021) study the impact of violence against civilians on counterinsurgency informing in Iraq. Using province-week data for the period June 1, 2007 to July 18, 2008, they find that government-directed (insurgent-directed) violence against civilians decreases (increases) such informing. Their violence measures in both cases are drawn from SIGACTS. For their analysis, the authors account for unobserved province-specific trends via a first-differences model that considers the province-level changes in their dependent and

²I.e., to civilian victims.

³Though the author notes that this measure likely undercounts attacks against civilians (p. 855).

independent count variables from week $t - 1$ to week t . Shaver and Shapiro (2021) then (i) include week fixed effects to capture unobserved week-specific effects and (ii) cluster their standard errors at the province level “to take into account the potential for serial correlation” (p. 19). Finally, Shapiro and Weidmann (2015) use a first difference approach to analyze the effects of cellphone coverage on district-month (and district-tower) insurgent attacks in Iraq. The latter (insurgent attack) dependent variable was normalized by population after being derived from SIGACTS for the February, 2004 through January, 2009 period. The authors include district and month fixed effects throughout their specifications, alongside district-clustered standard errors. For robustness, they also “include a spatial lag of the dependent variable to account in a rough way for spatial autocorrelation” (p. 265).

The above review only highlights a subset of the many studies that have employed Iraq’s SIGACTS data to date. Alongside the research reviewed above, a number of additional studies utilize SIGACTS’ Afghanistan data (e.g., Condra et al. 2018; Trebbi et al. 2020). Of note, several of these latter studies have used Afghanistan’s SIGACTS data for validation purposes, including validation in relation to GED (Weidmann 2015; Weidmann 2016).

3 Past Usage of GED and ICEWS to Study Iraq

A longstanding body of research has employed political event data for the study of conflict in the Middle East (e.g., Schrodtt and Gerner 1994, 1997, 2004; Goldstein et al. 2001). Building upon this seminal work, Iraq has come to feature prominently in contemporary studies utilizing GED and/or ICEWS. With respect to ICEWS, this includes analyses of the effects of the Islamic State of Iraq and the Levant (ISIL) on diplomatic relations in the Middle East (Tellez and Roberts 2019), studies of protest mobilization during the Arab Spring (Steinert-Threlkeld 2017), and efforts to predict forced migration in Iraq (Singh et al. 2019). Regarding GED, recent work has leveraged these data to report on contemporary trends in organized violence within Iraq and several additional select conflict hotspots (Pettersson and Öberg 2020), incorporated the GED into analyses of the interplay between fire, conflict, and land within Iraq’s Kurdistan region (Eklund et al. 2021), and considered the linkages between military activity and pollution in Iraq (Chudnovsky and Kostinski 2020).

4 Event Data Formatting and Aggregation

In this section, we describe our formatting decisions for SIGACTS, ICEWS (Boshee et al. 2016), and GED (Sundberg and Melander 2013). SIGACTS relies on Coalition forces to geolocate events to specific latitude-longitude coordinates via GPS technology. For our study, we utilize Silverman’s (2020) already-aggregated SIGACTS data. Per Silverman (2020, 855), these corresponding SIGACTS events denote insurgent attacks against Coalition and Iraqi security forces, against the Iraqi national government, and against civilians. Silverman’s data were assigned to individual Iraqi districts based upon their original latitude-longitude coordinates for half year periods from February 2004 through December 2008. The resultant half year-district SIGACTS event counts were then normalized by district population levels so as to denote SIGACTS rates per 1,000 residents within each individual district-half year

observation in our data. We utilize this population-normalized data as the gold standard records for validation purposes within our own analyses. This ensures that the district-half year period is our unit of observation for all ensuing comparisons.⁴

We next turn to our (machine coded) ICEWS data. Using large corpus of local and international news(wire) reports, ICEWS separately codes events in English, Spanish, French, and Portuguese for all countries of the world excluding purely domestic US events. Date ranges are specified alongside search query terms and discard terms. These steps employ mild deduplication during to minimize multiple stories from the same publisher, same headline, and same date (ICEWS 2015). ICEWS then applies automated event and actor extraction routines to code relevant socio-political events. This entails the application of shallow parsing—and related dictionary-based event and actor codings—to extract source and target actors and action types from the first six sentences of each news(wire) report considered. Event action types are coded based upon the CAMEO project’s 300-plus event type taxonomy (Schrodt, Gerner, and Yilmaz 2009). Identified source and target actors are evaluated against ICEWS’ own specialized entity dictionaries, which encompass over 50,000 named and time-indexed entities and over 700 generic agent names for matches to relevant source/target actors and country-assignments (Lockheed Martin 2021).

ICEWS then uses automated methods to geolocate identified events that were identified and coded via a hybrid approach that leverages dictionaries of location names and string matching algorithms. Named entity resolution (NER) is then used to match each candidate location name to a gazetteer’s geolocation names and coordinates, along with location specificity precision. The single most likely geolocation based upon a statistical ranking of all NER-identified matches is then assigned as an event’s geolocation (Lee, Liu, and Ward 2018). Evaluations of ICEWS’ subnational geolocation accuracy has previously determined that ICEWS geolocates roughly 85% of all events to the subnational level, and exhibits a subnational geolocation accuracy rate (i.e., to an appropriate country) of 78% (Lautenschlager, Starz, and Warfield 2017).

We used several of the extracted event fields described above to subset and aggregate the ICEWS’ data in a manner that maximized comparability to our SIGACTS violence measure. We first retained only those ICEWS events transpiring in Iraq for the February 2004 through December 2008 period. We then subset all resultant ICEWS events to include only those events that involved (i) “rebel,” “separatist,” and/or “insurgent” source actors and (ii) target actors associated with non-insurgent domestic and international state and non-state groups (excluding insurgents, rebels, criminals, and separatists). We then kept only those retained events that were geolocated to the city/town level(s) of geographic precision.⁵ With these city-specific Iraq events in hand, we next retained only ICEWS’ CAMEO category 18 (ASSAULT), category 19 (FIGHT), and category 20 (USE UNCONVENTIONAL MASS VIOLENCE) events with the following three or four digit CAMEO codes:

- 180: Use unconventional violence, not specified below
- 181: Abduct, hijack, or take hostage
- 182: Physically assault, not specified below

⁴Though for robustness, we also consider the district-month period of analysis.

⁵This was achieved by dropping events with missing information on ICEWS’ “city” field.

- 1821: Sexually assault
- 1822: Torture
- 1823: Kill by physical assault
- 183: Conduct suicide, car, or other non-military bombing, not specified below
- 1831: Carry out suicide bombing
- 1832: Carry out car bombing
- 1833: Carry out roadside bombing
- 184: Use as human shield
- 185: Attempt to assassinate
- 186: Assassinate
- 190: Use conventional military force, not specified below
- 191: Impose blockade, restrict movement
- 192: Occupy territory
- 193: Fight with small arms and light weapons
- 194: Fight with artillery and tanks
- 195: Employ aerial weapons
- 196: Violate ceasefire
- 203: Engage in ethnic cleansing
- 200: Use unconventional mass violence, not specified below
- 201: Engage in mass expulsion
- 202: Engage in mass killings
- 203: Engage in ethnic cleansing

Following this, we de-duplicated all remaining events to ensure that only one event(-type) was recorded per day, source, target, action type, and latitude-longitude coordinate. We next aggregated our deduplicated events to district half-year level counts by matching these events to shapefiles of Iraq’s district via latitude longitude coordinates. Finally, we divided each district half-year count by district population in a comparable manner to Silverman’s SIGACTS measure.

We then subset and aggregated the GED in a similar fashion. The GED is a (near-global) human-coded event dataset that draws on both news(wire) sources and non-governmental organization reports for its coding of individual events. A core set of English-language global newswires are first identified and human coded for relevant events. Based upon the degree of country and/or conflict coverage obtained from this step, a second evaluation can then be used to identify and add additional local coverage of a relevant conflict and/or country. In both cases, articles and reports are retrieved for coding in correspondence with the UCDP’s more broadly identified conflicts across the world, which require a threshold of at least 25 battle-related deaths for consideration (Sundberg and Melander 2013). However, proximate conflict-country years (ie., non-active years) that fall below this aggregate threshold are also coded for events by GED (Crociu and Sundberg 2015). Contra to ICEWS, events are then only coded if they exhibit at least one fatality and identifiable source and target actors.

GED uses human coders to identify relevant strings for geo-coding. Once these strings are identified, GED geolocates events via an implementation of NER in a human-directed and/or semi-automated fashion. In these cases, GED relies on a hybrid set of gazetteers as a reference set for georeferencing identified events, which it complements with local and historical maps

(Sundberg and Melander, 2013; Crociu and Sundberg 2015). These sources appear more expansive than that which is described for ICEWS above. GED’s geolocation routines then either rely on human coders to geolocate events or on semi-automated geolocation of events. In the latter case, “semi-automatic geocoding is employed in a number of cases (mainly in Europe and the former Soviet Union), using Google Geocoding API, Yandex, and Bing” with subsequent human and automatic vetting procedures to verify geolocation accuracy (Crociu and Sundberg 2015, 23). For GED, standardized geolocations are retained for each event’s single most precisely mentioned location. Finally, information on administrative divisions and geolocation precision are then added. Together, these steps are likely to ensure more precise geolocations than what is obtained in ICEWS.

To format our own GED-based events, we started by subsetting the GED to encompass only Iraq-located events for the February 2004 through December 2008 period. We next retained all non-state perpetrated cases of violence against civilians alongside comparable violence against government and military actors, while taking care to verify that our non-state source actors in this case encompassed only rebel and insurgent conflict initiators. These fatal, insurgent-directed events were then merged to Iraq district-level templates for our period of interest based upon latitude-longitude coordinates, while taking care to omit any GED events whose levels of geocoding accuracy were determined to be too ambiguous to fall within the district level.⁶ After these steps were complete, we generated counts of our retained GED events for each district-half-year unit during the 2004-2008 window. For our analysis, these counts were then divided by district population to generate a “GED events per 1000 persons” for comparison to our Iraq SIGACTS and ICEWS measures.

⁶Specifically, we only retained events that GED indicated were either (i) geolocated within 25km of a known location or (ii) whose exact location was recorded with latitude and longitude coordinates.

5 References

- Berman, E. and J.N. Shapiro, and J.H. Felter. 2011. "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq." *Journal of Political Economy* 119(4): 766-819.
- Boshee, E., J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward. 2016. "ICEWS Coded Event Data." <http://dx.doi.org/10.7910/DVN/28075>. Harvard Dataverse.
- Braithwaite, A. and S.D. Johnson. 2012. "Space-Time Modeling of Insurgency and Counterinsurgency in Iraq." *Journal of Quantitative Criminology* 28: 31-48.
- Chudnovsky, A. and A. Kostinski. 2020. "Secular Changes in Atmospheric Turbidity over Iraq and a Possible Link to Military Activity." *Remote Sensing* 12, 1526.
- Condra, L.N. and J.N. Shapiro. 2012. "Who Takes the Blame? The Strategic Effects of Collateral Damage." *American Journal of Political Science* 56(1): 167-187.
- Condra, L.N. J. Long, J. Shaver, and A. Wright. 2018. "The Logic of Insurgent Electoral Violence." *American Economic Review* 108(11): 3199-3231.
- Croicu, M. and R. Sundberg. 2015. "UCDP GED Codebook version 4.0," Department of Peace and Conflict Research, Uppsala University.
- Eklund, L., A.M. Abdi, A. Shahpurwala, and P. Dinc. 2021. "On the Geopolitics of Fire, Conflict and Land in the Kurdistan Region of Iraq." *Remote Sensing* 13, 1575.
- Goldstein, J.S., J.C. Pevehouse, D.J. Gerner, and S. Telhami. 2001. "Reciprocity, Triangularity, and Cooperation in the Middle East, 1979-97." *Journal of Conflict Resolution* 45(5), 594-620.
- Hagan, J., J. Kaiser, and A. Hanson. 2016. "The Theory of Legal Cynicism and Sunni Insurgent Violence in Post-Invasion Iraq." *American Sociology Review* 81(2), 316-346.
- Johnson, S. and A. Braithwaite. 2009. "Spatio-temporal distribution of insurgency in Iraq." In: Freilich J. and G. Newman (eds). Countering terrorism through SCP. *Crime Prevention Studies* 25, 9-32.
- Lautenschlager, J., J. Starz, and I. Warfield. 2017. "A statistical approach to the subnational geolocation of event data." In *Advances in Cross-Cultural Decision Making*, S. Schatz and M. Hoffman, Editor. Switzerland: Springer International Publishing.
- Lee, S., H. Liu, and M. Ward. 2019. "Lost in Space: Geolocation in Event Data." *Political Science Research and Methods* 7(4): 871-888.
- Linke, A.M., F.D.W. Witmer, and J. O'Loughlin. 2012. "Space-Time Granger Analysis of the War in Iraq: A Study of Coalition and Insurgent Action-Reaction." *International Interactions* 38(4): 402-425.
- Pettersson, T. and M. Öberg. 2020. "Organized Violence, 1989-2019." *Journal of Peace Research* 57(4): 597-613.
- Schrodt, P. and D. Gerner. 1994. "Validity Assessment of a Machine-coded Event Data Set For the Middle East, 1982-1992." *American Journal of Political Science* 38: 825-854.
- Schrodt, P. and D. Gerner. 1997. "Empirical Indicators of Crisis Phase in the Middle East, 1979-1995." *Journal of Conflict Resolution* 41(4): 529-552.

- Schrodt, P. and D. Gerner. 2004. "An Event Data Analysis of Third-Party Mediation in the Middle East and Balkans." *Journal of Conflict Resolution* 48(3): 310-330.
- Schrodt, P., D. Gerner, and O. Yilmaz. 2009. "Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World." In *International Conflict Mediation: New Approaches and Findings*. J. Bercovitch and S. Gartner (ed.). Routledge: New York.
- Shapiro, J.N. and N.B. Weidmann. 2015. "Is the Phone Mightier Than the Sword? Cell-phones and Insurgent Violence in Iraq." *International Organization* 69(2): 247-274.
- Shaver, A. and J.N. Shapiro. 2021. "The Effect of Civilian Casualties on Wartime Informing: Evidence from the Iraq War." *Journal of Conflict Resolution* 1-41.
- Silverman, D. 2020. "Too Late to Apologize? Collateral Damage, Post-Harm Compensation, and Insurgent Violence in Iraq." *International Organization* 74(4): 853-873.
- Singh, L, L. Wahedi, Y. Wang, Y. Wei, C. Kirov, S. Martin, K. Donato, Y. Liu, and K. Kawintiranon. 2019. "Blending Noisy Social Media Signals with Traditional Movement Variables to Predict Forced Migration." *KDD '19*, August 4-8, 2019, Anchorage, AK, USA: 1975-1983.
- Steinert-Threlkeld, Z. 2017. "Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring." *American Political Science Review* 111(2): 379-403.
- Sundberg, R. and E. Melander. 2013. "Introducing the UCDP georeferenced event dataset." *Journal of Peace Research* 50(4): 523-532.
- Tellez, J. and J. Roberts. 2019. "The Rise of the Islamic State and Changing Patterns of Cooperation in the Middle East." *International Interactions* 43(3): 560-575.
- Trebbi, F., E. Weese, A. Wright, A. Shaver. 2020. "Insurgent Learning." *Journal of Political Institutions and Political Economy* 1(3): 417-448.
- Weidmann, N.B. 2015. "On the Accuracy of Media-based Conflict Event Data." *Journal of Conflict Resolution* 59(6): 1129-1149.
- Weidmann, N.B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60(1): 206-218.
- Weidmann, N.B and I. Salehyan. 2013. "Violence and Ethnic Segregation: A Computational Model Applied to Baghdad." *International Studies Quarterly* 57(1): 52-64.