March 2012

# ESOC – Iraq Civil War Dataset

## Codebook for Version 3.0

Eli Berman
Luke N. Condra
Joseph H. Felter
Jacob N. Shapiro

Empirical Study of Conflict (ESOC) Project

## A Note to Users

This document serves as the codebook for the Empirical Study of Conflict Project's Iraq War dataset (ESOC-I), Version 3.

Users of the violence and reconstruction data should cite the article in which those data are first presented:

Eli Berman, Jacob N. Shapiro, and Joseph H. Felter, "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq,"*Journal of Political* Economy 119, no. 4 (August 2011).

Users of the civilian casualty data should cite the article in which those data are first presented:

Luke N. Condra and Jacob N. Shapiro, "Who Takes the Blame? The Strategic Effects of Collateral Damage," *American Journal of Political Science* 56, no. 1 (January 2012): 167-87.

Users of these data should contact Jacob Shapiro (jns@princeton.edu) with comments or corrections so that this dataset may be improved in further versions.

These data will be periodically augmented and updated. Users interested in calculating spatial lags of variables can contact ESOC for an adjacency matrix and Stata code to build spatial lags.

Joshua Borkowski, Zeynep Bulutgil, Tiffany Chou, Rob Fargo, Christopher Paik, Thomas Scherer, Peter Schram, Erin Troland, Choon Wang, and Nils Weidmann provided invaluable assistance at different points in developing these data.

**GIS Files**

Iraq_district_boundaries_UTM.shp
Iraq_governorate_boundaries_UTM.shp

We provide two sets of GIS data files:

1. District shape files that describe the district boundaries we used in Berman, Shapiro, Felter (2011). These boundaries were created to match, as much as possible, those used for the 2003, 2005, and 2007 World Food Program Food Security and Vulnerability Analysis.
2. Governorate shape files.

Users should note that many publicly available shape files for Iraq created before 2008 do not deal appropriately with the boundaries of Baghdad governorate that were changed substantially under the Coalition Provisional Authority.

As a general point analysts using pre-calculated time-series data on governorate or district level covariates for Iraq should be aware that the physical areas covered by named areas—Erbil district, for example—change substantially over time. Aggregating socio-economic data from published reports should therefore be done with care, particularly for the Kurdish regions and areas bordering them.

Background information

Our boundary files are designed to conform, as closely as possible, to the boundaries of districts and governorates recognized and observed in the household surveys administered by the United Nations World Food Programme in the second half of 2003 and published in *Baseline Food Security Analysis in Iraq* (WFP, 2004). This conformance is critical for weighting by district or governorate population. The 2003 WFP survey provides ante-bellum district-level population statistics by extrapolating known growth rates from the 1997 Iraqi census.

Unfortunately, despite repeated attempts, we were unable to secure the actual GIS files used by the WFP when enumerators administered the first household survey in 2003. We took two steps to match the physical boundaries in our files with those used by the WFP survey teams. First, we compared the boundary files we obtained from the Humanitarian Information Centre (HIC), an initiative of the UN Office for the Coordination of Humanitarian Affairs (OCHA), to the pictorial representations in the WFP survey report. Several districts in the HIC file were reassigned to different governorates based on this comparison or were divided in two to align with the WFP maps. Second, we communicated with the WFP Iraq Country Office in Amman. Their senior staff were reviewed our boundaries and confirm that, in their opinion, our boundaries matched those used by the WFP.

For areas not covered by WFP surveys we combined various sources to generate district boundaries that matched population figures provided by the Iraqi Kurdistan Regional Office and the NGO Coordination Committee in Iraq (NCCI). These were the only

population figures available for Kurdish areas as the 1997 Iraq census did not cover the three Kurdish governorates.
Since we originally constructed these files the government of Iraq has created a number of additional districts, mostly in the Kurdish areas. Analysts finding discrepancies between more recent shape files and ours should be aware that some processing of newer data may be required to match the district boundaries we employ.

**Violence Data**

esoc-iraq-v3_sigact_country-half.dta
esoc-iraq-v3_sigact_country-month.dta
esoc-iraq-v3_sigact_country-quarter.dta
esoc-iraq-v3_sigact_country-week.dta
esoc-iraq-v3_sigact_country-year.dta
esoc-iraq-v3_sigact_district-half.dta
esoc-iraq-v3_sigact_district-month.dta
esoc-iraq-v3_sigact_district-quarter.dta
esoc-iraq-v3_sigact_district-week.dta
esoc-iraq-v3_sigact_district-year.dta
esoc-iraq-v3_sigact_governorate-half.dta
esoc-iraq-v3_sigact_governorate-month.dta
esoc-iraq-v3_sigact_governorate-quarter.dta
esoc-iraq-v3_sigact_governorate-week.dta
esoc-iraq-v3_sigact_governorate-week.dta
esoc-iraq-v3_sigact_governorate-year.dta
esoc-iraq-v3_sigact_governorate-year.dta

The files we provide contain the total number of incidents recorded in the Multi-National Forces Iraq (MNF-I) SIGACT III database for a country, governorate, or district/time-period (week, month, quarter, half-year, and year). The source data were cleaned before aggregation. We removed possible duplicate incidents by dropping all incidents that had the same values on all variables.

The 'Significant activity' (SIGACT) reports by Coalition forces capture a wide variety of information about "…executed enemy attacks targeted against coalition, Iraqi Security Forces (ISF), civilians, Iraqi infrastructure and government organizations."[1] Unclassified data drawn from the MNF-I SIGACTS III Database provide the location, date, and time of attack incidents between February 2004 and February 2009. The unclassified data do not include any information pertaining to the Coalition Force units involved, Coalition Force casualties or battle damage incurred as a result of the reported incidents. Moreover, the data do not include successful coalition-initiated events such as raids where no one returned fire, or coalition-initiated indirect fire attacks not triggered by an initiating insurgent attack.

The SIGACT-derived violence data have limitations that any analysis must take into account. First, they capture violence against civilians and between non-state actors only when US forces are present and so dramatically undercount sectarian violence (GAO 2007, Fischer 2008, DOD 2007).[2] Second, several potentially useful variables in the data,

[1] GAO (2007), DOD (2008). The information provided in the Unclassified SIGACT data are limited to the fact of and type of terrorist/ insurgent attacks (including improvised explosive devices [IEDs]) and the estimated date and location they occurred.

[2] To address this weakness we have also collected geo-located data on civilian casualties recorded in the Iraq Body Count database. For 2006 the bivariate correlation between SIGACTs and incidents of civilian killings is approximately .855 at the governorate/month level. The correlation is lower at the district/month

type of attack and target of attack for example, are inconsistently coded over time. Third, these data almost certainly suffer from significant measurement error, though we have found little evidence that the error is non-random with respect to either aid spending or civilian casualties.[3]

We provide filtered and unfiltered versions of the SIGACT-III data in these replication data. Users should be aware that public reports by MNF-I and Multi-National Corps Iraq (MNC-I) apply different filters to the SIGACT-III data to generate their incident counts. Some of those filters involved data fields that remain classified and so cannot be reconstructed.

Using the SIGACT-III dataset, we created the following variables:

(1) *SIGACT* - The number of attacks recorded, includes all types of attack incidents.
(2) *SIG_1* – Number of attacks recorded, excluding incidents positively identified as criminal attacks, those in which civilians or insurgents were the primary target, and those related to sectarian violence, looting, and the like. This filer is designed to match as closely as possible those applied by MNF-I in 2007 to measure rates of insurgent attacks.
(3) *ied_total* – Total number of IED events recorded. This variable is consistent over the entire time-period, so analyses of IED attacks for the entire war should use it.
(4) *ied_attack* – Number of IEDs that exploded before being detected or defused. Only valid after September 1, 2006.
(5) *ied_clear* – Number of IEDs recorded as found and cleared. Only valid after September 1, 2006.
(6) *df* – Number of direct fire incidents recorded.
(7) *idf* – Number of indirect fire incidents recorded.
(8) *suicide* – Number of suicide bombings recorded.

level, .541, because many incidents of civilian killings in Baghdad governorate cannot be precisely located. The rate of undercounting at the governorate level is statistically significantly greater in mixed and Shi'ite governorates than in Sunni governorates. In mixed governorates this is likely due to the high rate of sectarian violence. In Shi'ite governorates the Coalition presence is less dense.

[3] Kilcullen (2008) reports that attempts to reconcile the SIGACT data with unit leaders' recollections show the accuracy of the data varies widely by unit. One source of these discrepancies is that the element responsibility for making initial SIGACT reports varies across units and over time. We should expect, for example, different reporting biases from a company headquarters than from a battalion intelligence officer (S-2).

**Reconstruction Data**

maaws_20091002_07feb08districts_public.dta
aggregate_maaws_26MAR12.do

The reconstruction dataset generated by ESOC from the U.S. Army Corps of Engineers Gulf Region Division's Iraq Reconstruction Management System (IRMS) is based on construction projects with start dates between January 2003 and December 2008. We drop 122 projects with coordinates more than 60 miles outside Iraq's borders,

We cleaned the IRMS reconstruction data to remove information on specific project names and locations below the district level, creating a project-level dataset entitled maaws_20091002_07feb08districts_public.dta. The date refers to the date of the shape file defining the district boundaries used to geo-locate projects.

In addition to the project-level data, we provide a .do file which will aggregate the data up to any desired geo-temporal unit as well as producing summary measures of spending in different categories that include the categories used in Berman, Shapiro, and Felter (2011). The rules defining categories were developed based on extensive conversations with individuals involved in administering reconstruction spending from 2004-2009. Users of the data should be aware that aggregating below the half-year level temporal entails making decisions about the allocation of spending over time on projects, as most projects last several months. The spending allocation rule built into aggregate_maaws.do distributes spending evenly between start and end dates. That may or may not be an accurate representation and so users wishing to analyze the impact of spending at the month level, for example, should exercise due caution.

(1) *uri* – Unique project code used in IRMS.
(2) *Executing_Agency* – Agency responsible for executing the project on the ground.
(3) *Construction_Manager* – Agency responsible for construction.
(4) *Program* – Program project falls under. Some common acronyms are: CERP = Commander's Emergency Response Program; DFI = Development Funds for Iraq; ESF = Economic Stabilization Fund; IRRF = Iraq Reconstruction Funds; ISFF = Iraq Security Force Funds. Each program had distinct accounting rules and associated deadlines. More details on programs can be found in various program guidance documents.
(5) *Fund_Type* – Congressional appropriation which paid for the project. CERP appropriations, for example, were made each year from 2004 through 2009.
(6) *Program_Manager* – Military command responsible for the program. MCNI, for example, is Multi-National Corps Iraq.
(7) *Status, Project_Status* – Project status, second variable is acronym.
(8) *Benchmark_Category* – Standard categorization of projects made by GRD staff.
(9) *Benchmark_Item* – Standard items money was spent on as identified by GRD staff.
(10) *Sector, Sub_Sector, and Sub_Sub_Sector* – Variables categorizing projects by sector of the economy they supported.

(11) *Worktype* – Type of work supported by the project.
(12) *Construction_Cost* – Total construction cost in dollars.
(13) *Actual_Pct_Complete* – Percent of planned work completed.
(14) *Forecast_Award* – Planned date of award when project first entered into the system.
(15) *Actual_Award* – Date contract actually awarded.
(16) *Forecast_Start* – Planned start date when project first entered into system.
(17) *Actual_Start* – Date construction actually started.
(18) *Forecast_Finish* – Planned completion date when project first entered into system.
(19) *Actual_Finish* – Actual completion date. Missing for projects not completed.
(20) *District* – District of project location based on project lat-long as recorded in IRMS. Note, some project categories were sometimes recorded in the location of the entering officials. Many USAID Community Action Program (CAP) projects, for example, were recorded as being done in the Green Zone. Users should be careful to apply a common-sense check to variables that disaggregate spending by project type.
(21) *Governorate* – Governorate of project location based on project lat-long as recorded in IRMS.
(22) *soi* – Projects that appear to have been payments to local militias to provide security services, e.g. 'Sons of Iraq' units during the Anbar Awakening. The ability to detect such spending increase massively in July 2007 when SOI payments became an authorized spending category.

The .do file aggregate_maaws.do creates a range of variables in addition to *district*, *governorate*, *year*, and *month.* Each variable comes in threevarieties: *"spent_"* provides the amount spent on that type of project; *"np_"* provides the count of projects of that type active in that district/month; and variables with the *"-noncerp"* stub appended are versions of the variable dropping all CERP projects of that type.

In the aggregates below we drop projects with missing start date, with missing reconstruction expenditures, and with actual start date after actual finish date. We also dropped projects funded by CERP, CHRRP, and OHDACA which were completed in less than 30 days and cost more than $900,000 or were terminated due to contractor default. We calculated the duration of each project by counting the number of days to complete it. Projects started and finished on the same day were assigned duration of one day.

To calculate the amount spent in any month, we allocated the reconstruction spending of each project by dividing it evenly over its duration to get a daily total before aggregating all daily-project spending up to district/month reconstruction spending totals. In our experience results using aggregates below the half-year level can be sensitive to this aggregation rule as many projects span multiple months. The median CERP project, for example, lasted 93 days.

To aggregate up as desired users just need to change the values in the local variables *i* and *t* prior to the collapse at the bottom of the .do file with their geographic unit (*i*) and temporal unit (*t*).

For each of the following variables we calculate the amount spent and number of projects:

(1) *np* – Total number of projects of all types active in that region/period.
(2) *spent* – Total spent in that unit.
(3) *soi* – Projects that appear to be payments to local militias.
(4) *cerp* – This variable captures CERP projects.
(5) *conditional* – Projects funded by CERP, CHRRP, or OHDACA, all of which were military-administered programs that delivered small-scale projects. This is the variable used for CERP projects in "Can Hearts and Minds Be Bought?"
(6) *unconditional* – Projects not funded by CERP, CHRRP, or OHDACA.
(7) *cap* – Projects that were part of the USAID Community Action Program.
(8) *csp* – Projects that were part of the USAID Community Stabilization Program.
(9) *large* – Projects that cost more than $100,000 and spent in excess of $5,000/day (designated as "Large" projects).
(10) *democracy* – Projects relating to democracy promotion and governance.
(11) *education*- Projects relating to education.
(12) *electricity* - Projects relating to electricity.
(13) *healthcare* – Projects relating to healthcare or hospitals.
(14) *pubbuild* – Projects relating to public buildings, repair or construction.
(15) *transport* - Projects relating to transportation or roads.
(16) *watersan* - Projects relating to water and sanitation, potable water or sewerage.
(17) *dfi* – Projects funded through the Development Fund for Iraq.
(18) *irrf* – Projects funded through the Iraq Relief and Reconstruction Fund
(19) *cerp_nonsoi* – CERP projects that were not SOI-related.
(20) *cerp_large* – CERP projects over $50,000.
(21) *cerp_small* – CERP projects $50,000 or less in total cost.
(22) *recon* – Projects coded as reconstruction projects.
(23) *recon_c* – CERP projects coded as reconstruction.
(24) *notrec_c* – CERP projects coded as non-construction.

**Election Data**

esoc-iraq-v3_elections.dta

The election dataset records governorate-level returns in the December 2005 election, the smallest geographic unit for which returns were reported. Governorate level totals were created for different types of parties based on the classification below which divided parties into Sunni, Shia, Kurdish, secular nationalist, and pro-government parties. Small parties not receiving more than 1% of the vote in any province were not coded as there is no consistent information on their affiliations..

In addition to the variable Governorate, there are twelve variables in the dataset:

(1) *turnout* - The percentage of voter turnout.
(2) *su_v* - The percentage of valid votes for Sunni affiliated parties. These parties include Iraqi Accord Front, Iraqi National Dialogue Front, and Liberation and Reconciliation.
(3) *sh_v* - The percentage of valid votes for Shia affiliated parties (United Iraqi Alliance and Progressives).
(4) *sn_v* - The percentage of valid votes for secular nationalist parties (National Iraqi List, Iraqi Turkuman Front, Yazidi Party, Assyrian Democratic Movement List, Iraqi National Common Council, Gathering for Ind. Iraqis, and Parliament, National Forces).
(5) *k_v* - The percentage of valid votes for Kurdish affiliated parties (Kurdistani Alliance, Islamic Union of Kurdistan, and Islamic Movement in Kurdistan).
(6) *progov_v* - The percentage of valid votes for pro-government parties (UIA and Kurdistan Alliance).
(7) *valid_v* - The total number of valid votes.
(8) *invalid_v* - The total number of invalid votes.
(9) *blank_v* - The total number of blank votes.
(10) *total_v* - The total number of votes cast
(11) *sect* - A categorical variable that takes the value 1, 2, 3 or 4 reflecting the "sectarian" identity of the party. When at least 66% of the population in a governorate voted for clearly Sunni affiliated or secular nationalist parties, *sect*=1; when at least 66% of the population in a governorate voted for clearly Shia affiliated parties, *sect*=2; when at least 66% of the population in a governorate voted for clearly Kurdish affiliated parties, *sect*=3; otherwise, *sect*=4.

**Community Characteristics**

esoc-iraq-v3_ilcs-district.dta
esoc-iraq-v3_ilcs-governorate.dta
esoc-iraq-v3_ilcs-sect.dta

We provide aggregated data on community characteristics using raw 2004 ILCS survey responses at the district, governorate, and sect levels. The 2004 ILCS instrument is called the "Iraq Multiple Indicator Rapid Assessment" and is available online. Raw ILCS responses are available from the Central Organization for Statistics and Information Technology of Iraq (COSIT). Similar data could be constructed from the 2007 IHSES, though we have not done so.

The following variables are included in our replication data:

1. *elec_instability* – Percent of households reporting instability in their main electrical source.
2. *network_primary* – Percent of households relying on shared network as primary source of electricity.
3. *net_instability* – Mean level of instability on a 5-point ordinal scale.
4. *sewer_prob* – Mean level of problems with sewage system on 3-point scale.
5. *refuse_index* – Index based on series of interviewer observations on binary questions about presence of waste and sewage in vicinity of home.
6. *street_light* – Percent of households that had functioning streetlights outside their home when survey conducted.
7. *street_light_02* – Percent of households that report having functioning streetlights outside their home in 2002.
8. *street_light_change* – Difference between (6) and (7).
9. *pub_garbage* – Proportion of households relying on some form of public garbage collection, either collection service or coordination on common disposal area.
10. *police_d* – Mean self-reported distance to police station.
11. *pub_serv_d* – Index of distances reported to basket of public services.
12. *phone_index* – Index capturing series of questions about quality of phone service.
13. *road_qual* – Average quality of roads based on road type leading up to household.
14. *kid_safety* – Average satisfaction with child safety.
15. *clan_ties* – Proportion of households reporting members of their clan living in same neighborhood.
16. *jamiyya* – Proportion of respondents participating in rotating savings institution called *jamiyya*.
17. *pub_serv_in* – Index of respondent satisfaction with schools, health services, public transportation, and water supply.
18. *damage* – Proportion of households reporting damage to their dwelling.
19. *renters* – Proportion of households having renters.
20. *wealthin* – Index based on binary variables about possession of various household goods.
21. *inc_2002, inc_2003, inc_2004* – Mean reported income in year.

22. *inc_2002_qcap, inc_2003_qcap, inc_2004_qcap* – Mean reported income quintile in year.
23. *dif_02_03, dif_03_04, dif_02_04* – Mean change in household income between years.
24. *dif_02_03_qcap, dif_03_04_qcap, dif_02_04_qcap* – Mean change in household income quintiles
25. *missing* – Percent of households reporting a member missing.
26. *share_generator* – Percent of households using a shared generator.
27. *percent_grid* – Percent of households relying on the national electrical grid as main source of power.
28. *Victim* – Percent of households in which a household member was victim of a crime in the four weeks before the survey.
29. *relatives, coalition, police, community, militia, nobody* – Percent saying they would go to each of the above if they were the victim of a crime.
30. *hh_count* – Number of households in given area.

**Population Data**

esoc-iraq-v3_population.dta

We provide population figures for 2003, 2005, and 2007. These figures come mostly from World Food Program (WFP) estimates and are described below. Given the vast internal and external violence-caused population movements these figures should be taken as rough estimates only. Our population file also includes cross-sectional population estimates derived from LandScan (2008) data.

2003 Population Figures

2003 figures are mostly sourced from the WFP (World Food Program) 2004 report. Because of district reassignment, we grouped some WFP districts together. All figures have been aggregated to district-year for the purposes of our analysis.

The WFP 2004 report does not have population figures for districts in Dahuk and Erbil (except population of Koisnjaq district, which was reported in Sulaymaniyah governorate in WFP 2004), and the population figures were sourced from NCC (National Council of Churches) Iraq. Akre district is missing in Ninewa governorate in WFP 2004; we used as an estimate the population figure from NCC Iraq. WFP 2004 reports population for Al Faris district in Salah al-Din governorate, and we treat Al Faris district as a part of Balad district. We combined the population figure of Kifri district with the population figure of Kalar district in Sulaymaniyah governorate.

2005 Population Figures

2005 population figures are mostly from the WFP 2005 report. Because of district reassignment, we grouped some WFP districts together. These WFP districts are Rawa (combined with Ana) in Anbar, Al Faris (combined with Balad) in Salah al-Din, Kifri (combined with Kalar) in Sulaymaniyah, Shahrazour (combined with Halabja) in Sulaymaniyah, and Al Azizia (combined with Al Suwaira) in Wassit. We also assigned WFP's Koisnjaq district in Ninewa governorate as a district in Erbil governorate. The population figures of the following districts are from NCC Iraq: Amedi, Dahuk, Sumel, and Zakho in Dahuk; Choman, Erbil, Makhmur, Mergasur, Shaqlawa, and Soran in Erbil; and Akre in Ninewa.

2007 Population Figures

2007 population figures are from the WFP 2007 report. As in our treatment of WFP 2004 and WFP 2005, we grouped some districts in WFP 2007 together and reassigned the governorates of some districts in WFP 2007. We grouped WFP's Rawa district with Ana district in Anbar governorate. WFP's Al Shekhan district in Dahuk governorate was grouped with Al Shikhan district in Ninewa governorate. WFP's Bardah Resh district and Akre district in Dahuk governorate were grouped together and assigned to Ninewa governorate as Akre district. Koisnjaq district in Erbil governorate includes WFP's Kardagh district and Said Sadik district in Sulaymaniyah governorate. WFP's Erbil district, Khabat district, and Dusty Howleer district in Erbil governorate were grouped into Erbil district in Erbil governorate. WFP's Al Digeel district was grouped with Balad

district in Salah al-Din governorate. WFP's Al Azizia district was grouped with Al Suwaira district in Wassit governorate.

LandScan Population Figures

We calculated district-level population figures using the LandScan (2008) gridded population data which provide population estimates at the 30 arc second level, approximately 1km by 1km at the equator. LandScan estimates are generated by combining available census data with information on land cover, elevation, slope, and other variables derived from overhead imagery. See http://www.ornl.gov/sci/landscan/landscan_documentation.shtml for more details. Using Hawth's Tools we aggregated these data to the district level.

**Household Income Variables and Unemployment Rates**

esoc-iraq-v3_econfactors.dta

Economic variables are captured from the WFP surveys and ILCS survey and are described below.

2004 district level household income variables and unemployment rates were calculated using the ILCS. 2005 and 2007 district level household income variables and unemployment rates were calculated using the WFP 2005 report and the WFP 2007 report. Note that the WFP unemployment rate is calculated as % (of household members) unemployed divided by the sum of % employee, % employer, % working on own account, % working pensioner, and % unemployed. Because certain WFP's districts were grouped together, each combined district's statistic is the population weighted average of the component districts' statistics.

The unemployment rate is the average fraction of unemployed household members divided by the average fraction of household members in the labor force. The employment rate is simply the fraction of the population employed divided by the total population.

The household income variables are the fractions of a district's households in each national income quintile. Thus, hhinc_q1 is the percent of district households with incomes in the top fifth of the national income distribution. From these, we create t2_prop and b2_prop as the percent of district households with income levels in the top or bottom 40% of the nation respectively. For variable values in this dataset for years 2008 and 2009, we linearly extrapolate from the 2003-2007 values using Stata's -ipolate- command.

**Natural Resource Data**

esoc-iraq-v3_oil.dta

Our natural resource data come from a variety of sources providing shape files on natural resource reserves and pipelines. From these data, we used GIS software to calculate measures at the district and governorate level. The following variables (all scaled to billions of barrels per unit) have been used in tables III, IV and V of "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq," *Journal of Political Economy* 119: 766-819.

*intrp_resv_p* – refers to total oil and gas reserves, weighted by price. Total reserves were calculated using hydrocarbon field size data from Horn (1999) as well as private data. For fields whose total reserves were not precisely known, the maximum and minimum values of the estimated range of field reserves were averaged to produce the reserves value assigned to the field. To identify reserves for fields lacking data, we used a linear prediction based on field size with the prediction based on a regression of reserve volume on field surface area for well-studied fields ($R^2 = 0.7012$). Total oil and gas reserves of individual fields lying across district boundaries were divided according to the proportion of their surface areas within each district. This approach is justified as the extraction technologies currently being used in Iraq do not permit substantial horizontal drilling; we can therefore assume that all wells located in a given district tap only resources within the district). Finally, total reserves per district were calculated by summing the total reserves of oil and gas fields (both whole and partial) lying within district boundaries.

*interp_resv_p_infadj* – calculated as above, but adjusted for inflation in the US dollar.

*oil_vol_p* – refers to total pipeline volume, aggregated by district-month, weighted by the price of oil.

*oil_vol_p_infadj* - refers to total pipeline volume, aggregated by district-month, weighted by price of oil, adjusted for inflation.

We provide a number of other variables in the data file which may be useful to analysts but which we feel do not represent the natural resource value of an area as well as the variables above for a number of reasons.

**Civilian Casualty Data**

esoc-iraq-v3_ibc.dta

Data on civilian casualties in Iraq are fully described in Luke N. Condra and Jacob N. Shapiro, "Who Takes the Blame? The Strategic Effects of Collateral Damage," *American Journal of Political Science* 56, no. 1 (January, 2012). Users of these data should cite this article in which the data are first presented. Code to aggregate as in that paper are available upon request.

The data are the product of a multi-year collaboration with Iraq Body Count (IBC), a non-profit organization that collects data on civilian casualties suffered in the Iraq War and makes them publicly available (http://www.iraqbodycount.org). The data provided here improve on the publicly available data in a variety of ways, most importantly in the specificity of the incidents' geocoding down to the district-level in all but about 12% of included incidents (2,612 of 21,086). The data are based on media reports of incidents involving civilian casualties between December 2003 and July 2009.

Civilian casualties are usefully divided into four different categories in the data: (1) Insurgent killings of civilians that occur in the course of attacking Coalition or Iraqi government targets; (2) Coalition killings of civilians; (3) Sectarian killings defined as those conducted by an organization representing an ethnic group and which did not occur in the context of attacks on Coalition or Iraqi forces; and (4) Unknown killings, where a clear perpetrator could not be identified. This last category captures much of the violence associated with ethnic cleansing, reprisal killings, and the like, where claims of responsibility were rarely made and bodies were often simply dropped by the side of the road.

Users should think carefully about attribution rules as many events involve multiple perpetrators. There is no way, for example, to distinguish which casualties incident to a firefight were due to insurgent action and which due to Coalition action.

Variables included in the data:

- Code – IBC incident code
- *Min* – minimum casualties attributed to that incident
- *Max* – maximum casualties attributed to that incident
- *Location* – location details of the incident taken from the media source.
- *Weapons* – description of weapons used during the incident.
- *Town* – town in which incident occurred, if known.
- *Cause_of_Death* – Formal categorization of cause of death with party responsible in [].
- *Morgue_Cumulative* – Indicates event included in the data because one confirming source was an aggregated morgue report.
- *Princeton_Disaggregated* – indicates incident that was disaggregated from original IBC coding as part of our coding effort. This would include morgue

reports that specify numbers by district, which would not have been disaggregated as part of the original IBC data.

- *Princeton_Supplement* – indicates incident added to original IBC coding as part of our coding effort.
- *governorate* – incident governorate.
- *district* – incident district.
- *alt_governorate* – alternate governorate for that incident, used when there is ambiguity in the press reports.
- *alt_district* – alternate district for that incident, used when there is ambiguity in the press reports.
- *coalition* – Dummy variable equal to '1' if incident involves casualties caused by Coalition forces.
- *insurgent* – Dummy variable equal to '1' if incident involves casualties caused by insurgents involved in combat with Coalition forces.
- *sectarian* – Dummy variable equal to '1' if incident involves casualties caused by identified sectarian militia not involved in combat with Coalition forces. This category thus combines violence by sectarian militias targeted at co-ethnics as well as violence against people from other groups.
- *unknown* – Dummy variable equal to '1' if incident involves casualties caused by unknown perpetrator not involved in combat with Coalition forces.
- *sect_type1* – Dummy variable takes '1' if sectarian casualties involve direct fire.
- *sect_type2* -- Dummy variable takes '1' if sectarian casualties involve indirect fire.
- *sect_type3* -- Dummy variable takes '1' if sectarian casualties involve bombs and explosions.
- *sect_type4* -- Dummy variable takes '1' if sectarian casualties involve selective violence.
- *sect_type4* -- Dummy variable takes '1' for sectarian violence not fitting those categories.
- *event* – Marker to enable counting of incidents
- *reasons* – reports why an incident could not be geo-located to the district. Codes are: 99 - No location information available in source; 88 - Distinct incident, no sub-governorate location information available in source; 77 - Distinct incident, sub-governorate location information indeterminate; 66 - Aggregated casualty report, no sub-governorate location information available in source; 55 - Aggregated casualty report, sub-governorate location information indeterminate.
- *ph* – Dummy variable to indicate incident with potentially higher body counts than reported. Typically means that an incident involved a number of people who could not be positively identified as civilians.

## Ethnicity Data[4]

esoc-iraq-v3_ethnicity.dta

We generated data on ethnic populations in Iraq by combining high-resolution population data (LandScan 2008)[5] and maps that bound areas in which ethnic groups reside to estimate ethnic group population numbers at the district administrative level.[6]

### Base Files

The boundaries of Iraq's districts and governorates we use are designed to conform as closely as possible to the boundaries recognized and observed in the household surveys administered by the United Nations World Food Programme. Our shapefiles were used in Berman, Shapiro, Felter (2011)[7] with very minor improvements to the alignment of district boundaries.[8]

### Map Sources

Many maps of ethnic and religious groups in Iraq are available, but most are derived from only a few original sources. We include an appendix that explains and lists all maps considered for this project. Out of the maps researched, we determined that six maps contain distinct and useful information and have traced these maps for use in GIS. The first five maps are country maps, while the last contains useful information on the urban area of Baghdad.

| Map source: | Map filename[9]: |
|---|---|
| CIA 1978 | iraq_ethnic_1978 |
| CIA 2003 | iraq_ethno_2003 |
| CIA 1992 | iraq_ethnoreligious_1992 |
| M. Izady | Iraq_Ethnic_lg.jpg |
| M. Izady | Central_Iraq_Ethnic_lg.jpg |
| M. Izady | Baghdad_Ethnic_2003_lg |

### Coding

We traced the maps to create a set of boundaries that delineate the areas where ethnic groups live. Areas identified as belonging to a single ethnic group were treated as homogeneous, mixed areas were treated as evenly split, and minor ethnic groups were not counted. The maps we use only mix two major groups, never three or more.

---

[4] The first version of these data were developed and described by Joshua Borkowski and Zeynep Bulutgil. Subsequent improvements were made by Zach Romanow and Christopher Paik.

[5] Oak Ridge National Laboratory, Oak Ridge, TN. LandScan™ Global Population Database. Available at https://share.ornl.gov/sites/landscan/2008/default.aspx. This information is gridded at a scale of approximately 1 km$^2$ (one cell is 30 arc seconds).

[6] For another application of this approach see, NB Weidmann, JK Rød, LE Cederman, "Representing Ethnic Groups in Space: A new Dataset," *Journal of Peace Research* 47, no. 4 (July 2010): 491-99.

[7] E Berman, JN Shapiro, J Felter, "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq," *Journal of Political Economy* 119, no. 4 (August 2011): 766-819.

[8] Iraq's governorates and districts are first- and second-level administrative divisions, respectively. Out most recent file, dated 03-17-2010, is included locally in this file's replication folder: \ESOC_Iraq_boundaries\Iraq_district_boundaries_UTM.shp

[9] Please see Appendix I for full citations.

Using the ArcGIS Intersect tool we created a new shapefile of district fragments, representing the areas of each district that are inhabited by one major or a mixture of ethnoreligious groups. Each fragment area is defined by (1) which district it forms a part, and (2) its ethnic mix. For each population fragment we used LandScan (2008) data to calculate its population. These were then aggregated by ethnicity to provide district totals by ethnic group. Note, these totals by district do not exactly match the totals calculated by district area alone, mostly due to the zero-coding of small ethnic groups.

Baghdad

Because of its complexity and density, the city of Baghdad was treated separately from the other three maps. A boundary for the city's urban area was traced from the Humanitarian Info Centre for Iraq. This boundary includes all of Al Resafa district, and parts of Adhamiya, Al Sadr, Karkh, and Khadamiya districts. The boundary shapefile was broken into 21 neighborhood-level polygons that were individually coded. As before, these areas were coded with a proportion of the ethnic groups who reside within each one.

Replication

Full replication files including source maps and shape files are available from the ESOC Iraq Ethnicity files.

List of Variables:

The final tabular file includes our counts by district and by ethnic group.

| | |
|---|---|
| district | District name |
| Sunni_pop_CIA_1978 | Sunni Arab population by district, according to the 1978 CIA map boundaries. |
| Kurd_pop_CIA_1978 | Kurdish population by district, according to the 1978 CIA map boundaries. |
| Shia_pop_CIA_1978 | Shia Arab population by district, according to the 1978 CIA map boundaries. |
| Total_pop_CIA_1978 | Sum of Sunni Arab, Shia Arab, and Kurdish population by district, according to the 1978 CIA map boundaries. |
| Sunni_pop_CIA_2003 | Sunni Arab population by district, according to the 2003 CIA map boundaries. |
| Kurd_pop_CIA_2003 | Kurdish population by district, according to the 2003 CIA map boundaries. |
| Shia_pop_CIA_2003 | Shia Arab population by district, according to the 2003 CIA map boundaries. |
| Total_pop_CIA_2003 | Sum of Sunni Arab, Shia Arab, and Kurdish population by district, according to the 2003 CIA map boundaries. |
| Sunni_pop_CIA_1992 | Sunni Arab population by district, according to the 1992 CIA map boundaries. |
| Kurd_pop_CIA_1992 | Kurdish population by district, according to the 1992 CIA map boundaries. |
| Shia_pop_CIA_1992 | Shia Arab population by district, according to the 1992 CIA map boundaries. |

| | |
|---|---|
| Total_pop_CIA_1992 | Sum of Sunni Arab, Shia Arab, and Kurdish population by district, according to the 1992 CIA map boundaries. |

landscan_pop – district population (WFP)
shiapop – district's Shia population based on Landscan data
sunnipop – district's Sunni population based on Landscan data
kurdpop – district's Kurdish population based on Landscan data
xtianpop – unit's Christian population based on Landscan data
turcpop – unit's Turkoman population based on Landscan data
mixedpop – district's mixed population based on Landscan data