Joint Special Operations University
Department of Strategic Studies

Big Data Conflict Forecasting:
Operationalizing the Data Science Team

By Diane M. Zorri and Mihhail Berezovski

JSOU PRESS
OCCASIONAL PAPER
JUNE 2021

On the cover: Public utilities and power grids are often compromised, attacked, or controlled by malign actors. Photo by SHUTTERSTOCK.

The views expressed in this publication are entirely those of the author and do not necessarily reflect the views, policy, or position of the United States Government, Department of Defense, United States Special Operations Command, or the Joint Special Operations University.

This work was cleared for public release; distribution is unlimited.

JSOU Press Publications are available for download at: http://jsou.libguides.com/jsoupublications.

ISSN 2572-9020 (print)
ISSN 2572-9039 (online)

About the Authors

Dr. Diane Maye Zorri is an assistant professor of security studies at Embry-Riddle Aeronautical University in Daytona Beach, Florida and a non-resident senior fellow with Joint Special Operations University. Prior to Embry-Riddle, Dr. Zorri served as a visiting professor at John Cabot University in Rome, Italy and as an affiliated scholar with George Mason's School for Conflict Analysis and Resolution. Prior to her work in academia, she served as an officer in the United States Air Force and later worked in the defense industry doing foreign military sales, integrated communications, and proposal development for an Italian defense conglomerate. She is a graduate of the U.S. Air Force Academy and Naval Postgraduate School and earned a Ph.D. in political science from George Mason University's Schar School of Policy and Government.

Dr. Mihhail Berezovski is an associate professor of mathematical sciences and program coordinator for the data science track within the bachelor of science in computational mathematics program at Embry-Riddle Aeronautical University. Prior to Embry-Riddle, Dr. Berezovski was a postdoctoral scholar in the Department of Mathematical Sciences at Worcester Polytechnic Institute in Worcester, Massachusetts. His research centers on two disciplines: computational mathematics for material science and data-enabled industrial mathematics. His research in computational mathematics for material science focuses on adaptive algorithms for numerical simulations of dynamic energy redistribution in advance and novel materials. His latest research area of interest is mathematical and statistical modelling, and numerical methods and algorithms for data-enabled problems in social studies and real-world industrial problems. He is a graduate of the Tallinn University of Technology where he spent several years in the Institute of Cybernetics.

Big Data Conflict Forecasting: Operationalizing the Data Science Team

United States Special Operations Command (USSOCOM) continuously identifies emerging threats to U.S. Special Operations Forces (SOF). The nature of the global system has prompted a focus on how the "confluence of information, technology, and innovation" will affect the force now and in the future.[1] As technologies mature, the military is more likely to use enhanced data techniques to consolidate and analyze information, leverage geospatial capabilities, and use the material to provide alternative planning constructs. Ideally, data analytics and modeling can (a) give special operators the ability to conduct more rigorous scenario-based planning on a country given varying initial conditions and (b) use those scenarios for alternative planning constructs in order to pre-position appropriate resources given variations in a country's resource and governance capabilities.

At the heart of data analytics—commonly referred to as big data—is the data science team, which is typically comprised of subject matter experts (SME) on the topic under investigation, and data modelers, whose expertise lies in mathematics and computer programming. The data science team is a prerequisite for using big data effectively because no dashboard or single algorithm can accurately work across locations. Each situation has a unique context that requires subject matter expertise to decipher, and access to accurate data is not always possible, meaning modelers have to use professional judgment on how to employ the appropriate statistical methods given the data in hand. Identifying emerging threats using advanced technology requires a group of SMEs and technical experts to brainstorm on what to model, how to model it, and what decision-making insights can be derived from the modeling results.[2]

To demonstrate how data-enabled intelligence and planning could be employed through a data science team, this occasional paper explores the practicality of using big data analytical techniques on open-source, publicly available, archived datasets. SOF are oftentimes deployed to locations with intensifying conflict and with limited knowledge of prior political history or conflict dynamics. While archival datasets often appear unhelpful, they could in fact identify local conflict patterns with operational-level consequences.

In this case study, the authors demonstrate how a data science team can be operationalized by combining SME appreciation of a conflict zone with a mathematics and computer programming

[1] *Special Operations Research Topics 2020*: *Revised Edition for Academic Year 2021* (Tampa: JSOU Press, 2020), https://jsou.libguides.com/ld.php?content_id=55347911.

[2] For more on how to compose a data science team, see Joan Peckham and Andrew Geyer, "Making Big Data Models Work Right," in *Big Data for Generals…and Everyone Else over 40*, ed. David C. Ellis (Tampa: JSOU Press, forthcoming).

expert. Based on prior operational experience (2008–2009) and subsequent research expertise on Iraq, the content SME recognized the doctrinal importance of the stabilizing phases of armed conflict on the counter-Islamic State of Iraq and Syria (ISIS) campaign from 2014–2017. Theoretically, stability operations should reduce the probability of continuing armed conflict by restoring vital essential services and infrastructure to meet the basic survival needs of the people and, in so doing, restoring support for the government. Lacking the technical expertise in big data analytics, the content SME then engaged a professor of mathematics and computer programming who could utilize technology to combine and clean disparate, sometimes incomplete, archival data.

This project provides a step-by-step demonstration of how to operationalize big data on a conflict zone about which SOF might have limited direct experience. Certainly, SOF have extensive knowledge and experience with Iraq, and this conflict zone was chosen in part for its familiarity so the focus could remain on the data analytics approach instead of the conflict zone itself. First, the authors illustrate how a doctrinal assumption or theory of conflict can frame a research hypothesis. Second, they provide a brief review of the Iraq conflict dynamics to which the doctrinal assumptions could be applied to generate a context-relevant model. Third, they discuss how the data science team cleans the data by integrating several archived datasets, determines the research methodology, and performs multi-step data analysis based on the hypotheses derived from steps one and two. Fourth and finally, they show how to analyze the data and draw insight from the data science approach.

Step 1: Determine a Research Hypothesis—U.S. Military Doctrinal Assumptions on Essential Services and Conflict

Basic needs, such as shelter and environmental protection, are essential for human survival.[3] Renowned psychologist, Abraham Maslow, conceptualized these needs in a hierarchy, where safety and physiological needs are at the foundation of human motivation. According to Maslow, physiological needs—such as breathing, food, water, and sleep—must be met before humans can focus on the needs of safety and security as well as higher-order physiological needs like love, esteem, or intimate relationships.[4] Figure 1 provides a well-known pictograph of Maslow's Hierarchy of Needs model coupled with the hierarchy and foundation for how the U.S. military prioritizes essential services during the stabilization phases of armed conflict.

Physiological needs form the foundation of Maslow's Hierarchy of Needs model. Likewise, the U.S. military has identified several components of essential services. Amongst these are water, food, security, medicine, housing, electricity, trash, and sewer services. Hot summer days in countries like Iraq, where temperatures can regularly climb above 105°F, make public utilities, especially electricity, an important basic survival need for humans.

[3] Abraham Maslow, *Motivation and Personality*, 3rd ed. (New York: Addison Wesley Educational Publishers, 1987), xxiii.

[4] Maslow, *Motivation and Personality*, 45.

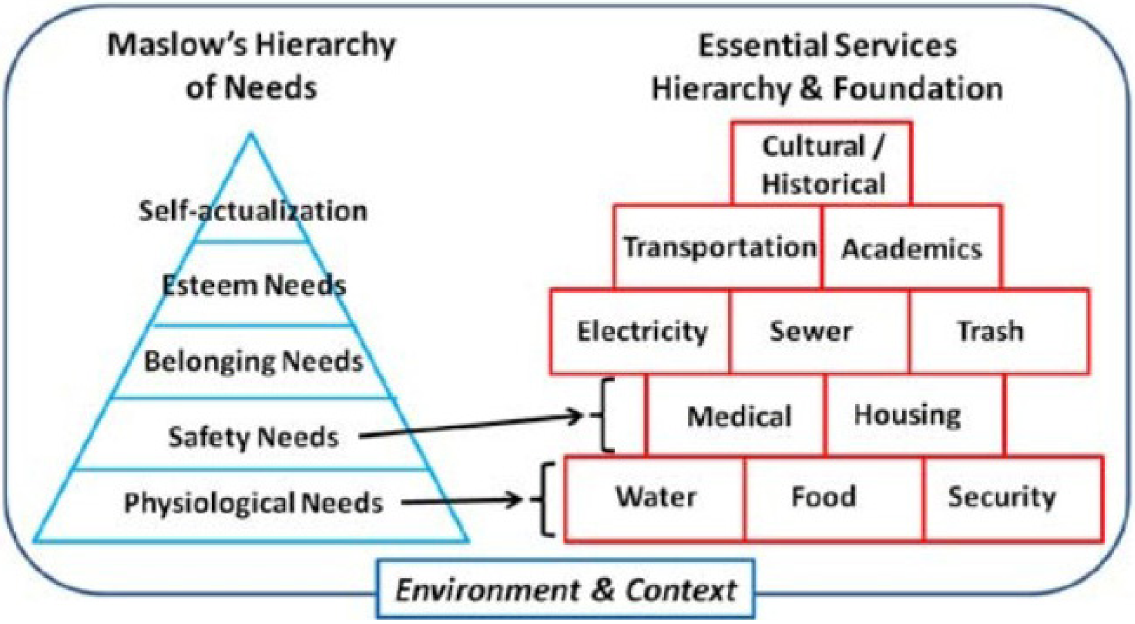

Figure 1. Maslow's *Hierarchy of Needs* model and the U.S. Army Corps of Engineers essential services components. Source: Anthony Barbina/Comparing models for the restoration of essential services during counterinsurgency operations.

During periods of violent conflict, public utilities and power grids are often compromised, attacked, or controlled by maligned actors.[5] In the stability phases after violent conflict, restoring essential services such as public utilities not only increases government legitimacy and restores a sense of normalcy, in theory it will decrease the population's inclination towards criminal activity.[6] When a government, regime, or commanding force does not restore essential services or uses essential services as a tool against the people, populations may resist the regime in power. For this reason, the restoration of essential services to the population's expectation is a critical component of the U.S. counterinsurgency and counterterrorism strategies, which advocate population-centric tactics, the use of small maneuver units, and engaging the civilian population by leaving forward-operating bases and dispersing forces throughout urban centers and villages. Likewise, the U.S. Army Corps of Engineers' Sewer, Water, Electricity, Academics, and Trash model was widely used in Iraq and Afghanistan during recent operations as a guideline for implementing stability and restoring essential services.[7]

[5] James Glanz, "The Reach of War: Infrastructure; Sabotage Cuts Power to More Than 100 Key Electrical Lines," *New York Times*, 11 June 2004, https://www.nytimes.com/2004/06/11/world/reach-war-infrastructure-sabotage-cuts-power-more-than-100-key-electrical-lines.html

[6] Anthony Barbina, *Comparing Models for the Restoration of Essential Services during Counterinsurgency Operations* (Fort Leavenworth: School of Advanced Military Studies, United States Army Command and General Staff College, 2011).

[7] Department of the Army, *Counterinsurgency,* FM 3-24 (Fort Leavenworth: Department of the Army, 2006): 8-4; Department of the Army, *Stability Operations,* FM 3-07 (Washington, D.C.: Department of the Army, 2008): 3-9.

While the full nature of the relationship and interplay between environmental factors, essential services, the government, and populations under duress is indeterminate, researchers across several scientific disciplines have found the onset of violent conflict is more likely in warmer years than in colder.[8] Despite this finding, the complex variables of governance and basic utility needs are often overlooked when addressing the human element of violent conflict. For instance, years after the initial coalition invasion of Iraq, electricity remained scarce, and many reconstruction efforts failed to meet the basic needs of the Iraqi people.[9] U.S. military and coalition planners closely monitored restoration efforts, but even into the fifth and sixth years after the invasion, some sectors of the country only received four hours of power (HOP) during the country's hottest months.[10] Likewise, violent conflict and terrorist incidents ebbed and flowed, but each year they peaked during the hottest months.[11] Electricity was a highly sought commodity, yet sectarian infighting and stretched budgets stymied the Iraqi government's efforts to modernize and upgrade its energy sector.[12] Survey data revealed the Iraqi public was largely dissatisfied with the availability of basic needs such as water, power, and sanitation,[13] and by 2006, components of the U.S. military recognized the issue and deemed electricity to be the highest infrastructure priority.[14] Survey data also revealed that during this period Iraqis actually understood the term reconstruction "in the narrow sense of getting the power restarted."[15]

Scientists have assessed a broad range of weather and climate data and their effects on human behavior. For instance, researchers have found that long periods of sustained stable, warm weather are statistically associated with the outbreak of civil war.[16] Another study looked at average monthly temperatures when correlating the outbreak of civil war;[17] and another looked at planetary-scale climate patterns and the stability of modern societies.[18] Similarly, a number of academics have noted the positive correlation between temperature and the outbreak of conflict or climate change and

[8] Solomon Hsiang, Carl Meng, and Mark Cane, "Civil Conflicts are Associated with the Global Climate," *Nature* 476 (2011): 438-441.

[9] Barbina, *Comparing Models.*

[10] Ibrahim Karim, Power Generation Database, Iraqi Transition Assistance Office (unpublished data file, 2009).

[11] Global Terrorism Database, National Consortium for the Study of Terrorism and Responses to Terrorism (2009), https://www.start.umd.edu/gtd.

[12] Luay al-Khatteeb et al., *Turn a Light On: Electricity Sector Reform in Iraq* (Doha: Brookings, 2016), https://www.brookings.edu/wp-content/uploads/2016/06/alkhatteeb-Istepanian-English-PDF.pdf.

[13] Frederick Barton et al., *Progress or Peril? Measuring Iraq's Reconstruction* (Washington, D.C.: Center for Strategic and International Studies, 2004), https://www.csis.org/analysis/progress-or-peril-measuring-iraqs-reconstruction-0.

[14] Barbina, *Comparing Models.*

[15] Barton et al., *Progress or Peril?*

[16] Steven Landis, "Temperature Seasonality and Violent Conflict: The Inconsistencies of a Warming Planet," *Journal of Peace Research* 51, no. 5 (2014): 603-618.

[17] Landis, "Temperature Seasonality," 603-618.

[18] Hsiang, Meng, and Cane, "Civil Conflicts," 438-441.

conflict.[19] Researchers and policy makers have also made note of how inadequate essential services can act as a motive for conflict and civil insurrection.[20]

Step 2: Establish the Local Context—Post-Saddam Iraq

In March 2003, when Saddam Hussein's Ba'athist regime fell, so did the facade of public order. Without the regime in place to enforce public security, the Iraqi people took to the streets, looting and burning the city of Baghdad. Initially, coalition forces were not ordered to contain the violence. This proved to be a critical error in decision-making. The impact of the lawlessness compounded the already dilapidated state of Iraq's infrastructure, making it far more difficult for the transitional governments to provide essential services such as electricity, sewage, and water. Likewise, the looting made it difficult for the coalition to operate under the plans they had devised for the post-war occupation and recovery. The subsequent de-Ba'athification of the internal security police and military forces exacerbated the problems. The result of these decisions was widespread criminal activity and civil chaos.

The abolition of the Iraqi government, military, and security forces by the coalition forces created a power vacuum that was quickly filled by Shi'a militias, Sunni insurgents, former Ba'ath party loyalists, and al-Qaeda in Iraq operatives. Retaliatory killings, torture, and kidnappings greatly increased across the country, and the rift between all armed factions proliferated. Iraq's Shi'a political factions were split with regard to their support for U.S. political objectives. The political blocs with institutional longevity—like the Da'wa party and the Islamic Supreme Council of Iraq—generally worked with the Americans and coalition partners.

Populist Shi'a groups like Jaysh al-Mahdi (JAM), who supported the clerical teachings of Moqtada al-Sadr, put up a resistance front to U.S. and coalition efforts as well as the Shi'a-led government in Baghdad.

As such, by 2006 American policy makers feared that if U.S. forces pulled out of the region too soon, the nascent Iraqi government would be faced with overwhelming opposition from the insurgency and an ethno-sectarian civil war.[21] In order to counter al-Qaeda in Iraq and the most radicalized elements of Iraqi society, U.S. and coalition forces stayed in the country much longer

[19] See: Ole Theisen, Nils Gleditsch, and Halvard Buhaug, "Is Climate Change a Driver of Armed Conflict?" *Climate Change* 117 (2013): 613–625; Craig Anderson, "Heat and Violence," *Current Directions Psychological Science* 10, no. 1 (2001): 33-38; Drago Bergholt and Paivi Lujala, "Climate-related Natural Disasters, Economic Growth, and Armed Civil Conflict," *Journal of Peace Resolution* 49, no. 1 (2012): 147–162; Alexander Bollfrass and Andrew Shaver, "The Effects of Temperature on Political Violence: Global Evidence at the Subnational Level," *PLoS ONE* 10, no. 5 (2015); Halvard Buhaug, "Climate Not to Blame for African Civil Wars," *Proceedings of the National Academy of Sciences* 107, no. 38 (2010): 16,477–16,482; Cullen Hendrix and Steven Glaser, "Trends and Triggers: Climate, Climate Change and Civil Conflict in Sub-Saharan Africa," *Political Geography* 26, no. 6 (2007): 695–715; Solomon Hsiang, Marshall Burke, and Edward Miguel, "Quantifying the Influence of Climate Change on Human Conflict," *Science* 341, no. 6151 (2013): 1–14.

[20] Department of the Army, *Counterinsurgency*; Barbina, *Comparing Models.*

[21] Marina Ottaway, *Back from the Brink: A Strategy for Iraq* (Washington, D.C.: Carnegie Endowment for International Peace, 2005), https://carnegieendowment.org/2005/11/28/back-from-brink-strategy-for-iraq-pub-17724.

than initially expected. The conundrum for Iraq was that in order for the nation to gain security, Iraqi-led forces had to generate enough control of their own territory, but without U.S. and coalition forces in the region, the nation would have quickly spiraled into civil chaos. Planned in 2006 and deployed in 2007, the U.S. surged military forces to the region to counter the massive resistance movement by Iraqi armed factions like JAM and al-Qaeda in Iraq.

Yet equally important to the surge in coalition forces was the coalition and Iraqi government's ability to restore essential services to the people. One Army colonel present in Iraq during the surge recounted that the reconstruction projects were just as important in helping the Iraqis regain a sense of normalcy.[22] The surge in forces, as well as the military's doctrinal changes, eventually led to a decrease in violence and a strategic pause, which enabled U.S. policy makers to negotiate—albeit not actually agree to—a Status of Forces Agreement with the Iraqi government and develop a plan for complete withdrawal of U.S. forces.[23] However, it is impossible to assess the success of the surge without also looking at the simultaneous reconstruction efforts by the U.S. Army's Civil Affairs units, the U.S. Army's Corps of Engineers, and the State Department's Provincial Reconstruction Teams. Billions of U.S. taxpayer dollars were spent on reconstruction projects, which went directly toward civil capacity building. The importance of capacity-building tasks is clearly outlined in Joint Publication 3-07, *Stability*:

> Detailed planning enables staffs to integrate and synchronize activities in time and space, identify complementary and reinforcing actions, and prioritize efforts within and across the stability sectors. The stability sectors represent the five key focus areas for civil-military efforts. The stability sectors are security, justice and reconciliation, humanitarian assistance and social well-being, governance and participation, and economic stabilization and infrastructure.[24]

Yet, building civil capacity was not always within the jurisdiction or control of coalition forces. Both the Iraqi government and insurgent forces sought to control and sabotage coalition objectives, including efforts to build electrical capacity. Insurgents cut power supplies into major Iraqi cities, and the central government routinely cut the power supply on a sectarian basis.[25] As such, by 2007 the Iraq energy sector was only meeting about half of its demand on electricity.[26]

A side benefit of having U.S. and coalition forces located across Iraq was that reasonably good data was captured on the availability of public services. In 2014 and 2015, Iraq once again faced an onslaught of violence and civil strife as militants from the Salafist terrorist organization, Islamic State, rapidly gained territory across the country. Like 2007 and 2008, in 2014 and 2015 violence

[22] Retired Lieutenant Colonel Dale Kuehl, interviewed by Diane M. Zorri, May 2014, transcript.

[23] Thomas Ricks, *The Gamble: General Petraeus and the American Adventure in Iraq* (New York: Penguin, 2008), 122.

[24] Joint Chiefs of Staff, *Stability,* 3-07 (2016), https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_07.pdf.

[25] James Glanz, "High Voltage Sabotage Cuts Power to Baghdad," *New York Times,* 19 December 2006, https://www.nytimes.com/2006/12/19/world/africa/19iht-electric.3951190.html.

[26] "Electricity in Iraq Factsheet," Inter-Agency Information and Analysis Unit, 2010, https://reliefweb.int/sites/reliefweb.int/files/resources/F409BC15DE5570AC8525777C006B1430-Full_Report.pdf.

increased during Iraq's arid summer months.[27] Moreover, years of chaotic politicking had "stymied the Iraqi government's efforts to modernize and upgrade their energy sector."[28] The question here is: Could intelligence and planning personnel have forecasted the levels of conflict in 2014 and 2015 by modeling the data from 2007 and 2008? Doing so would have improved decision-making on how to posture U.S. and coalition support across the interagency for a maximum counter-ISIS effect in Iraq.

Step 3: Establish the Modeling Framework

Based on the extant research and experience with the Iraqi context, the authors hypothesized there could be a relationship between essential services, temperature, and the probability of conflict. To better understand the interplay between these widely disparate factors, and understanding that—in theory—a human security perspective should deeply inform counterinsurgency and counterterrorism approaches, this research project integrated several layers of information about unmet demand on electrical power, air temperature, and the outbreak of conflict in Iraq. Although this project is specifically focused on Iraqi conflict dynamics—in theory—frameworks such as this could help researchers and military planners better decipher the spatial and temporal patterns of human behavior under a state of duress. In addition, using environmentally based geographic data in conjunction with human demographics, governance, and conflict information could contribute towards a better understanding of emerging security issues for SOF in rapidly changing environments.

Yet equally important to the surge in coalition forces was the coalition and Iraqi government's ability to restore essential services to the people.

This study uses three seemingly disparate, obsolete, or archived datasets to assess the relationship between air temperature, the availability of essential services, and instances of conflict and violent activity. As an example of how to use the archival data for an operational purpose, the authors assess the demand on public utilities in Iraq and the daily temperature across the country, and then compare these instances to variations in conflict within an active war zone.[29]

Step 3a: Clean and Reconcile Disparate Data

It is important to note the many limitations when reconstructing patterns of human behavior from disparate data sources, especially when using data that was not collected expressly for the purpose of the new modeling project. Preexisting datasets often exist in a variety of formats and languages and contain diverse dimensionality. Because of this, the data is often very low quality. When looking at disparate data, it is common for there to be "missing values, inconsistencies, ambiguous records, noise, and high data redundancy."[30] Therefore, it is a challenge to integrate and analyze

[27] Global Terrorism Database.

[28] al-Khatteeb et al., *Turn a Light On*.

[29] "Iraq Daily Temperature," National Oceanic and Atmospheric Administration, National Center for Environmental Information, 2019.

[30] Lidong Wang and Randy Jones, "Big Data Analytics for Disparate Data," *American Journal of Intelligent Systems* 7, no. 2 (2017): 39-46.

disparate data from various sources. Unfortunately, wartime metrics are often disparate, dynamic, untrustworthy, and inter-related, which makes the cleaning and reconciling step a prerequisite. As technology and computing power improves, more complex data analytics can be used to analyze correlation between factors and detect patterns or uncover unknown trends in widely disparate and/or archived data and data that is seemingly unrelated to the subject matter under investigation. The dataset in this study is relatively small yet is widely disparate. For this reason, many of the techniques developed from big data analytics and Bayesian statistical probability analysis are appropriate.

The key dataset for understanding essential services was the HOP per region per day, which was supplied by the Iraqi Transition Assistance Office and the Multinational Force-Iraq/Multinational Corps-Iraq Fusion Cell for the periods of May–August 2007 and May–August 2008. It shows the variation of the unmet demand on public utilities in Iraq during a wartime environment. This dataset was difficult to obtain, and the time period available in this dataset is the only parameter available for understanding the importance of essential services in Iraq. The records in the dataset are presented in an interval scale of hours.[31]

Instead, the machine learning algorithm was able to accomplish the same task in a fraction of the time and cost.

The data on the incidence of attacks comes from the University of Maryland's Global Terrorism Database (GTD). The GTD is an open-source data set with over 190,000 cases from 1970–2018.[32] The GTD also includes data columns on variables such as attack type, target type, weapon type, perpetrator, casualties, fatalities, and injuries. When the information was uncertain, the column was coded as "unknown."

The air temperatures by day and by city were provided through the National Oceanic and Atmospheric Administration at the National Center for Environmental Information on an interval scale in Fahrenheit.[33] Because so many Iraqi weather stations were not included in the data collection, most of the data had to be scaled from the city level to the provincial level. Thus, the task involved identifying the closest station to each attack location in the GTD and assigning a temperature on that specific day. For the purposes of this study, the authors recorded the maximum daily temperature. The algorithm used to compute the closest station is the Vincenty solution of geodesics on the ellipsoid, which calculates geodesic distances between a pair of latitude/longitude points on the Earth's surface.[34]

In an important example of why to use a big data approach on archival data, this dataset required the transliteration of several Iraqi cities from the Iraqi Fusion Cell's dataset. To accomplish this, the authors used a machine learning algorithm to identify the name of the city and assign its matching location from the GTD. In addition, the dataset only contained information for 2007 and 2008.

[31] Karim, Power Generation Database.

[32] Global Terrorism Database.

[33] "Iraq Daily Temperature."

[34] "Vincenty Solutions of Geodesics on the Ellipsoid," Movable Type Scripts, accessed 11 January 2021, https://www.movable-type.co.uk/scripts/latlong-vincenty.html.

Thus, while the preliminary analysis could integrate the demand on public utilities as a potential covariate, analysis performed on later years could not. In short, cleaning and reconciling the disparate datasets and aligning the HOP with the separate conflict and temperature databases by hand would have taken hundreds of man hours. Instead, the machine learning algorithm was able to accomplish the same task in a fraction of the time and cost.

Through this application of big data analysis, the authors created a master database chronicling (by day and geolocation) maximum daily temperature, demand for public utilities, and the outbreak of violent conflict. When combined, the dataset for 2007 alone yielded over 8,850 data points. This data was uploaded, compiled, normalized, and harmonized. One issue with looking at this kind of data is that the domain types are different, and the measurements are not uniform. Thus, the authors needed to brush the data to normalize for environmental factors. The datasets also contained missing values, inconsistencies, and indications of ambiguous record keeping. Yet, despite these irregularities, the process of data fusion enabled a more robust analysis than that provided by any single data source.

Step 3b: Establish the Research and Methodology

The research questions in this study investigate whether there is a statistically significant correlation between demand on public utilities, air temperature, and the outbreak of conflict in the Iraqi context. To answer this question, the study aggregates data on essential services in conjunction with the incidence of attacks. In particular, it statistically analyzes Iraq in the summers of 2007 and 2008, forecasts the probability of attacks for 2014 and 2015 when the country experienced a marked increase in violence, and then compares the model's results to the actual 2014 and 2015 attack and temperature data.

The dependent variable in this study is the incidence of attack. It is measured on a nominal scale as "attack" or "no attack" listed in the GTD.[35] The authors then used algorithms to create data layers that provide contextualization for the information. Through the use of machine learning and automation, the data was integrated to create a geospatial picture of human conflict with relationship to environmental variables. Once the data was fully incorporated into the master database, researchers and data scientists began computation on the aforementioned hypothesis using a variety of statistical methods, including Bayesian analysis, to create probabilities.

The period under study contains a wide variety of environmental factors and conditions that are impossible to control. For instance, the majority of attacks in the country occurred in Baghdad province. The wartime situation was dynamic, and political realities were constantly shifting. It is not possible to replicate any of the conditions or factors in this study, and the information it reveals is unique to this time and place. The approach taken by the researchers is a fundamental shift from control-based experimental analysis. Instead, this analysis is fundamentally forward-thinking and simply seeks to identify the likelihood of events based on historical data at widely divergent scales. The main takeaway here for emphasis on data science teams is that models from one conflict cannot

[35] Global Terrorism Database.

be applied to another conflict environment. While the general data science approach is replicable, other countries could require a different set of SMEs, conflict hypotheses, and databases to analyze.

Step 4: Analyze the Results

The results were categorized into a framework encompassing four steps in the application of data: descriptive and exploratory analysis, predictive statistics, detection, and evaluation/prescription.[36] The first step—descriptive and exploratory analysis—showcases the context of the data. The second step—predictive—uses pattern analysis to identify the probability of different outcomes. The third step—detection—gives greater context to the anomalies in the dataset, and finally, the evaluation portion of the results provides recommendations on how to use the interpretations from this dataset in future stability operations.

The first set of hypotheses asks whether there is a positive correlation between unmet demand of public utilities and the incidence of attack. The second set looks at whether there is a correlation between temperature and the incidence of attack. Finally, the third set looks at whether there is a correlation between both unmet demand and air temperature on the incidence of attack. The first two hypotheses seek to determine an independent correlation while the third hypothesis seeks to determine if there is a compounding effect when combined as social experiences. Unlike the first two hypotheses, instead of resulting in a purely linear extrapolation of the potential correlation, the third hypothesis assesses compounding changes to human conflict as a function of unmet demand of public utilities and air temperature.

> H_1: There is a positive correlation between unmet demand of public utilities and the incidence of attack.
> H_0: There is no correlation between unmet demand of public utilities and the incidence of attack.
>
> H_2: There is a positive correlation between air temperature and the outbreak of human conflict.
> H_0: There is no correlation between air temperature and the outbreak of human conflict.
>
> H_3: There is a positive correlation between both unmet demand on public utilities and air temperature and the outbreak of human conflict.
> H_0: There is no correlation between both unmet demand on public utilities and air temperature and the incidence of attack.

[36] This framework for the application of big data is outlined in a United Nations report on integrating big data, and further emphasized in a JSOU report on the same topic. See Michael Bamberger, ed. Tamara Karaica and Felicia Vacareulu, *Integrating Big Data into the Monitoring and Evaluation of Development Programs* (UN Global Pulse Report, 2016): 22, 57-59; Tammy Low, *Exploitation of Big Data for Special Operations Forces* (Tampa: JSOU Press, 2018): 110.

Step 4a: Descriptive and Exploratory Analysis

The subsequent analysis provides an illustration of the outcome generated through the process of descriptive and exploratory analysis. This analysis showcases the results from the analysis from May–August 2007 as an example of the type of information preliminary analysis can yield.

The first set of figures used in describing datasets are typically frequency tables and graphs. While machine learning algorithms were necessary for creating the database to conduct the descriptive and exploratory analysis, big data techniques are not needed for this first level of analysis. For this study, the frequency tables and figures show the number of attacks by month, the average temperature by month, and the average HOP in the region by month for the period of May–August 2007. Tables 1 and 2 and figures 2–5 give a description of the dataset and offer summary statistics of the datasets involved in the summer of 2007.

Table 1. Dataset frequency table by month. Source: Authors

	MAY 07	JUN 07	JUL 07	AUG 07
Average HOP	13.41758	12.57143	12.93548	13.66359
Number of Attacks	27	197	56	25
Average Maximum Temperature	99.36964	105.5167	110.0079	109.0171

Table 2. Dataset frequency table by day, summer 2007. Source: Authors

	MEAN	STANDARD DEVIATION	MIN	MAX	COUNT
HOP	13.17632	3.875452131	4	24	777
Maximum Temperature	106.2618	6.593001532	77	120.2	826
Attacks	1.859756	1.542395536	1	10	305
Number of Attacks without HOP*					68

**1 week missing from dataset*

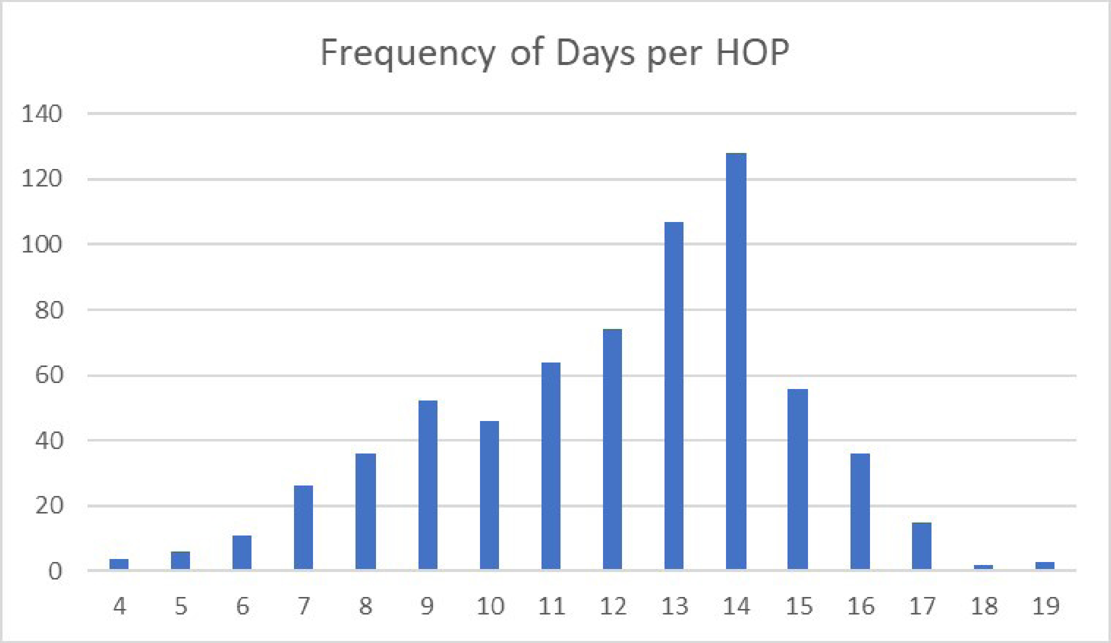

Figure 2. Frequency of days per HOP. Source: Authors

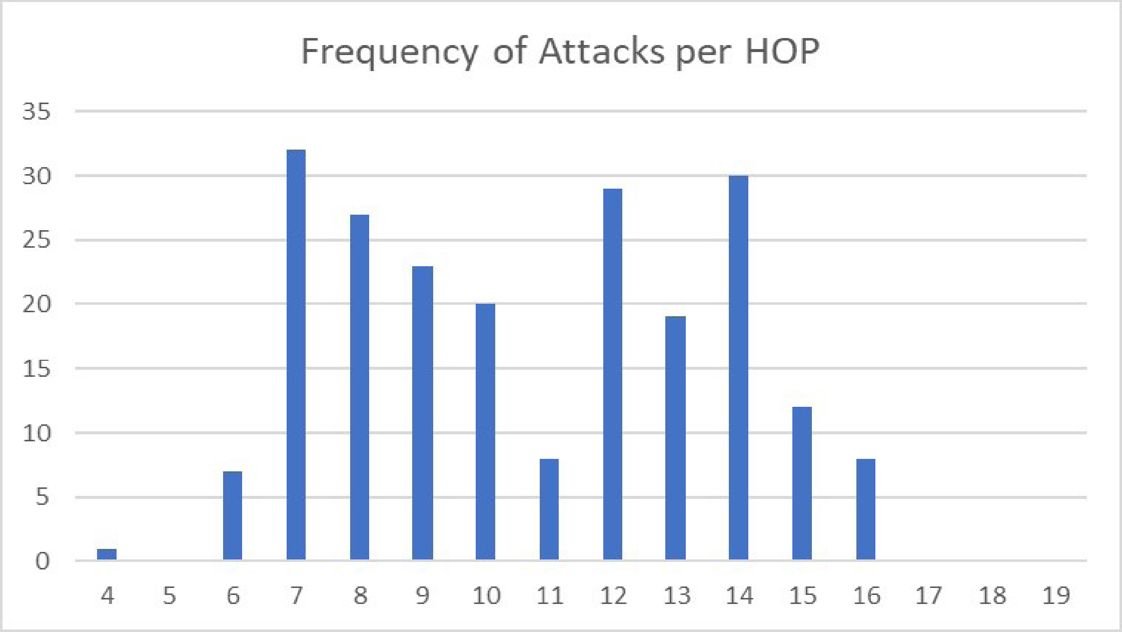

Figure 3. Frequency of attacks per HOP. Source: Authors

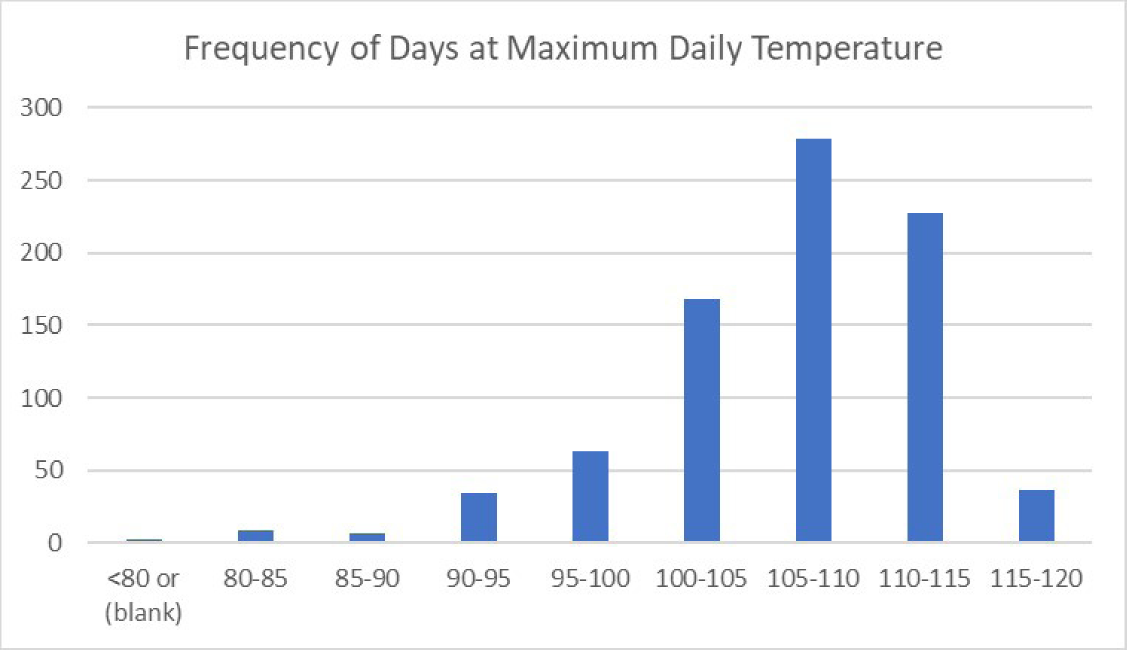

Figure 4. Days versus temperature. Source: Authors

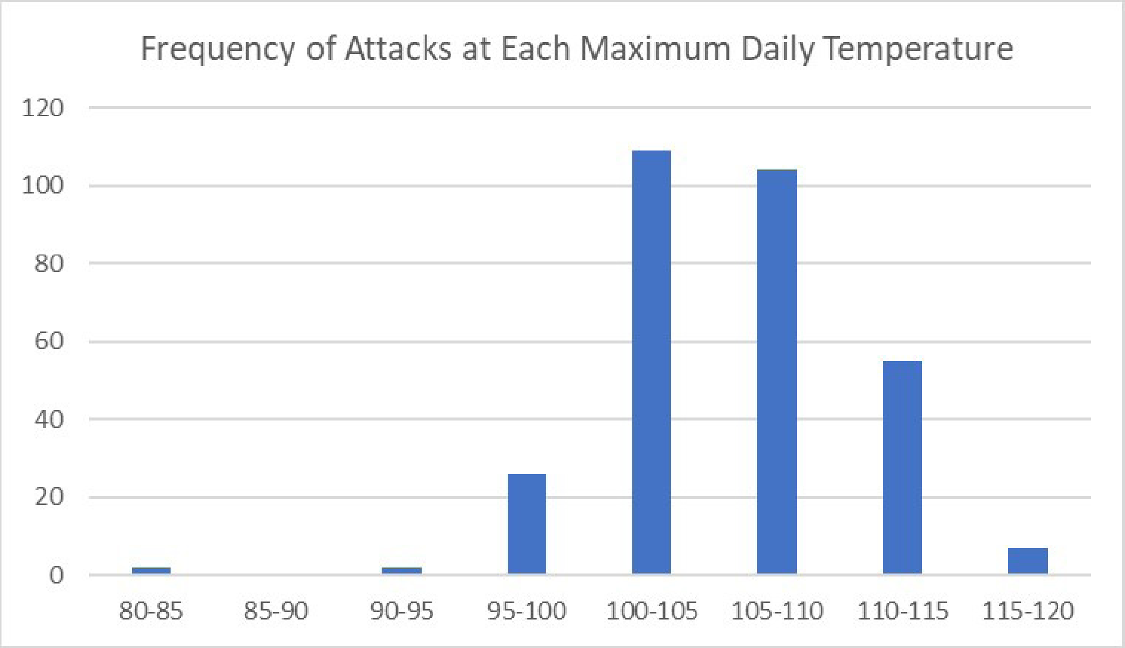

Figure 5. Frequency of attacks at each maximum daily temperature. Source: Authors

Specifically, tables 1 and 2 are dataset frequency tables that show the number of attacks, average maximum temperature, and average HOP by month, as well as the number of days with attacks, and a summary of the means and standard deviations from the dataset. Figures 2–5 show the frequency of days in Iraq across the seven most attacked provinces at certain HOP levels and the frequency of attacks at those same HOP levels, and the frequency of days in the dataset across the seven most attacked provinces in Iraq within the given temperature range and the frequency of attacks at those same temperatures.

The dataset reveals that HOP ranged from a minimum of 4 to a maximum of 24 hours per day. The maximum daily temperature in Iraq during this period ranged from 77°F to a high of 120.2°F. In the summer of 2007, there were a total of 305 recorded attacks over a 118-day period across the seven most attacked Iraqi provinces.[37]

Figure 6 shows a plot of the 75 cities in Iraq during the 118-day period under observation. Each circle represents one day in the study for a total of 8,850 data points.[38] The days are plotted on a grid by their daily HOP and their maximum daily temperature. The circles are coded red if there was an attack on that day. The intensity of the color indicates a higher number of days with an attack. This exploratory analysis reveals the majority of attacks were concentrated between 6 and 14 HOP and 100–110°F.

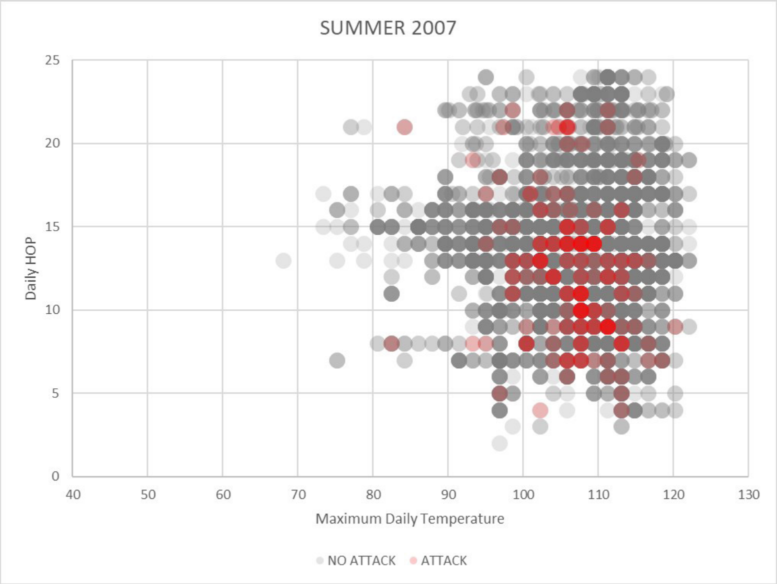

Figure 6. Data plot summarizing all 8,850 data points: daily HOP across 75 Iraqi cities versus maximum daily temperatures in Iraq, May–August 2007. Source: Authors.

[37] The researchers assessed al-Anbar, Babil, Baghdad, Diyala, Kirkuk, Ninawa, and Salah al-Din provinces. The provinces of al-Qadisiyah, Dhi Qar, Erbil, Karbala', Maysan, al-Muthanna, al-Najaf, and Wasit had fewer than three attacks during the time period under observation. The province of al-Basra was missing one week of HOP data; observation and detection analysis revealed the data was likely not missing at random because of the high frequency of attacks during that week.

[38] There are 118 days under observation across 75 Iraqi cities, bringing the total number of data points to 8,850.

The data in figure 6 reveals that the maximum temperature was frequently above 100°F, with the most frequent range between 105°F and 110°F. The most frequent number of HOP the Iraqis received in one day was 14, but there was a wide variation in how much power one could expect on any given day. While one would expect the majority of attacks to occur on the most frequent temperature range, a quick visual interpretation shows that, while the majority of days in Iraq had a maximum temperature between 105°F–110°F, most of the attacks occurred between 100°F–105°F. Similarly, one would expect that the most attacks would take place on days with 14 HOP, yet the frequency graph of days of HOP in Iraq takes on a normal (bell-curve) distribution; the frequency of attacks at their corresponding HOP is widely distributed with a spike at 7 hours, a second spike at 12 hours, and a third spike at 14 hours (which was the most common HOP in the dataset).

To show geocoded representation of the data, figure 7 shows the frequency of attacks on a map of Iraq. The more intense color represents a higher frequency of attacks. As expected, this analysis shows that the majority of attacks took place in areas with high population density (e.g., Baghdad and Mosul) but also in areas along major roadways in the country. Similarly, figure 8 shows the frequency of attacks by province and by the week during the year. The first week is week 19, which represents the first full week of May 2007, and ends with week 35, which was the final week of August 2007. When the data is presented using this method, it becomes apparent that the majority of attacks in the country took place in Baghdad, Kirkuk, Ninawa, and Salah al-Din provinces early in the summer, and there was also a spike in attacks in Baghdad during weeks 27–30.[39]

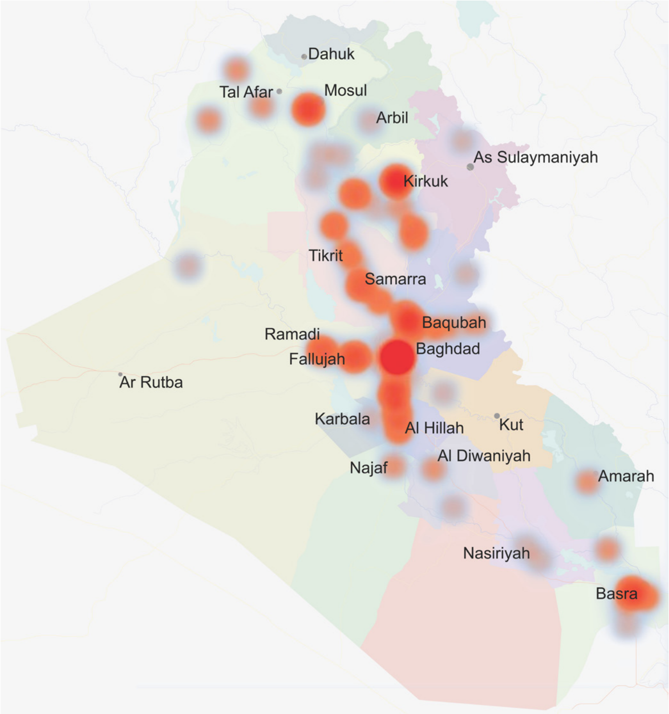

Figure 7. Geocoded frequency analysis of attacks in Iraq, summer 2007. Source: "Iraq regions map" by Peter Fitzgerald, https://creativecommons.org/licenses/by/3.0/deed.en, modified from original.

[39] Due to limitations in the dataset, Iraq's southern provinces have been omitted from this analysis.

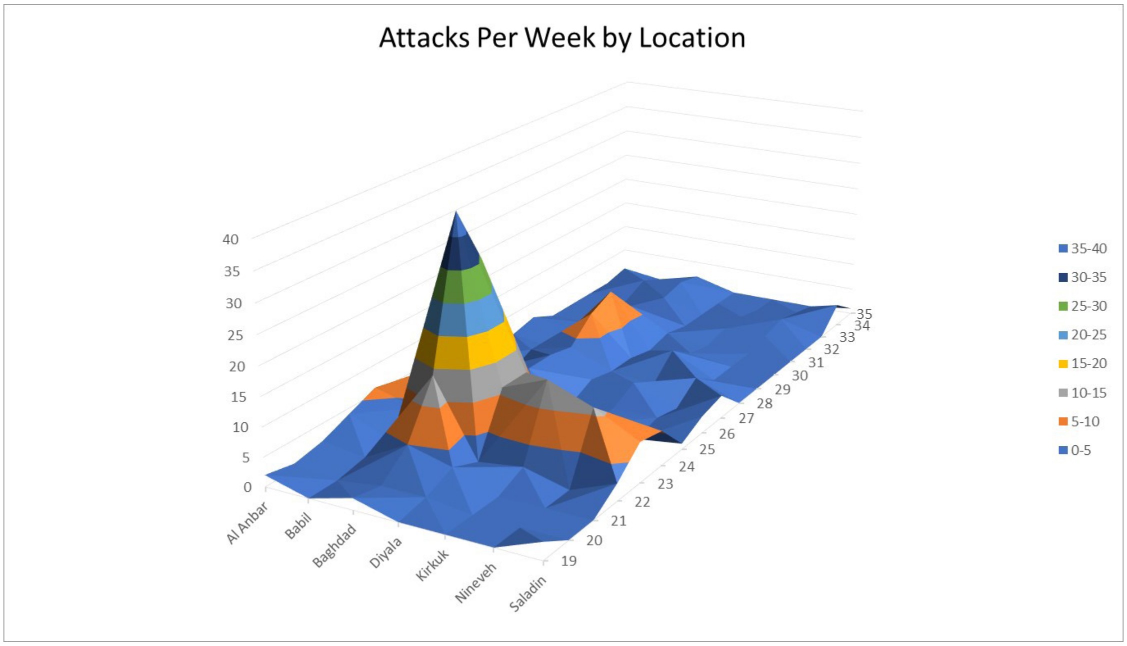

Figure 8. Attacks by week and location in Iraq, summer 2007. Source: Authors.

Step 4b: Predictive Probability Statistics

To create more predictive analytics from this dataset, the researchers adjusted the hypothesis to a wartime scenario and asked questions that were similar to the aforementioned hypotheses but based on probability.

Question 1: Does the probability of attack increase at a certain temperature?

Question 2: Does the probability of attack increase at a certain HOP range?

Question 3: Does the probability of attack increase at a certain temperature and HOP range?

Taking the seven most attacked provinces of Iraq into consideration, and using Bayesian techniques, it is possible to graph the probability of an attack given the temperature and the HOP per day. Figure 9 shows a somewhat normal distribution peaking at seven HOP. If the probability of attack were independent of the HOP, one would expect that the probability would be uniform across the graph. Yet, when controlling for the density of HOP in the study, this analysis suggests the probability of attack is nearly 20 percent at seven HOP and, as expressed in hypothesis 1, shows a decreasing fluctuation as HOP increases.

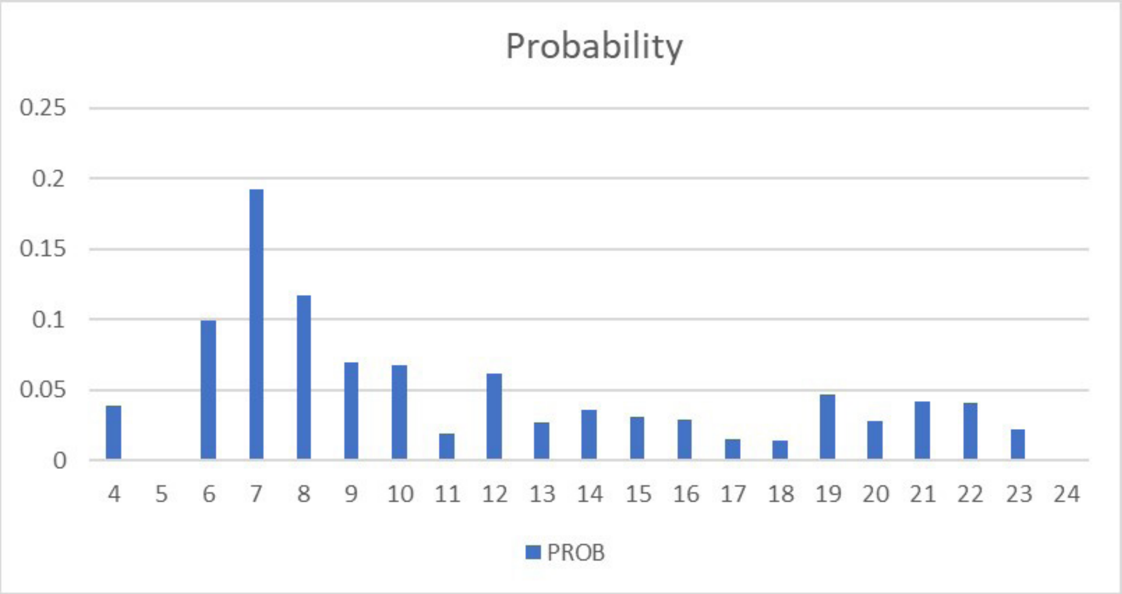

Figure 9. Probability of attack based on HOP, 2007. Source: Authors.

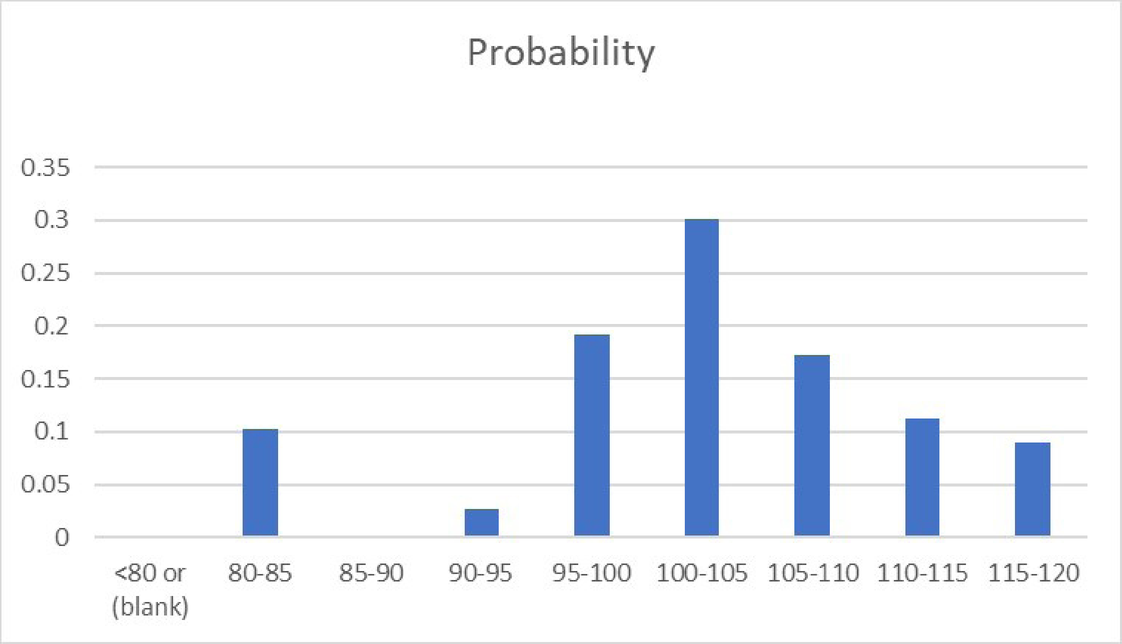

Figure 10. Probability of attack based on temperature, 2007. Source: Authors.

Meanwhile, figure 10 shows that the probability of an attack is not distributed normally, nor is the probability of attack independent of the temperature. If the probability of attack is not correlated with the temperature, one would expect that each probability would be the same. When controlling for the frequency of days at specific temperatures, the analysis reveals there is an outlier of probability between 80°F–85°F, low probability between 85°F–90°F, and a normal distribution peaking at 30 percent between 100°F–105°F. The analysis reveals that when the temperature is between 100°F–105°F, there is a 30 percent chance of attack, which leads to a rejection of the null hypothesis in H2 that temperature is not correlated to incidence of attack.

Given the fact that the data is significantly skewed towards Baghdad, which suffered the majority of attacks, the researchers then controlled for the city and attack type. It is also possible to control for the density of the population as an intervening variable. This analysis reveals a 10 percent higher chance of attack when HOP is seven and between the temperatures of 105°F–110°F. In 2008, there is a 16 percent greater chance of attack when the HOP is seven and the temperature is between 110°F–115°F. See figures 11 and 12 below for more information.

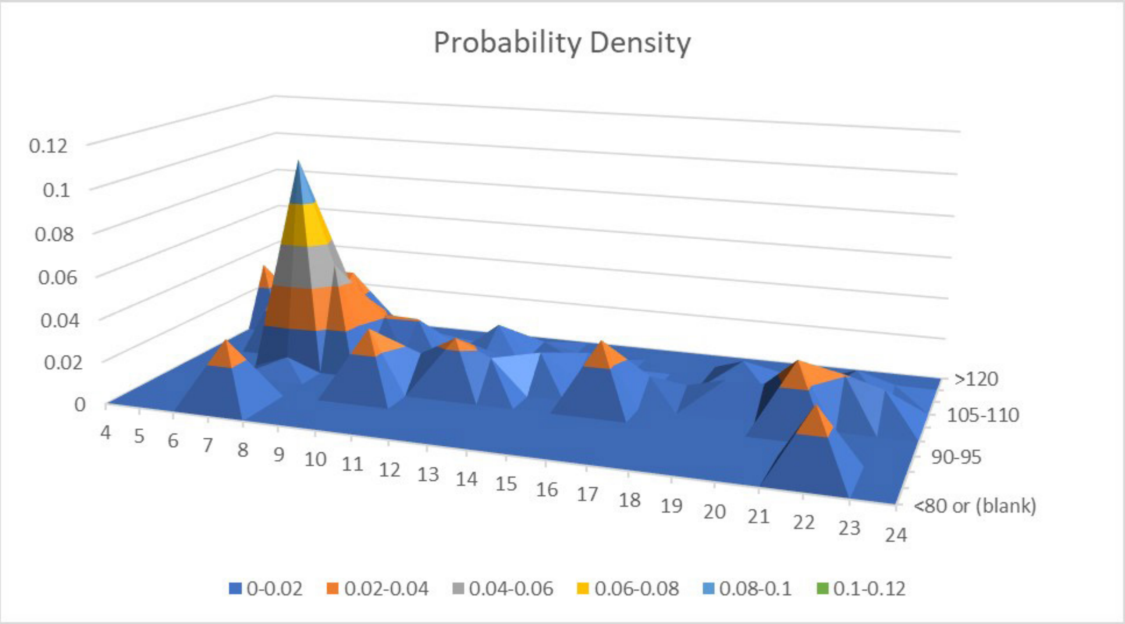

Figure 11. Probability of attack controlling for population density, 2007. Source: Authors.

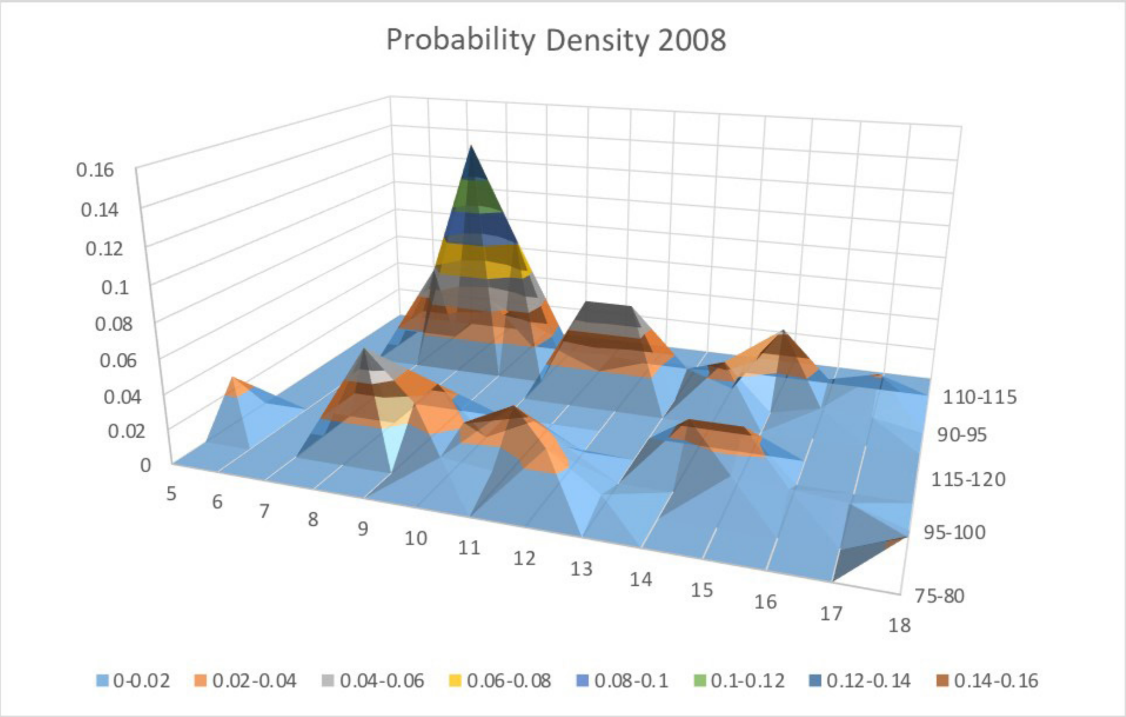

Figure 12. Probability of attack controlling for population density, 2008. Source: Authors.

Step 4c: Detection

The aggregated dataset revealed several anomalies and outliers. The first particularly startling anomaly was that an entire week was missing from the Iraqi Transition Assistance Office's database. While this could have been missing data at random, when looking at the complete dataset, the missing week included the hottest days on record in 2007, and it was during the densest weeks in terms of total attacks (68 or 22 percent of the attacks in the period under observation). A second anomaly in the dataset was that one Iraqi province, Anbar, was consistently above 14 HOP. The researchers surmised this anomaly was due to the region having a power source outside the compromised areas on Iraq's main grid.[40]

Another interesting finding is in the outliers of the dataset. The GTD dataset has a column for the perpetrator of the attack. While the majority of the 305 attacks in the 2007 study listed an unknown perpetrator, the majority of the incidents that fell drastically outside the expected HOP and temperature ranges listed the perpetrator. One hypothesis for this anomaly is that these attacks were better organized and planned far in advance of shifting weather and HOP fluctuations, thus they were independent of these variables. These outliers could also account for the anomalies in the 80–85-degree temperature range.

[40] The Haditha Dam in western al-Anbar province was built to stabilize Iraq's power grid; it is the second largest of eight hydroelectric dams in Iraq.

Step 4d: Evaluation and Prescription—Operational Significance

The archival data indicates unique sociocultural patterns in the context of Iraq. Without knowing anything about the current state of infrastructure, the archival data can reveal patterns with significant operational value to a SOF planner tasked with analyzing support to Iraqi counter-ISIS operations. For instance, a SOF planning team in 2014 could have used the archival data to model broader conflict dynamics not tied to specific actors. To showcase this phenomenon, the researchers looked at data from 2007–2008 to forecast the dynamics of 2014–2015.[41] Looking only at the data from 2007—2008, a SOF planner could anticipate several key operational factors.

First, the planner may want to know which temperature ranges are more likely to experience attacks on coalition personnel. The data derived from 2007–2008 shows that the 105–110°F temperature interval consistently yields a volume of attacks outside the expected range (figure 14). When observing the percentage of attacks by temperature interval for the periods under observation, the number of attacks is strongly correlated with the number of days at that temperature (figure 13). One may expect the 100–105°F temperature range would yield the highest probability of attack because it was the most frequently occurring temperature range, yet, when controlling for the frequency of days with those temperatures, the probability of attack shows a marked increase at the 105–110°F temperature interval for each of the years under observation (figures 13 and 14). As the 2007–2008 dataset predicted, this also holds true in the aggregate (all four years), stacked by year, and in the summary by province.

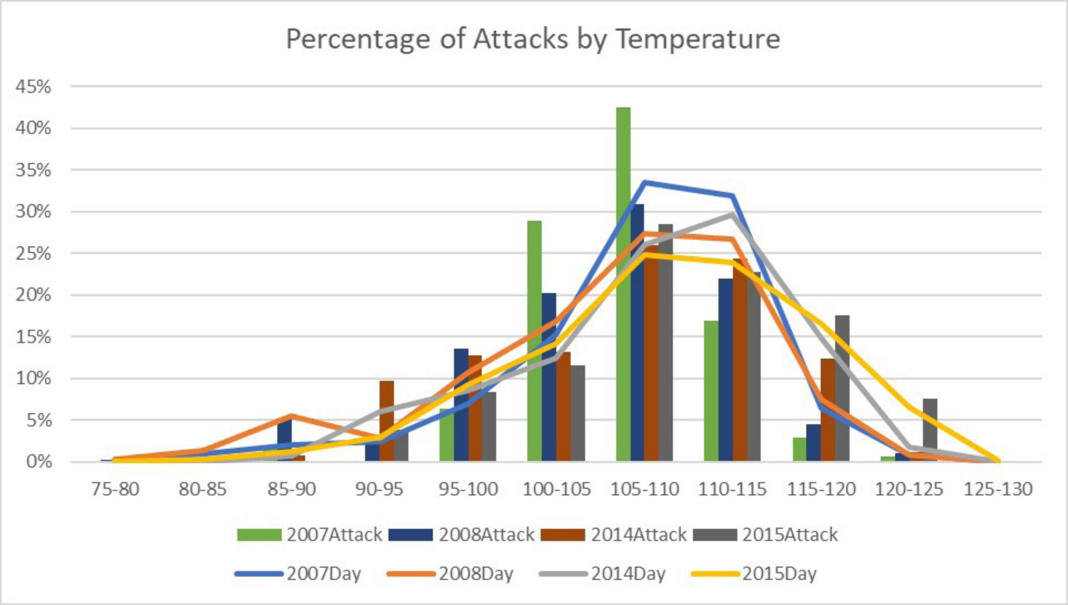

Figure 13. Percentage of attacks by temperature, Iraq. Source: Authors.

[41] It is important to note that the data analysis performed on the 2014 and 2015 datasets could not account for HOP as a potential covariate, thus compounding the difficulty of assessing the importance of essential services towards reducing violence. Instead, the researchers looked to other covariates in the datasets as plausible indicators of human conflict.

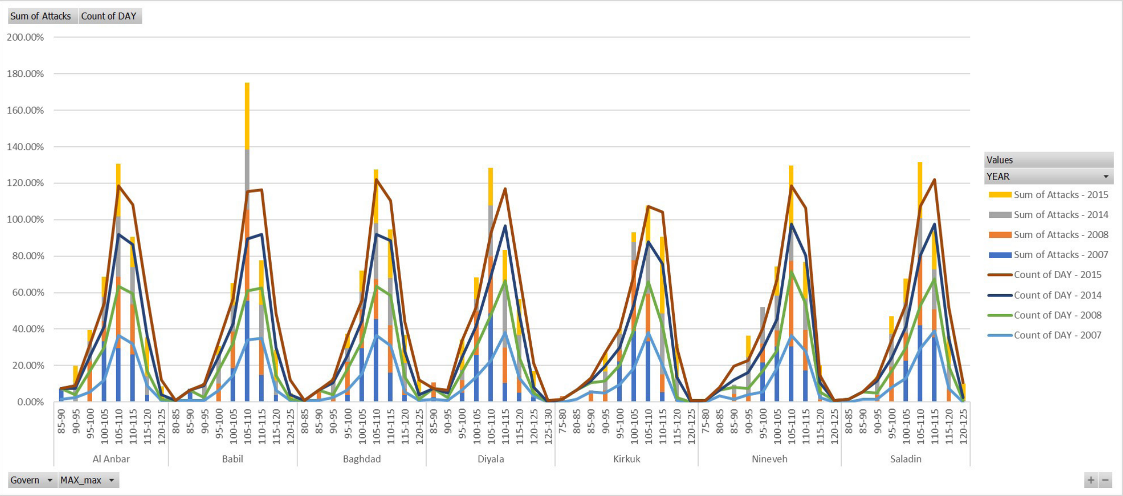

Figure 14. Sum of attacks by province by temperature interval. Source: Authors.

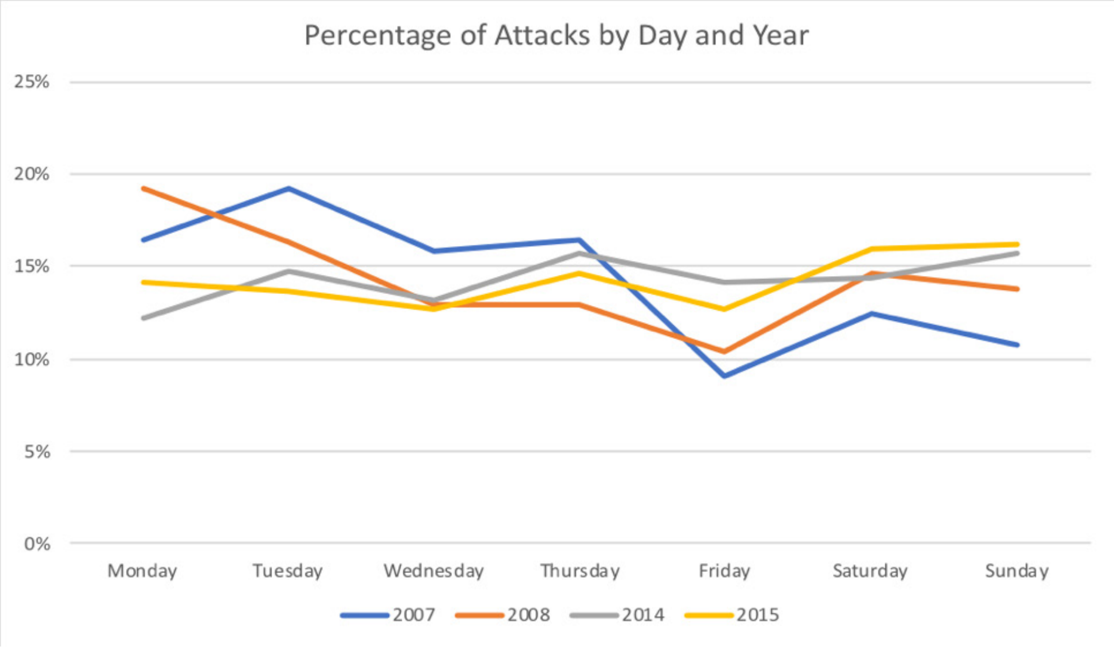

Figure 15. Percentage of attacks by day and year. Source: Authors.

Another key insight for the SOF planner may be in looking at the days of the week that yield the most attacks on coalition personnel. Looking at the data from 2007–2008, the analysis revealed marked downward shift in the number of attacks that took place on Fridays (figure 15). Culturally, Fridays in Arab countries are days of worship for adherents to the Islamic faith. Yet, looking at the data from 2014, this trend does not hold. The daily pattern of attacks in 2014 appears relatively consistent across the week, where 2015 sees a return to the ebb in Friday attacks and subsequent uptick in attacks on Saturdays. Religious observations may have influenced the steady downward trend in attacks on Fridays the years under the period of observation, where 2014 shows a clear deviation from the standard.

Finally, the SOF planner may look to see which weeks of the year were more violent in terms of attacks on coalition personnel. In 2007 and 2008, week 24 (late June) was significantly more violent than other weeks during the year. This could be due to a variety of factors, including annual holiday periods, special events, or crop and harvest cycles. The SOF planner would find that, like 2007 and 2008, week 24 of 2014 was significantly more violent than the other weeks of the year, with attacks on coalition personnel and other terrorist attacks following similar ebb and flow patterns to years past.

Conclusion

This project contains several important elements for using data science for operational-level intelligence and planning. First, it illuminates how archived wartime data and seemingly disparate metrics can be used in current analysis that may benefit special operations in the future. As researchers draw upon the data and theories outlined here, the ultimate objective is to create a framework for utilizing data science teams to generate locally relevant theories of conflict that will explain the relationship between critical factors of good governance, as exemplified in the Iraq case through essential services. Other environments will undoubtedly have different stressors, and only SMEs in those operating environments will be able to generate insight into the drivers of conflict.

Second, the massive shift in data analytics requires USSOCOM to keep pace with peers and competitors. Most researchers acknowledge there is a glaring need to identify best practices for identifying, categorizing, and integrating the most relevant and crucial information from large swaths of streaming data, yet archived data can also yield important insights for intelligence analysts and planners. Furthermore, there is a need for special operations domain-specific analytic platforms and a cadre of SOF data scientists that understand both the computational and the social science lines of inquiry.[42] While artificial intelligence and automated computing can fill some of the gaps, human insight is still required to pose hypotheses, interpret data analytics, and aggregate widely disparate data. Moreover, it is imperative that SOF planners create a feedback mechanism with their data science teams so the models can be updated or amended with new hypotheses and theories informed through intelligence gathered on the ground. Sensitivity to local conditions is the first step in generating better intelligence questions for data analysis.

Finally, this project offers the beginning of a modeling project for predictive analysis on the correlation between essential services and the incidence of attack in an active wartime environment. By creating data layers from existing information from essential services and comparing those data points with instances of attack, this research ultimately seeks to provide better models to forecast patterns of conflict in different sociopolitical contexts. It is feasible to overlay this analysis

[42] The National Academies of Sciences, Engineering, Medicine, introduction to *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions* (Washington, D.C.: National Academies Press, 2017).

with areas where SOF currently operate or in areas of concern. It is also possible to integrate this dataset with many more datasets. In the future, researchers could build a neural network and use Bayesian techniques to identify the most important factors from hundreds of thousands of potential intervening variables. This would provide a contextual visualization of the data specific to USSOCOM objectives and enable future operators with data-driven decision-making.

Acronyms

GTD	Global Terrorism Database
HOP	hour(s) of power
ISIS	Islamic State of Iraq aand Syria
JAM	Jaysh al-Mahdi
SME	subject matter expert
SOF	Special Operations Forces
USSOCOM	United States Special Operations Command

JOINT SPECIAL OPERATIONS UNIVERSITY
DEPARTMENT OF STRATEGIC STUDIES
7701 TAMPA POINT BLVD.
MACDILL AFB, FL 33621

JOINT SPECIAL OPERATIONS UNIVERSITY

JOINT SPECIAL OPERATIONS
JSOU
UNIVERSITY
PRESS