

Lesson (un)replicated: Predicting levels of political violence in Afghan administrative units per month using ARFIMA and ICEWS data

Tamir Libel* 

Department of Leadership and Command & Control, Swedish Defence University, Stockholm, Sweden

*Corresponding author. E-mail: tamirlibelphd@gmail.com

Received: 29 November 2021; **Revised:** 15 August 2022; **Accepted:** 29 August 2022


Key words: Afghanistan; forecasting; georeferencing; political violence; time-series analysis

Abstract

The aim of the present article is to evaluate the use of the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model in predicting spatially and temporally localized political violent events using the Integrated Crisis Early Warning System (ICEWS). The performance of the ARFIMA model is compared to that of a naïve model in reference to two common relevant hypotheses: the ARFIMA model would outperform a naïve model and the rate of outperformance would deteriorate the higher the level of spatial aggregation. This analytical strategy is used to predict political violent events in Afghanistan. The analysis consists of three parts. The first is a replication of Yonamine's study for the period beginning in April 2010 and ending in March 2012. The second part compares the results to those of Yonamine. The comparison was used to assess the validity of the conclusions drawn in the original study, which was based on the Global Database of Events, Language, and Tone, for the implementation of this approach to ICEWS data. Building on the conclusions of this comparison, the third part uses Yonamine's approach to predict violent events in Afghanistan over a significantly longer period of time (January 1995–August 2021). The conclusions provide an assessment of the utility of short-term localized forecasting.

Policy Significance Statement

The current article demonstrates the feasibility, advantages, and limitations of combining the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model with data from the Integrated Crisis Early Warning System (ICEWS) in predicting spatially and temporally localized political violent events. By outlining analytical strategy that is tested on data concerning violent events in Afghanistan between 1995 and 2021, the article introduces an approach to predict violent events in the subnational level (i.e., province and district levels) 1 month in advance. This could be of help to relevant policy actors, for example, humanitarian or monitoring agencies, in devising timely mitigation or prevention measures. In addition, the article supports the emerging focus in the forecasting literature on local rather than national prediction of violent events.

 This research article was awarded an Open Materials badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Introduction

The aim of the present article is to evaluate the use of the Autoregressive Fractionally Integrated Moving Average (ARFIMA) in predicting spatially and temporally localized political violent events, in accordance with the operationalization suggested by Yonamine (2013), and using the Integrated Crisis Early Warning System (ICEWS) (Lockheed Martin Advanced Technology Laboratories (ATL), 2021) that was not available for him. This article is, therefore, part of the “predictive turn” sweeping over conflict studies literature in the past decade. Drawing upon the growing recognition that explanatory modeling, that is, statistical significance, is rarely associated with predictive power, an increasing number of studies have pursued the development of methods that improve predictive accuracy. This effort was enabled, in turn, by the new availability of high-frequency georeferenced “big data” repositories, the advancement of machine learning algorithms, and scalable cheap high-performance computing. Gradually, a consensus emerged that short-term, local (i.e., subnational) predictions may provide the most promising venue for further development of forecasting accuracy.

Arguably, Yonamine’s (2013) was the first study to use the emergence of the first close-to-real-time repositories—Global Database of Events, Language, and Tone (GDEL)—for subnational forecasting. He sought to demonstrate the utility of the ARFIMA model, shown in other fields with univariate time-series data, in the provision of “more accurate and consistent forecasts than other time-series models” (Yonamine, 2013, p. 9). Although his work has received much attention over the years, his analytical strategy was not replicated. Nevertheless, such a replicatory exercise could be both timely and advantageous for scientific and policy stakeholders alike, as it offers a more feasible and straightforward approach than other more complicated algorithms at a time of greater availability of more rigorously screened data sources.

Therefore, following a review of the relevant literature, the present study focuses on replicating Yonamine’s (2013) analytical strategy. Special attention is given to testing his two hypotheses: the ARFIMA model would outperform a naïve model and the rate of outperformance would deteriorate the higher the level of administrative unit (i.e., aggregation). Since the original article focused on predicting political violence in Afghanistan, the replication of its analytical strategy is conducted on ICEWS data for Afghanistan. The events of August–September 2021, marking the end of the American involvement there and the Taliban’s regaining of control, have also rendered the analysis presented here an overview of the temporal-spatial dispersion of political violence in Afghanistan over the last two decades.¹

Methodologically, the replication exercise consists of three parts. The first is a replication of Yonamine’s (2013) study for the period beginning in April 2010 and ending in March 2012. Similarly to the original study, the data for this timeframe was divided into an initial training set (ITS) (September 2004–March 2010) and test set (TS) (April 2010–March 2012). The second part compares the results to those of Yonamine (2013). The comparison was used to assess the validity of the conclusions drawn in the original study, which was based on GDEL data, to the implementation of this approach to ICEWS data. Building on the conclusions of this comparison, the third part uses Yonamine’s (2013) approach to predict violent events in Afghanistan over a significantly longer period of time (January 1995–August 2021). The division into an ITS and TS was based on backtesting. The conclusions of this section provide an assessment, drawing on the methods and data source used, of the utility of short-term localized forecasting.

Literature Review

In 2010, Shmueli (2010, p. 290) observed that “In many scientific fields, and especially the social sciences, statistical methods are used nearly exclusively for testing causal theory.” Put differently, she

¹ The implications of this tragic development are beyond the methodological focus of the current article and its timeframe. Similarly, while the current article traces the evolution of violent events in Afghanistan, its focus is on validating a particular prediction approach—that is, ARFIMA. Hence, it is not engaged with the vastly excellent literature analyzing the drivers, causes, and implications for the failure of state building in Afghanistan. On this topic, see Barfield (2016) and Murtazashvili (2022).

pointed out that the majority of social scientific literature focuses on *explanatory modeling*,² resulting in the neglect of *predictive modeling*.^{3,4} Naturally, “the consequence of neglecting to include predictive modeling and testing alongside explanatory modeling is losing the ability to test the relevance of existing theories and to discover new causal mechanisms” (Shmueli, 2010, p. 304). Coincidentally in the same year, several prominent conflict studies scholars (Ward et al., 2010, p. 363) observed that the same was true in the field of conflict studies: “[I]t is not uncommon in conflict research to conduct statistical analyses and then draw policy inferences from the statistical information, without ever trying to make specific predictions” (Ward et al., 2010, p. 363).⁵

Moreover, they warned that this state of affairs had severe implications for theory building and conflict prediction alike: “[B]asing policy prescriptions on statistical summaries of probabilistic models (which are [italic in the original] predictions) can lead to misleading policy prescriptions if out-of-sample predictive heuristics are ignored” (Ward et al., 2010, p. 364). Replicating two prominent studies in the field, the authors demonstrated the inadequacy of explanatory modeling for prediction purposes (Ward et al., 2010, pp. 365–366, 372; Chadeaux, 2017, p. 8; Bara, 2020, p. 179).⁶ There are two reasons for this situation: “First, the information needed for prediction may simply be impossible to obtain... Another reason why theories might explain without predicting is the idiosyncrasy of the phenomenon” (Chadeaux, 2017, p. 8). As a result, scholars and practitioners alike had shelved the potential contribution that predictive modeling could make to model comparison and selection (Chadeaux, 2017, p. 8; Cranmer and Desmarais, 2017). In this sense, predictive modeling may be used as a validation measure for explanatory modeling (Cranmer and Desmarais, 2017, p. 146; Dowding and Miller, 2019, pp. 1002–1003).

However, over the last decade the situation has changed significantly, culminating in what may be termed a “predictive turn” in the literature (Chadeaux, 2017, p. 7; Colaresi and Mahmood, 2017, p. 93; Dowding and Miller, 2019, p. 1005).⁷ The change was characterized as:

“In the past decade, the field of conflict studies has seen increasing debates over the distinction between explanation and prediction and the implications of this distinction for research. From these

² Following Shmueli (2010, p. 290), explanatory modeling is defined hereby as: “The use of statistical models for testing causal explanations.” Cranmer and Desmarais (2017, p. 146) relied on Shmueli (2010) to provide a similar definition that is adapted to the field of political science (albeit with redefining “explanatory modeling” as “inferential modeling”): “In inferential modeling, the statistical model is constructed as an operationalization of a theoretical model. The specification is important because deviations from the theoretical model in operationalization inhibit our ability to use the statistical model to test hypotheses. The coefficients are the objects of interest, which is to say that the *statistical model itself is the object of interest* [italic in the original]. In inferential modeling, we use the data to learn about the statistical model.”

³ Following Shmueli (2010, p. 291), predictive modeling is defined hereby as: “The process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations”. Cranmer and Desmarais (2017, p. 146) relied on Shmueli (2010) to provide a similar definition that is adapted to the field of political science: “...in pure predictive modeling, the objects of interest are the variables rather than the parameters: we use the available data to produce the best possible predictions of the outcome variable. It does not matter, for purely predictive exercises, whether the statistical model used is a close operationalization of a causal theory, because the only metric for the quality of a model here is its predictive performance.”

⁴ In line with Colaresi and Mahmood (2017, p. 193), a *model* as opposed to *modeling* is defined hereby as: “...a stylized representation of the underlying process of interest.”

⁵ The seminal role of this article was captured succinctly by Bara (2020, pp. 179–180) who observed that “...it was not until an article by Ward et al. (2010) that the fundamental distinction between explanation and prediction reached a broader audience of conflict researchers... What drove the message home is that the authors subjected two of the most widely cited models of civil war onset to out-of-sample testing and found that their contribution to predicting future conflict was massively less impressive than the theoretical claims the authors had made on the basis of their statistically significant findings. Since then, out-of-sample predictions are a crucial part of what could be termed a conflict prediction paradigm.”

⁶ This point is summarized neatly by Chadeaux (2017, p. 8): “This focus on theoretically-motivated approaches could be justified if, as is typically assumed, statistical models with high explanatory power also had high predictive power. Yet this is often not the case, and relying on *p*-values for the purposes of policy formulation is risky at best.”

⁷ This opinion about the state of conflict studies is not, however, universally shared. See, for example, Cederman and Weidmann (2017, p. 474): “...prediction remains highly controversial in academic conflict research. Relatively few conflict experts have attempted explicit forecasting of conflict.” In spite of this stance, they go on to survey efforts to correct this situation.

debates, a subfield has emerged within conflict studies that is dedicated to prediction specifically.” (Bara, 2017, p. 177)

It was facilitated by advancement in data availability, computing power, and analytic methods (Cederman and Weidmann, 2017, p. 474; Dowding and Miller, 2019, p. 1008; Bara, 2020, pp. 183, 186). Methodologically, scholars began to turn increasingly away from the traditional focus on using statistical inference methods to analyze structural variables that are usually measured in country-year units, which vary little and slowly if at all (Cederman and Weidmann, 2017, p. 474; Chadeaux, 2017, p. 10; Dowding and Miller, 2019, p. 1013; Bara, 2020, p. 182). While these studies contributed significantly to understanding the causes and dynamics of conflicts, that is, explanation, they generally failed to pinpoint which of them would actually break into war and when, that is, prediction (Chadeaux, 2017, p. 10; see also Cranmer and Desmarais, 2017, p. 163).⁸ As an alternative, the “predictive turn” literature turned to resampling techniques and out-of-sample validation in order to “...guard against the inclusion of long lists of explanatory factors that may worsen predictive performance”(Cederman and Weidmann, 2017, p. 475). These efforts were supported by policy actors, for example, humanitarian organizations and intelligence agencies, that hoped to capitalize on their hopefully improved prediction accuracy (Chadeaux, 2017, p. 7; Bara, 2020, pp. 184–185).

Combined, these advancements in literature led to a noticeable improvement in the prediction accuracy of conflicts and violent events. However, the abovementioned change in methods used, which often involved ones that were complicated and hard-to-interpret, rendered the explanation and communication of the results a considerable challenge (Dowding and Miller, 2019, p. 1015). In spite of this difficulty, the use of such methods, for example, artificial neural networks (ANNs), was almost inevitable as “...the interactions of risk factors generating violent outcomes are inductively inferred from the data, and this process typically requires highly complex models” (Cederman and Weidmann, 2017, p. 474). The need, as described below, also emerged from the growing availability of close-to-real-time georeferenced, automatically generated, datasets. For the first time, all of these innovations paved the way for short-term, localized (i.e., subnational level) conflict forecasting.

As mentioned above, the improved ability to forecast conflict, measured as prediction accuracy, came at the expense of the ease of explaining how the algorithm worked. This was a considerable difficulty for the policy communities that supported and partially financed the advancement in this field. In addition to an early warning on an impending crisis, stakeholders must be able to justify a call for action in order to mobilize decision makers and legitimize action (Cederman and Weidmann, 2017, p. 476). It was also a growing problem for the forecasting scientific community concerning identification of the relative and combined contributions of variables to the outcome. As a partial solution, scholars have recently begun to increasingly rely on theoretical-informed reasoning in the planning of predictive modeling (Blair and Sambanis, 2020).

All of these developments in “predictive turn” literature bear directly on what is perhaps the main venue of this research program—tackling a key shortcoming of the traditional conflict forecasting literature:

“A growing body of research has suggested that an additional shortcoming of existing studies of civil wars rests with the level of analysis. While most large-n studies of intrastate conflicts use country-level data, most conflicts are more local than that. By disaggregating the focus from the country-level to regions or localities, further progress can be made.” (Ward et al., 2010, p. 373)

Indeed, “predictive turn” literature tackled not only the challenge of spatially but also temporally localized prediction, that is, short-term prediction. This was made possible by the combination of advanced analytical tools and the larger, (much) higher-frequency, georeferenced datasets mentioned above. Combined, these developments led to improved accuracy, rendering studies more methodologically

⁸ For a recent concise review of the relevant literature, see: Blair and Sambanis (2020, pp. 1888–1890).

sound than traditional literature (Cederman and Weidmann, 2017, pp. 474–476; Chadefaux, 2017, p. 10; Bazzi et al., 2019, p. 2; Bara, 2020, p. 182).

In spite of the progress made by “localized prediction” literature, the approach faces several challenges. First, the “big data” repositories it has relied upon thus far may well contain inherent biases:

“Because political violence is often coded from secondary sources such as news articles, a high level of observed violence can be due to a high level of actual violence or a higher probability of reporting (or both). This makes prediction difficult. If anything, scaling up the size of data set—as in several projects that made use of automated event coding—is likely to exacerbate this problem, due to reliance on the same secondary sources.” (Cederman and Weidmann, 2017, p. 476)

Second, some of the methods used by scholars who work on this venue faced sharp criticism, for example, the use of a univariate time series for conducting temporally and spatially localized predictions (Bara, 2020). Third, more recently scholars have begun to explore the potential use of combining violent event data with rich information micro-level covariates to improve prediction accuracy (Bazzi et al., 2019, pp. 2–3). However, the unavailability of such data for the majority of the world makes these studies of limited scientific and policy-oriented value. On the other hand, the growing public availability of close to real-time (almost) globally georeferenced event data with close to global coverage makes the development of reliable methods for localized conflict forecasting using univariate time-series data a task of utmost importance.

This task was taken on relatively early in the “predictive turn” by Yonamine (2013). He successfully demonstrated the feasibility of relying on univariate time-series data and relatively simple-to-interpret statistical models for short-term, localized conflict predictions. Despite receiving significant attention, his study was neither accompanied by the releasing of a code to ensure reproducibility nor any effort to replicate his assumedly straightforward research design. Thus, the following sections are devoted to replicating his analytical strategy using a different, more reliable data source.

Methods

Data acquisition

The data for the present article were downloaded from the ICEWS repository on Harvard University’s Dataverse using the ICEWS R package (Beger, 2021). The ICEWS “is an early warning system designed to help US policy analysts predict a variety of international crises to which the US might have to respond (...).” This project was created by the Defense Advanced Research Projects Agency but has since been funded (through 2013) by the Office of Naval Research (Ward et al., 2013, p. 1). The ICEWS is based on fully automated machine coding of news items from “over 6,000 international, regional, national and local news sources” (Boiney and Foster, 2013, p. 48). It is an example of what is known as automated event coders: “[A]utomated event coders code natural language text (usually from news sources) as categories of events” (Kotzé et al., 2020, p. 1). The coding was conducted “to extract <who, did-what, to-whom, when, where>, and performs updates in near-real time” (Boiney and Foster, 2013, p. 48).

The full ICEWS dataset retrieved on August 18th, 2021 consists of 19,140,780 observations for the time period January 1995 through August 2021. It was subset as described below into the main dataset used in the current study for the same timeframe. That dataset included 120,299 observations. A smaller dataset was extracted for the timeframe April 2010 to March 2012 as described below. It included 71,036 observations. The following sections describe the analysis conducted. The first part is a replication of Yonamine’s (2013) study for the period of April 2010–March 2012. The data for this timeframe were divided similarly to the original study to an ITS (September 2004–March 2010) and a TS (April 2010–March 2012).⁹ The second part compared the results to those of Yonamine (2013). The comparison was

⁹Due to this division, April 2004 was also used, as predictions were made based on information for March 2004.

used to assess the validity of the original study's conclusions, which were based on GDELT data, to the employment of this approach on ICEWS data.

Drawing on the conclusions of this comparison, the third part employed Yonamine's (2013) approach to predicting violent events in Afghanistan over a significantly longer period of time (January 1995–August 2021). The division into ITS and TS was based on backtesting. The conclusions of this section provide an assessment, based on the methods and data source used, of the prevalent argument in literature whereby, for the time being, the best predictions for which we can hope are near future, georeferenced ones.

Data preprocessing

The present article focuses on events in Afghanistan, and therefore, the first step was to subset all events that took place in Afghanistan using the ICEWS “country” variable, which was set to “Afghanistan.” Moreover, concentrating on physical violent events required a way of identifying and extracting them. ICEWS enables this as it classifies data according to the Conflict and Mediation Event Observations (CAMEO) typology (Ward et al., 2013, p. 1).¹⁰ The CAMEO ontology captures the action which took place, as well as its direction, between two actors. Each observation is identified by a unique code of either three or four digits. The identifiers were used to assign a quad code to each observation. These quad codes were used to create “quad counts” of the events in the following categories (Chiba and Gleditsch, 2017, p. 278). Each category's quad code classification appears alongside it in brackets: verbal-cooperation (a); material-cooperation (b); verbal-conflict (c); and material-conflict (d).

Despite their demonstrated utility, some argued that quad counts “are useful for distilling complex data sets into simple variables but are too coarse to capture the dynamics of escalation and de-escalation that procedural theories propose” (Blair and Sambanis, 2020, p. 1890). This critique does not pertain to the current study, however, as it does not focus on capturing conflict dynamics but on event prediction only. However, this may hinder the use of quad counts for engagements with the literature on state-building failure and violence in Afghanistan, which are focused on identifying and explaining the mechanisms driving violence.

Because the article centers on the prediction of political conflict events in their spatial settings, events had to be matched to the places in which they occurred. To do so, use was made of the observations' longitude and latitude variables,¹¹ which are provided as part of the ICEWS data. The coordinates were used to geolocate the event within its administrative unit, that is, district or province, requiring reliable, detailed data on the division of Afghanistan into its respective administrative units. This information was retrieved from the Global Administrative Areas (GADM) dataset (Global Administrative Areas, 2012). Its country data files on Afghanistan were used to assign the coordinates of events to the required administrative units via a customized function.¹² The latter conducts reversed geocoding that assigns an administrative unit's name (e.g., district or province) to a new variable (i.e., either “district2” or “province2”). Overall, 318 districts and 34 provinces¹³ were identified by their unique names.¹⁴ Subsequently, as a final step, the observations which were classified as “quad 4,” that is, material-conflict, were subset.

¹⁰ For background on CAMEO, see: Gerner et al. (2002). Moreover, the ICEWS team has “improved the CAMEO ontology, largely by resolving overlaps and clarifying guidelines for each extant type of event” (Ward et al. (2013, p. 3).

¹¹ These are provided in the ICEWS data as variables “latitude” and “longitude.”

¹² Only 316 districts had actual data and were included in the analysis. Thus, the “n” of the districts was set to 316 rather than 318.

¹³ In addition to the 34 provinces identified by their unique names, there was one “NA.”

¹⁴ The post-2001 Government of Afghanistan recognized 34 provinces as the first sub-national administrative unit. Therefore, the current study is in line with the official data, unlike Yonamine's (2013, p. 7), which argued that “Afghanistan is spatially divided into 32 provinces and 317 sub-provincial-level districts.” However, neither his claim that Afghanistan had 317 districts nor the identification of 318 by names and subsequent successful extraction of information on 316 districts in the current study match any official listings of districts. It seems that, in practice, the former regime lacked an agreed upon, formal list of its second sub-national administrative division (Ruttig, 2018). Thus, the predictions contained in the present study were based on those administrative units for which information was successfully obtained—34 provinces and 316 districts.

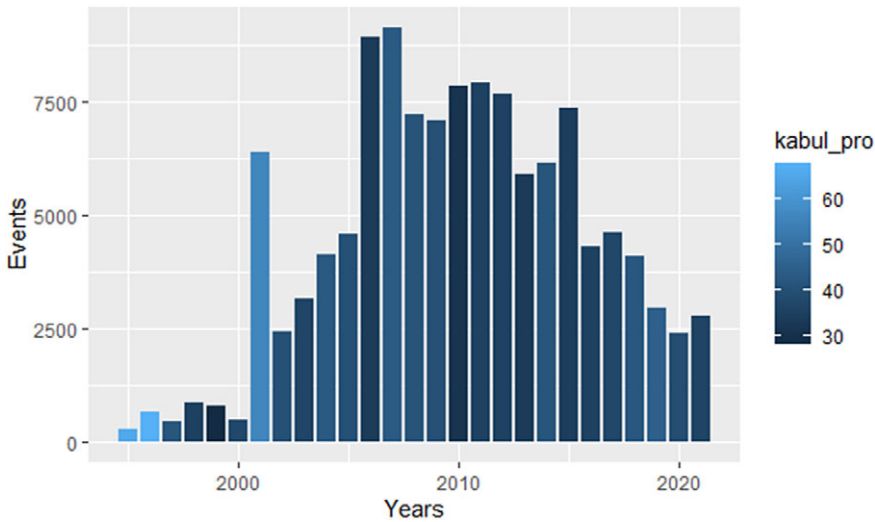


Figure 1. Number of violent events per year, 1995–August 2021. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

The resulting dataset contains 120,299 observations of material conflict for the period beginning in January 1995 and ending in August 2021.¹⁵ In order to provide some insight into the dataset, [Figure 1](#) presents the total number of cases per year, that is, at the country level. The variable `Kabul_pro` presents the proportion of total events in Afghanistan that occurred in the district of Kabul City.

As shown in [Figure 1](#), the height of each bar signifies the total number of quad four events for Afghanistan each year. The color of the bars shows the proportion of these events that took place in the district of Kabul City, where lighter colors correspond to greater proportion. Overall, there are relatively few events up to 2001, which saw a significant increase following the American invasion and subsequent occupation of Afghanistan. Moreover, the events from 1995, 1996, and 2001 were particularly concentrated in Kabul.

In order to verify that the American invasion of Afghanistan had indeed constituted a turning point in the number of violent events, [Figure 2](#) presents the number of violent events per month for 2001.

[Figure 2](#) indicates that the invasion in October 2001 was indeed a turning point with a major increase in the number of violent events from September onwards. [Figure 1](#) further indicates that the number of events in the following years is influenced by the turbulent years of American involvement in Afghanistan. Moreover, the smaller number of events prior to 2000 was probably also influenced by the relative lack of media attention to the developments in Afghanistan before the 9/11 attacks. To some degree, a similar pattern has also impacted the number of events reported in later years, when the attention of world media turned elsewhere.¹⁶ The diagram clearly shows that a relatively high proportion of the total violent events reported in Afghanistan took place in one district—Kabul City. This could be explained both by the Taliban and Islamic State’s strategic decision to concentrate efforts on Kabul in order to destabilize the

¹⁵ The following problem raised by Blair and Sambanis (2020, p. 1910, footnote 8) could potentially apply to the data used for this timeframe: “We opt not to use earlier Integrated Crisis Early Warning System (ICEWS) data based on advice from the Political Instability Task Force (PITF) as the ICEWS actor and event dictionaries were overhauled in 2000.” In addition, although the ICEWS system seems to use some internal filtering mechanisms to avoid or decrease duplication of stories, no clear information exists on this issue. One possible relevant control measure is the “One-A-Day” filter for post-processing suggested by Schrodtt and Analytics (2015). The author is grateful to Andreas Beger for bringing it to his attention.

¹⁶ The ICEWS relies on Factiva as a data source. The author was not able to find information on the latter with regard to which media outlets, that is, local versus international, were used to generate the data on Afghanistan. Had they relied on local media outlets, a different picture of events may have emerged.

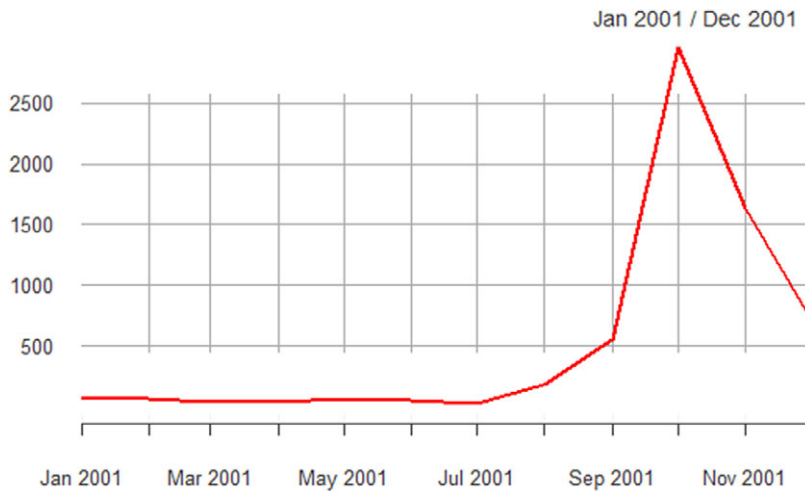


Figure 2. Events per month in Afghanistan in 2001. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

American-backed regime and the prevalence of international media in the city. The latter has since eased its reporting on the attacks that took place in the city (Kamran Bokhari, personal communication, August 17, 2021).

Out of the total number of violent events, the high proportion attributed to Kabul is an example of the inherent bias in ‘big data’ repositories mentioned above. However, one could also argue that it mostly resulted from the aggregation of data at higher levels. An alternative visualization of the data demonstrates the benefits offered by the georeferenced character of ICEWS data, providing the opportunity to present the spatial distribution of the annual number of material conflict events.

Figure 3 does so at the district level.

Figure 3 shows that, until 2001, there were no data available on the vast majority of districts. The districts for which data were available show a yearly total of several dozens of events. The sole exception is the district of Kabul that includes the Afghan capital. It shows an annual total of over 100 material conflict events even during these relatively calm years. Please note that the visualization method used has an inherent weakness: the legend of the heatmap is automatically adapted to the scale of events in each individual year. Therefore, a casual inspection may miss the differences in scales of number of events between years.

From 2001 to 2017, the number of districts for which no data were available dropped considerably. After that, the trend seems to begin reversing. The lack of data could be interpreted either as evidence of the absence of attacks resulting in no reporting as there was nothing to report, an indirect result of the gradual advancing of the Taliban that prevented media reports, or a failure of local and international media to report from many regions about the ongoing violence. While the first option may seem counter-intuitive, it could result from the Taliban’s and Islamic State’s strategic decision to focus their attacks on certain regions they deemed crucial for destabilizing the American backed-regime (Kamran Bokhari, personal communication, 17 August 2021), that is, Kabul.

Figure 4 presents the aggregation at the next level of administrative unit—the province level. The visualization at the province level is very similar to that of the district level. The aggregation also led to very few provinces for which no data were available. These appeared more frequently in the early years and cease to appear after 2006. Similarly to the district level diagram, here too Kabul province stands out as the hottest spot, with Kandahar province as another notable location after 2001.

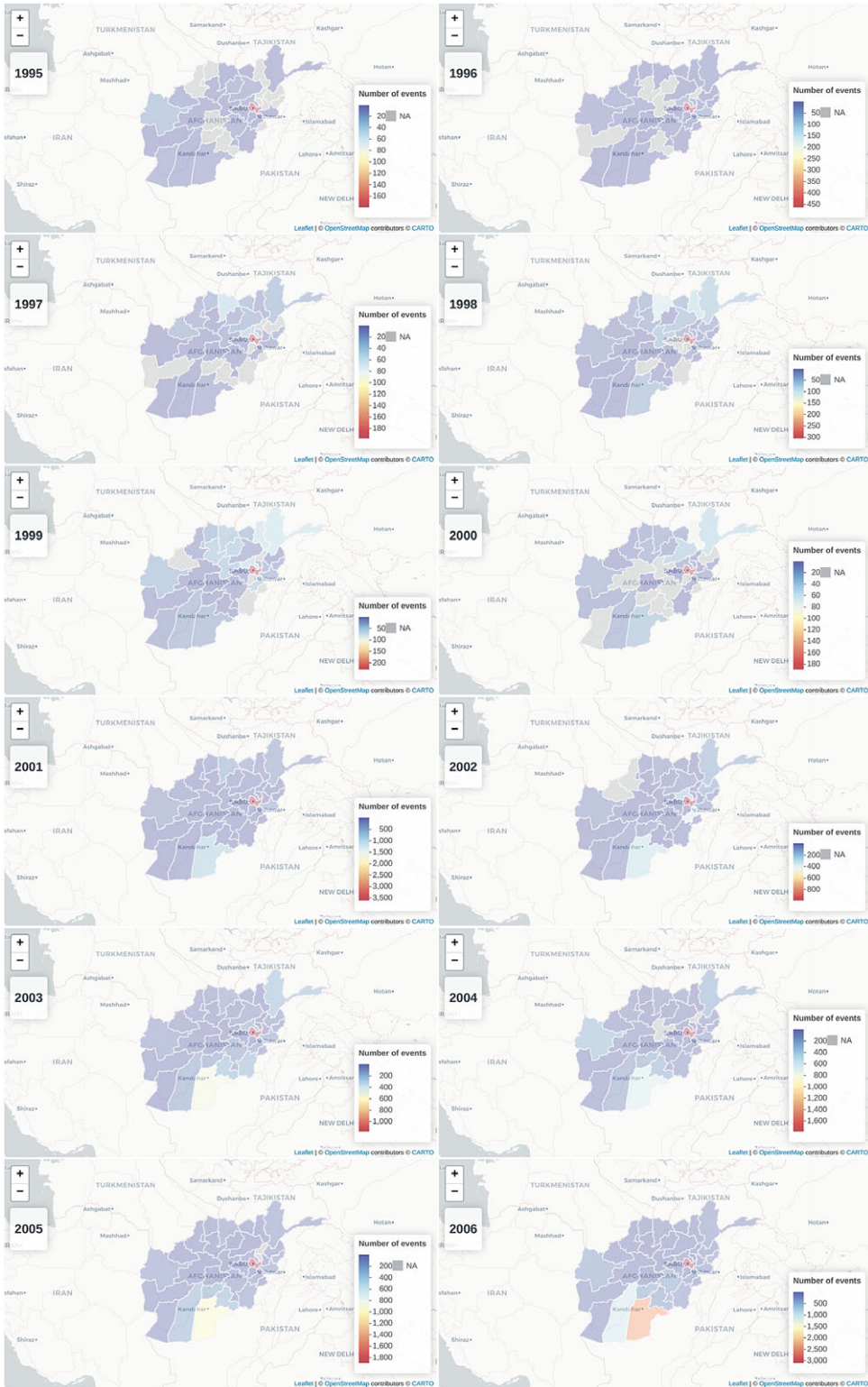


Figure 3. Annual number of material conflict events in districts, 1995–2020.¹⁷ Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

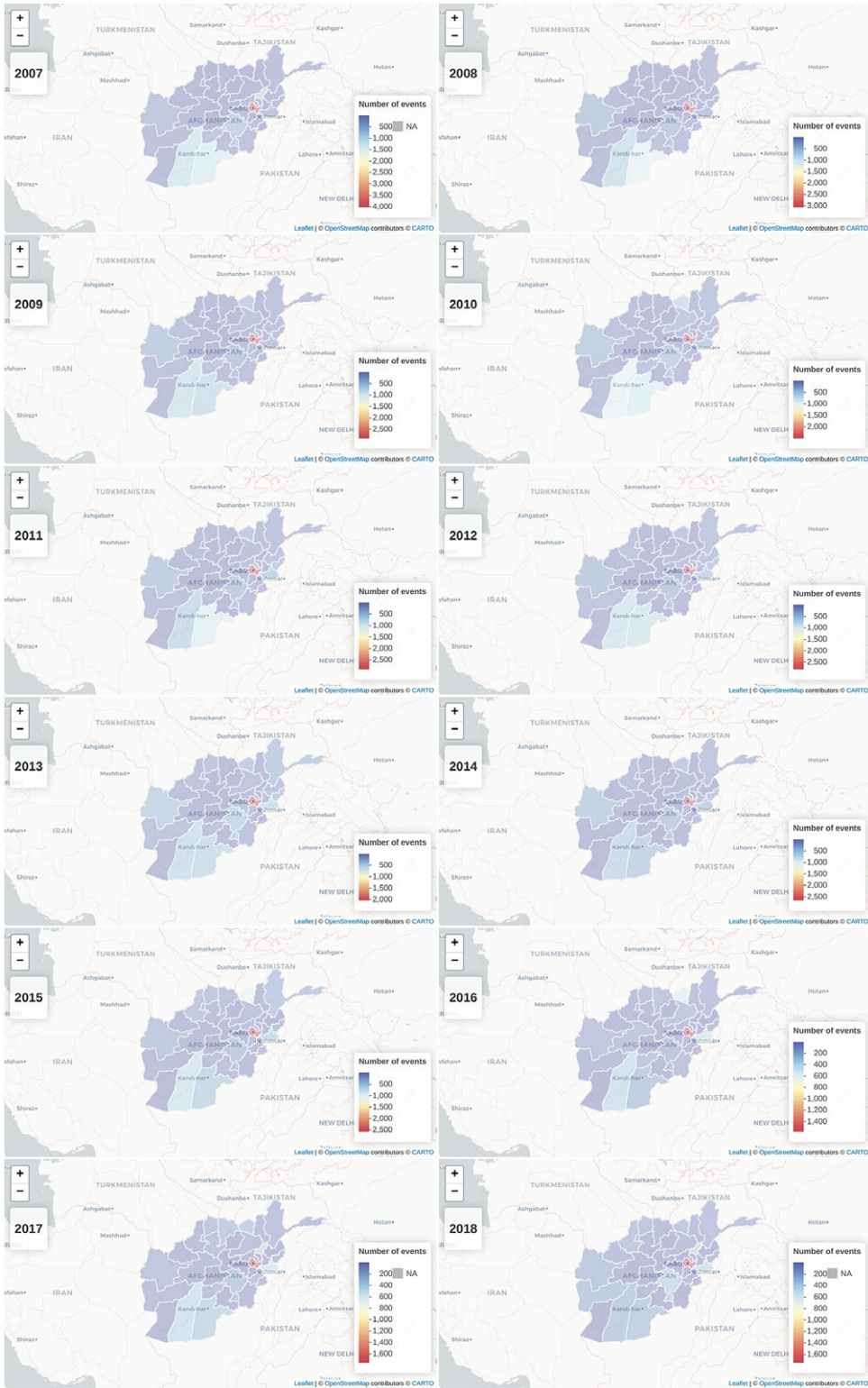


Figure 3. Continued.

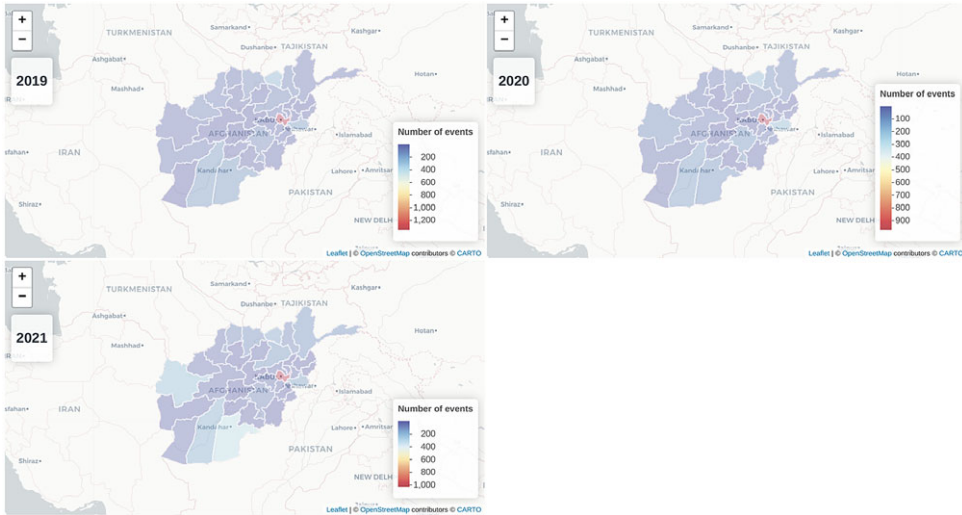


Figure 3. Continued.

Description of the proposed methods

Yonamine (2013, p. 1) pursued to create a policy-relevant forecasting tool by relying on GDELT, which was released that same year. At that time, GDELT was the only dataset that met the five attributes he defined for “building nuanced predictions on a global scale: broad spatial coverage; density; geo-coding; accuracy; and future availability in real-time” (Yonamine, 2013, p. 2). Moreover, GDELT also offered sub-state geo-coding that fit Yonamine’s (2013, p. 3) aim of presenting “the first study to ever use open-source, machine-coded event data to build forecasts of political violence at a sub-state level of geospatial aggregation.”

Due to resource limitations, he focused on “forecasting conflict in sub-state geospatial units in a single country” (Yonamine, 2013, p. 3). Afghanistan was chosen for two reasons: it offered a rich and varied number of relevant events over a long time period; and a precedent that demonstrated the ability to implement a similar approach to study Afghanistan (Zammit-Mangion et al., 2012).¹⁹ In practice, he built predictions 1 month in advance, with district-month ($N = 317$), province-month ($N = 32$), and country-month ($N = 1$) as units of analysis.²⁰ These units reflect an ascending order of administrative units, with provinces aggregating districts and the country level aggregating provinces. Therefore, using these units of analysis provided Yonamine (2013, p. 3) with “a rudimentary test of the effects of geo-spatial aggregation on forecast accuracy.”

¹⁷ The use of data aggregation means that a district for which there was data available, even for just 1 month in a given calendar year, will be counted as non-NA. Therefore, districts for which no data were available—represented in the heatmap as grey—are counted as NA throughout the year.

¹⁸ The use of data aggregation means that a province for which there were data available, even for just 1 month in a given calendar year, will be counted as non-NA. Therefore, provinces for which no data were available—represented in the heatmap as grey—are counted as NA throughout the year.

¹⁹ Despite the abovementioned similarity in approaches between Zammit-Mangion et al. (2012) and Yonamine (2013), there were notable differences between their articles, backing the latter’s claim to present a first of its kind. The former used leaked data from Wikileaks to construct an in-sample training set that builds predictions for the year following the last one for which leaked data were available. Therefore, the forecast accuracy of the out-of-sample test set had to be assessed using another dataset. As Yonamine (2013, pp. 5-6) had indicated, the Wikileaks dataset was neither updated nor did it provide coverage beyond Iraq and Afghanistan. As a result, this dataset did not meet his requirements of a dataset nor his article’s declared aim.

²⁰ Claiming that aggregation of events at the monthly level is a common practice in conflict studies literature, Yonamine (2013, p. 7) justified his choice by stating that: “I choose to use the month as my level of temporal aggregation because this provides sufficient variation throughout the time-series data while reducing the level of noise that is present at daily or weekly levels.” This practice was therefore kept in the current study as part of the replication.

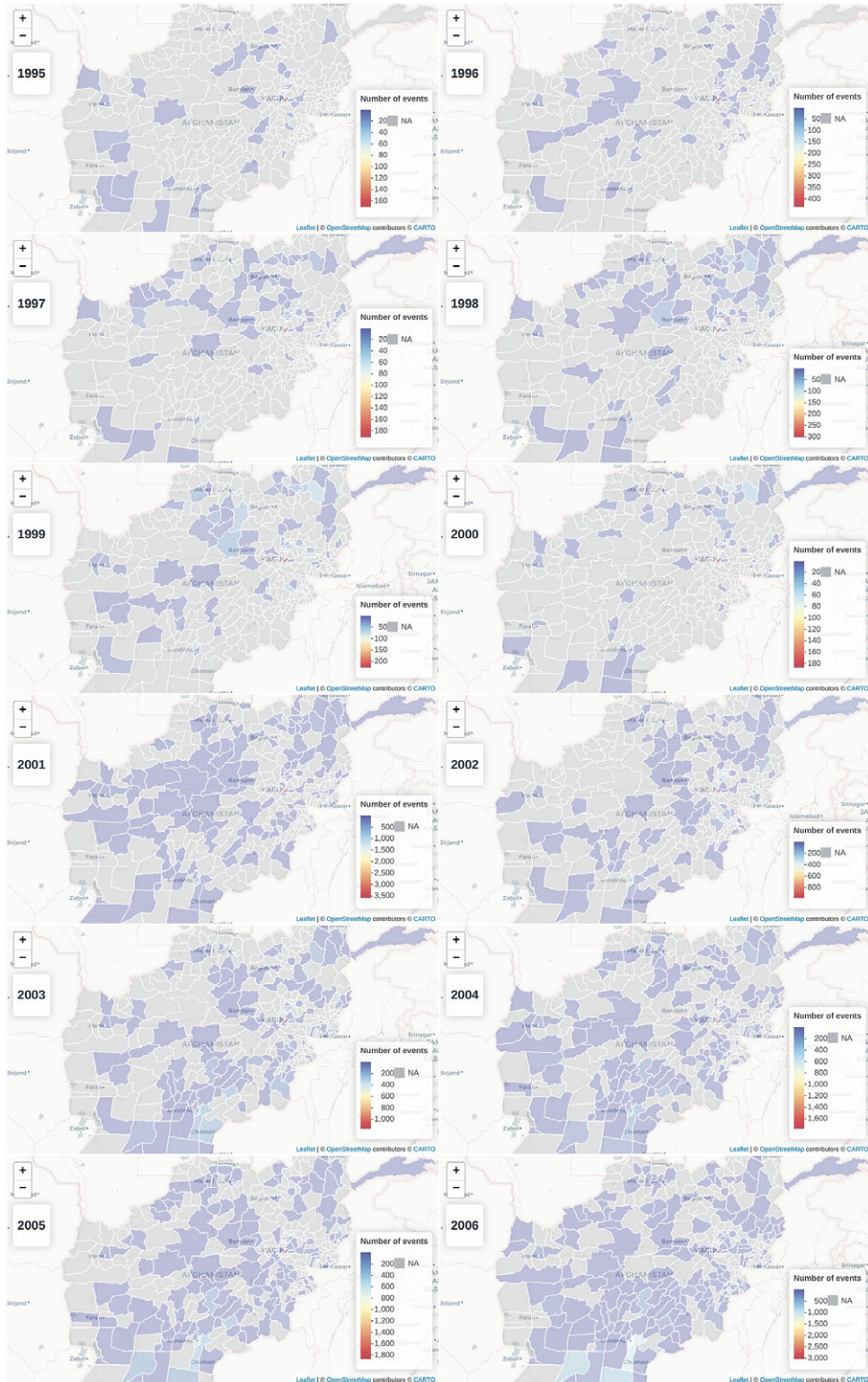


Figure 4. Annual number of material conflict events in provinces, 1995–2020.¹⁸ Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

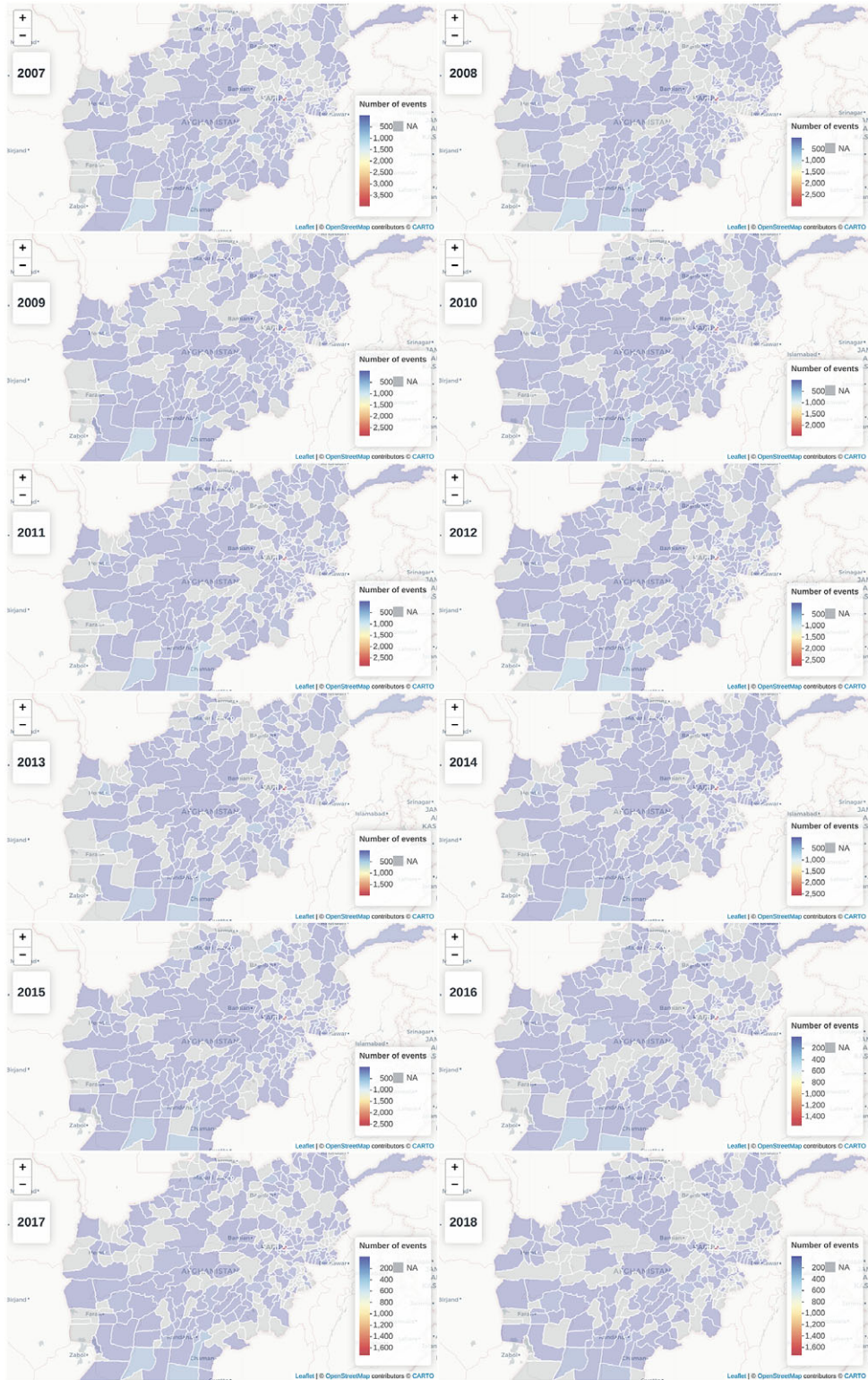


Figure 4. Continued.

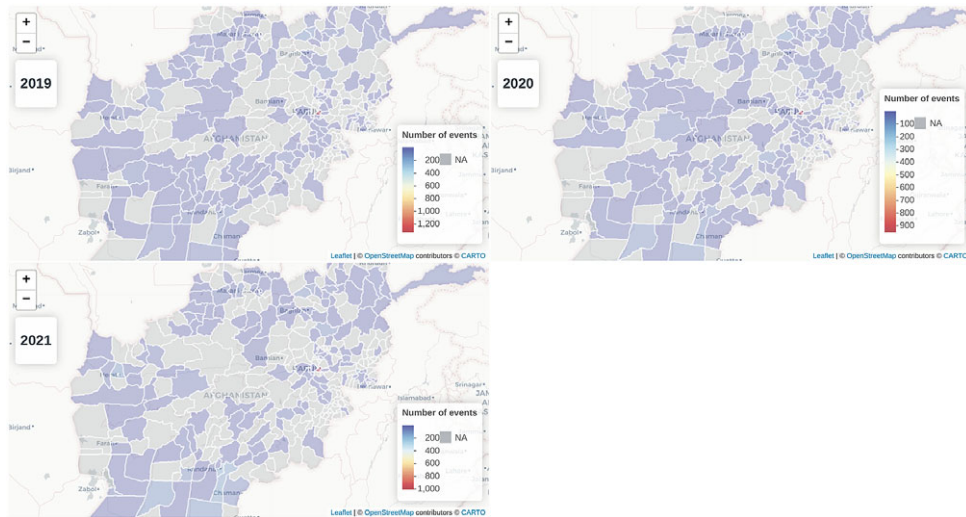


Figure 4. Continued.

The core of his analysis²¹ is a comparison of “forecasts of levels of material conflict one-month-in advance” using the ARFIMA model and a naïve model. The latter assumes that the current real-world level of violence at a given location is identical to that of the previous month. The former model, which is explained below, outperformed the naïve model in terms of forecasting accuracy; however, its performance relative to the naïve model decreased as geographical aggregation increased (Yonamine, 2013, p. 3).²² This validation approach is aligned with the advice provided by Cederman and Weidmann (2017, p. 476) for assessing the added value of a given approach: “...analysts need to do a better job comparing their forecasts from complex prediction machinery to simple baseline models. In its purest form, such a baseline model simply predicts no change from the past.”

Choosing the ARFIMA model enabled Yonamine to use “univariate data comprised solely of the counts of material conflict events” (Yonamine, 2013, p. 7). Other data structures could not be created as the long timeframe and level of analysis on Afghani districts made it impossible to find appropriate exogenous variables to combine with the GDELT data in order to predict conflict levels (Yonamine, 2013, p. 8). Yonamine (2013, p. 1) therefore focused on measuring the accuracy of the models’ predictions, as he believed that to be the best way to assess whether a model reflects “a real-world data generating process, or simply fitting noise.”

Yonamine’s (2013) work may be seen as an early example of both the “predictive turn” in literature (Ward et al., 2010) and the growing focus on localized predictions. His emphasis on forecasting accuracy is also grounded in his interest to present a useful prediction tool for the variety of actors involved in humanitarian and conflict resolving work (Zammit-Mangion et al., 2012). Yonamine’s (2013, p. 9) decision to employ ARFIMA for the first time in order to predict political conflict drew upon the experience in other fields that demonstrated the model’s “ability to generate more accurate and consistent forecasts than other time-series models.”

The ARFIMA is a commonly used method to model a time series with long memory (Reisen et al., 2001, p. 3; Liu et al., 2017, p. 5). A useful way of understanding the meaning of “long memory” in this context is through its opposite—a Markov Chain: “Rather than Markov processes where the current state

²¹ Yonamine (2013, pp. 15–18) also conducted an analysis of what he defined as “two logical extensions to the univariate ARFIMA model” (Yonamine, 2013, p. 3) whereby he included additional variables and added data on drug prices to the ARFIMA model. These are not replicated in the current article due to time and resource limitations.

²² The selection of the naïve model was based on the common finding that: “[A]s forecasters have long recognized, simple models often outperform complex ones in out-of-sample tests” (Blair and Sambanis (2020, p. 1903).

of a system is enough to determine its immediate future, the fractional Gaussian noise model requires information about the complete past history of the system” (Graves et al., 2017, p. 437). The ARFIMA process is based on three parameters: p , q , and d . While p stands for the number of lags in the autoregressive part to be modeled and q for the moving average lags, d indicates the degree of integration of the time series and can take on any value between 0 (stationary) and 1 (integrated) (Box-Steffensmeier and Tomlinson, 2000, p. 64). The ARFIMA model may be represented by the following equation:

$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t$$

where $\Phi(B)$ and $\Theta(B)$ represent the autoregressive and moving average components with B as back-shift operator (i.e., lag) such that $BX_t = X_{t-1}$. $(1-B)$ is the difference operator ∇^d equal to $(1-B)^d$, and ε_t an error term that is normally distributed with $E(\varepsilon_t) = 0$ and variance σ^2 (Box-Steffensmeier and Tomlinson, 2000, p. 64; Reisen et al., 2001, p. 3).

Yonamine’s (2013) analytical strategy consisted of the following steps. First, he trained the model on an in-sample training set that included all the data between February 2001 and April 2008. Second, he ran an out-of-sample prediction for May 2008. Third, he added the prediction for May 2008 to the in-sample training set. Fourth, he trained the model again on the new in-sample training set (i.e., containing now data between February 2001 and May 2008). Fifth, he ran a prediction for June 2008. Next, he repeated the above-mentioned steps “until a final prediction is made for April 2012... using a model trained on February 2001 through March 2012” (Yonamine, 2013, p. 11). The present article replicated the general analytical strategy of Yonamine (2013) rather than its exact steps. This replication was based on a nested loop, containing two for-loops (one for calculating ARFIMA predictions and the other—naïve predictions), which went through the data files of violent events in Afghanistan on the district, province, and country levels.

The predictions for both models were calculated using the “forecast” R package’s ARFIMA and naïve functions, respectively (Hyndman and Khandakar, 2008). The advantage of this ARFIMA function is that it estimates and sets the model’s parameters automatically. However, by using this function, the current article differed from Yonamine’s (2013) article, which used the “ARFIMA” R package (Veenstra, 2012). In line with the original article, a month-district/province/country unit that was not assigned a value of a conflict event was associated with a “0” (Yonamine, 2013, p. 7). For the first two of the three administrative unit levels—district and province—the following steps were taken. For the third level of administrative unit, that is, country, the workflow was somewhat different as there was only one data file.

The comparison that Yonamine (2013) conducted between the forecasting accuracy of ARFIMA and the naïve model required a suitable error or difference statistic, a measure that “can be used to compare model-produced estimates with independent, reliable observations” (Willmott and Matsuura, 2005, p. 79). His error statistic of choice was the mean absolute error (MAE), which is a common measure in literature (Chai and Draxler, 2014). MAE is a dimensioned measure of average performance error in the sense that “it expresses average model-prediction error in the units of the variable of interest” (Willmott and Matsuura, 2005, p. 79). It is also used to reflect average difference rather than error when “no set of estimates is known to be the most reliable” (Willmott and Matsuura, 2005, p. 79). Generally speaking, MAE sums up the absolute values of errors in a model and divides the results by n (Willmott and Matsuura, 2005, p. 80).

$$\text{MAE} = \left[n^{-1} \sum_{i=1}^n |e_i| \right]$$

Although the use of MAE is widespread, Yonamine (2013) implemented it in a unique manner due to his need to compare the results of multiple univariate time-series analyses with regard to three different units of analysis (i.e., month-district, month-province, month-country):

“For each of the 48 months that iteratively serve as the out-of-sample test, I calculate the error rates for the naïve model (naïve_error) and the ARFIMA model (arfima_error rate), which reflect the MAE across the N cross-sections ($N = 317$ for the district-month model, $N = 32$ for the province-month model and $N = 1$ for the country-month model).” (Yonamine, 2013, pp. 13–14)

As part of the replication pursued, the present study adopted the use of MAE as an error statistic, employing it in the same fashion as Yonamine (2013).

Results and discussion

The first part of the current study replicated Yonamine’s (2013) analytical strategy using ICEWS data. Yonamine (2013, p. 3) argued that the ARFIMA model’s outperformance of the naïve model will also be reflected in a smaller MAE size. Figure 5 presents the current study’s MAE results of ARFIMA and naïve models at the district level.

Figure 5 confirmed Yonamine’s (2013) assumption—the ARFIMA’s MAE, presented in red, is overall lower than those of the naïve’s MAE, presented in green. The five intersections of the two lines visualized the few exceptions: August 2008, January 2009, April 2010, March 2011, and September 2011. Next, Figure 6 presents the comparable results at the province level.

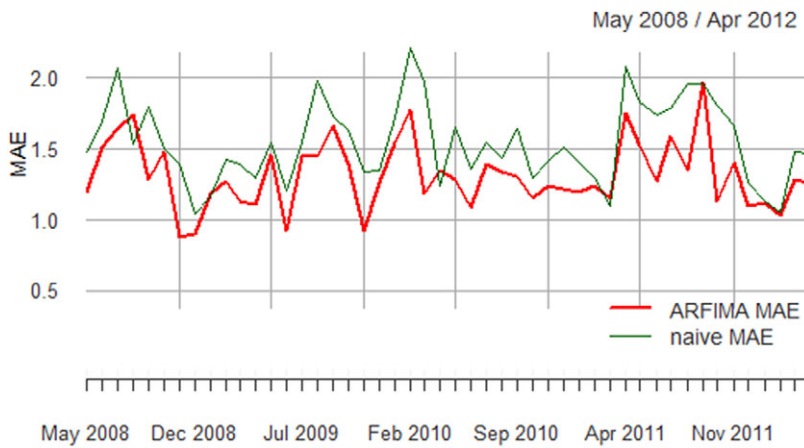


Figure 5. Assessing accuracy at the district level.²³ Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

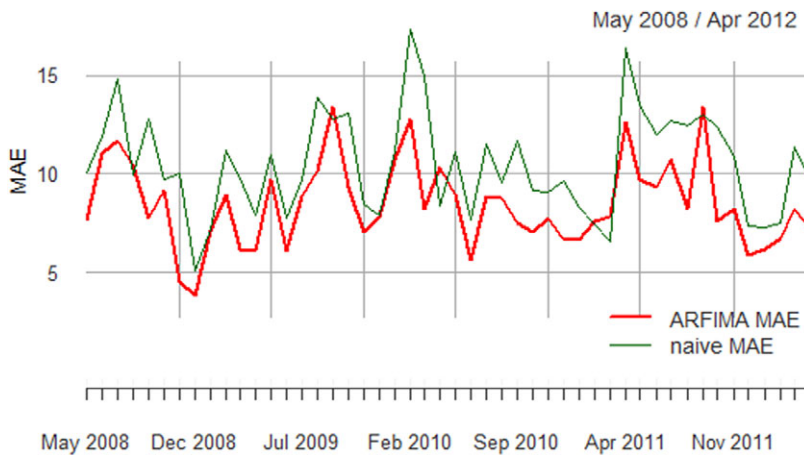


Figure 6. Assessing accuracy at the province level.²⁴ Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

²³ The title for this figure is an adaptation of Yonamine’s (2013, p. 30) title for his Table 1.

²⁴ The title for this figure is an adaptation of Yonamine’s (2013, p. 31) title for his Table 2.

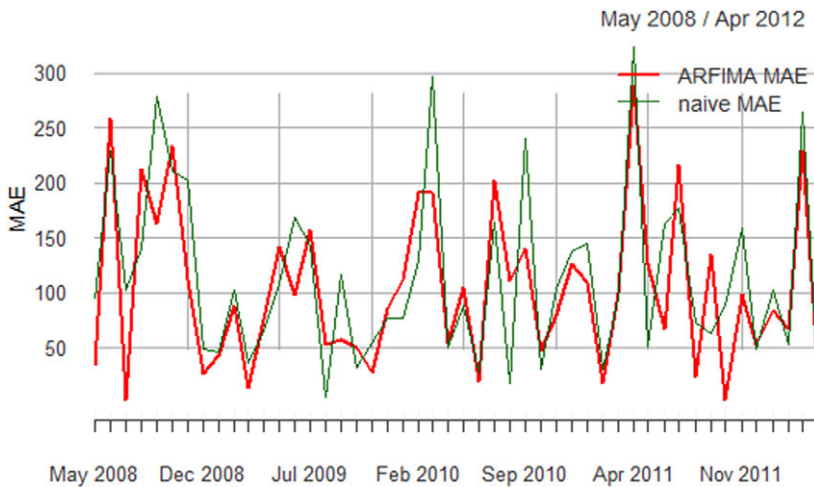


Figure 7. Assessing accuracy at the country level.²⁶ Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

As shown, the ARFIMA model's MAE results (i.e., red line) are lower than those of the naïve model (i.e., green line) at the province level too. Six intersections, indicating the months at which the naïve results are greater than those of ARFIMA, may be observed: August 2008, September 2009, April 2010, February 2011, March 2011, and September 2011. Figures 5 and 6 thus provide support for Yonamine's (2013) assumptions that the ARFIMA MAE would be smaller than the naïve's, and that the former's performance relative to the latter's will deteriorate the higher the level of aggregation is.²⁵ However, the rate of deterioration observed was miniscule compared to the original study. Next, Figure 7 provides the data at the country level.

At the country level, presented in Figure 7, the rate at which the ARFIMA model outperformed the naïve model deteriorated substantially. The 48 month-country units were divided almost equally between the ARFIMA and naïve models—25 and 23 months, respectively. The aggregation of the data at the country level also provided an opportunity to compare the observed (i.e., real events) and the predictions by both ARFIMA and naïve models for the TS period—May 2008 to April 2012. These results are presented as a line graph in Figure 8.

As expected, the predictions by the naïve model, marked here by the dark blue line, constituted a delayed version of the actual data. In turn, the ARFIMA predictions successfully captured the general trends from early 2009 to late 2010, and again from February 2012. As the country level presented the worst performance by ARFIMA compared to the naïve model, an additional visualization was provided for the district level, that is, the level at which the ARFIMA model performed best. Therefore, Figure 9 presents a comparison between the observed and predicted results for the Kabul district—the one with the highest number of events per year.

As can be seen, the behavior of the naïve model in Figure 9 is similar to the one displayed in Figure 8, that is, at the country level. In contrast, the performance of the ARFIMA model was much better at this level than at the country level. In other words, for the most part, it successfully predicted the trend in the observed data. This finding may be seen as support not only for Yonamine's (2013) assumptions but also for the argument in the forecasting literature whereby predictions may be most accurate for localized predictions.

²⁵ That is, district level < province level < country level.

²⁶ The title for this figure is an adaptation of Yonamine's (2013, p. 32) title for his Table 3.

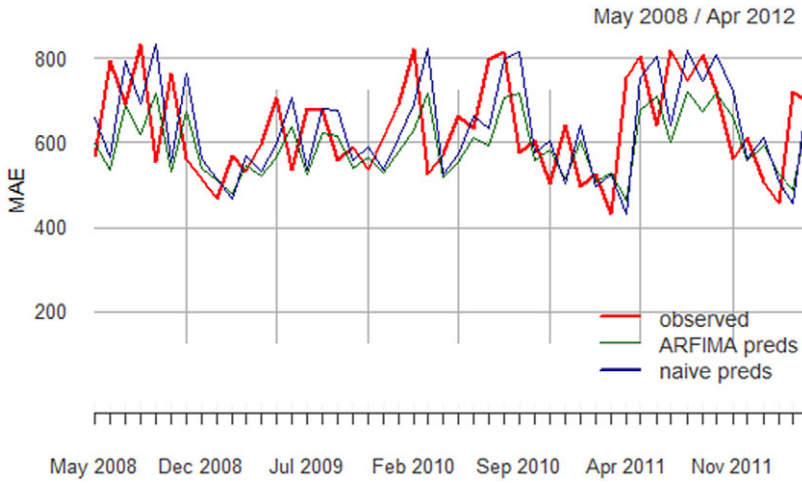


Figure 8. Comparison of observed, ARFIMA predictions, and naïve predictions at the country level. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

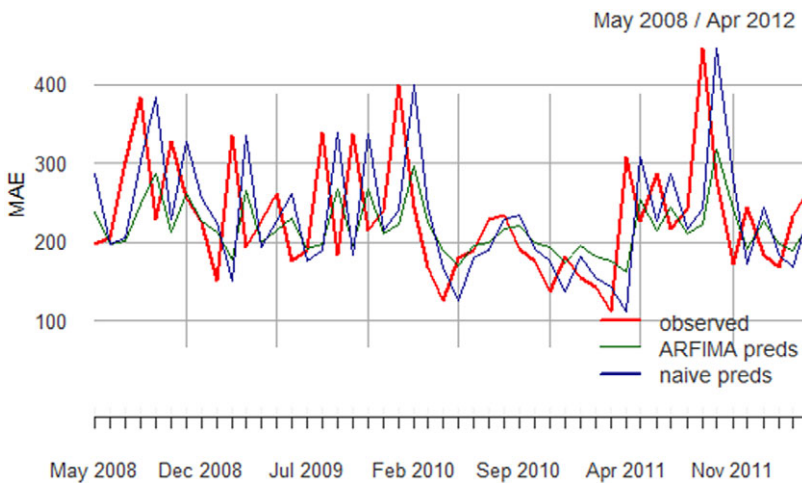


Figure 9. A comparison of observed, ARFIMA predictions, and naïve predictions for the Kabul City district. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

In the next section, the results of the present study are compared to those of Yonamine (2013) to provide a comparison of the use of a shared analytical strategy with two main data sources—GDELT and ICEWS. The comparison is presented in Table 1.

Table 1 shows that, at the district level, Yonamine’s (2013) results were clearly better, with 47 out of 48 months for which the ARFIMA model outperformed the naïve model. In contrast, for the current study, the figure was just 43 out of 48 months. The situation is reversed at the province level, with 42 out of 48 months for the current study, and 40 out of 48 months for Yonamine (2013). This was contrasted once more at the country level, at which the figures for Yonamine (2013) and the current study were 30 out of 48 and 25 out of 48, respectively. An additional comparison of the MAE results obtained by Yonamine (2013) and the present study is provided in Figures 10–12 for the district, province, and country levels, respectively.

Table 1. A comparison of the MAE results of Yonamine (2013) and the current study

	Yonamine (2013)				Current study			
	AE < NE		Sum of MAE		AE < NE		Sum of MAE	
	TRUE	FALSE	AE	NE	TRUE	FALSE	AE	NE
District	47	1	129.76	155.07	43	5	63.13	74.21
Province	40	8	1,004.56	1,151.50	42	6	406.62	505.5
Country	30	18	16,439	16,612	25	23	4,982	5,497

Data source: Yonamine (2013) and Lockheed Martin Advanced Technology Laboratories (ATL) (2021).
 Abbreviation: AE, ARFIMA error; MAE, mean absolute error; NE, naïve error.

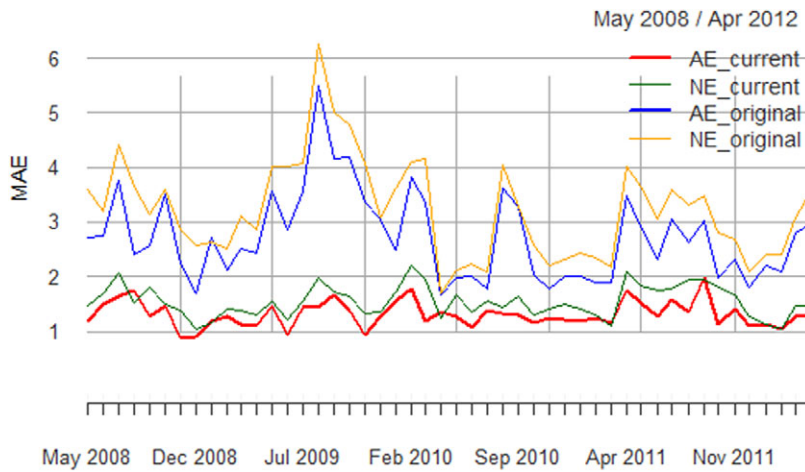


Figure 10. A comparison of accuracy assessments between Yonamine (2013) and the current study at the district level. Legend: original, Yonamine (2013); current, current study; AE, ARFIMA error; NE, naïve error. Data source: Yonamine (2013) and Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

Both Yonamine’s (2013) predictions and those made in the current study seem to follow similar trends. However, the former set’s trends are more similar to each other, with two exceptions: May 2009–October 2009 and November 2010–April 2011. What is also noticeable is the difference in scale between the two sets. Yonamine’s (2013) values for ARFIMA and naïve MAE range from 1.66 to 5.5 and 1.7 to 6.25, respectively. In contrast, the values for the current study are 0.88–1.97 and 1.04–2.21, respectively. The MAE results for the province level are provided in Figure 11.

The abovementioned difference in scale was more noticeable at the province level. For Yonamine (2013), the values of MAE ranged from 6.91 to 93.09 and 10.38 to 91.66 for the ARFIMA model and naïve model respectively. In contrast, the values for the current study ranged from 3.08 to 13.43 and 5.08 to 17.32, respectively. The difference probably originated in the different data sources, that is, GDELT and ICEWS, respectively. Moreover, while both sets of predictions seem to follow the same respective trends, Yonamine’s ARFIMA and naïve predictions tend to be more in sync.

Figure 12 shows both sets of the ARFIMA and naïve models as less in sync than at the previous levels. However, the original study’s sets were more closely related despite involving greater divergence and fluctuations. The increased scale of difference between the original and current studies probably resulted from the further aggregation of data. An additional assessment of the prediction accuracy was made by

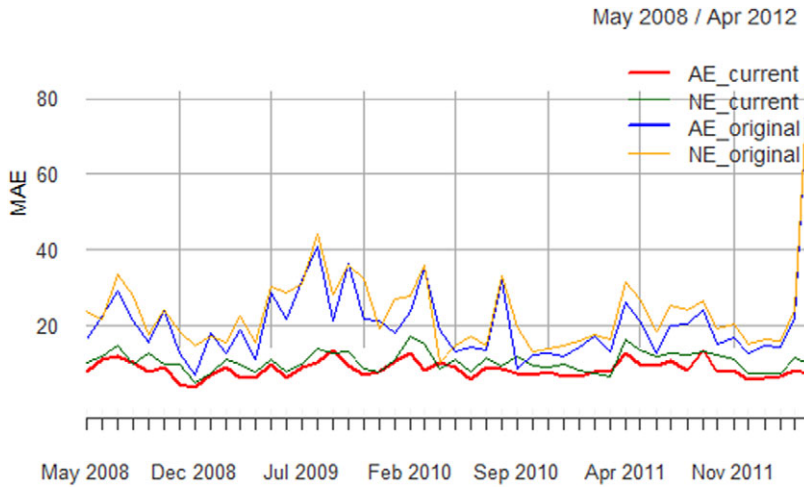


Figure 11. A comparison of accuracy assessments between Yonamine (2013) and the current study at the province level. Legend: original, Yonamine (2013); current, current study; AE, ARFIMA error; NE, naïve error. Data source: Yonamine (2013) and Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

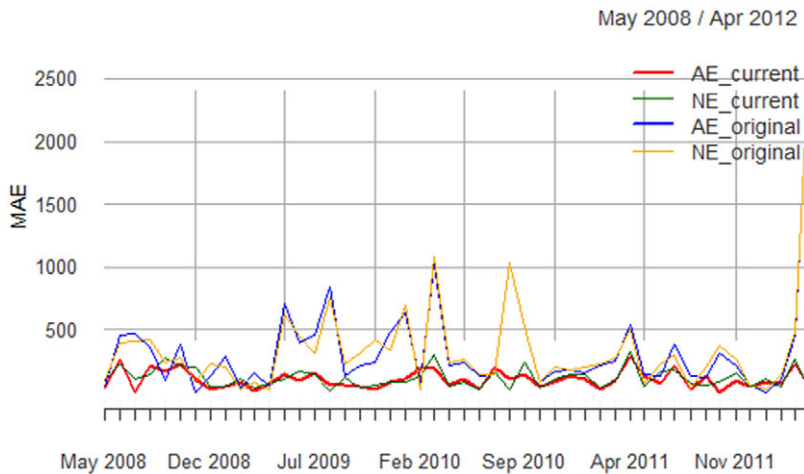


Figure 12. A comparison of accuracy assessments between Yonamine (2013) and the current study at the country level. Legend: original, Yonamine (2013); current, current study; AE, ARFIMA error; NE, Naïve error. Data source: Yonamine (2013) and Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

evaluating the extent to which the ARFIMA model reduced the MAE size compared to that of the naïve model. This measure was calculated using the following equation:

$$100 - \frac{AE \times 100}{NE}$$

where AE stands for “ARFIMA error” and NE stands for “naïve error.” The results are summarized in Table 2.

Table 2. *A comparison of the reduction in MAE size*

	Original study (%)	Current study (%)
District	16.33	14.94
Province	12.77	19.57
Country	1.05	9.37

Data source: Yonamine (2013) and Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

The results presented in Table 2 support the findings shown in Figures 10–12. The proportion of the ARFIMA model’s outperformance of the naïve model was better at the district level in the original study. In the original study, the ARFIMA model reached an MAE size that was 16.33% smaller than that of the naïve model. The comparable figure for the current study at that administrative level was 14.94%. The situation was reversed at the province level, for which the current study reached a better result than that of the original study—19.57% compared to 12.77% in the original study. For both studies, the AFRIMA model’s performance was the worst at the country level, at which the proportion of reduction in MAE size was the smallest for both studies—1.05% in the original study and 9.37% in the current one.

The results presented so far called for a qualification of Yonamine’s (2013) assumption that the ARFIMA model’s performance comparing to that of the naïve model deteriorates with each level of aggregation. This seemed to be true for the highest level of aggregation but not necessarily for the interim one in the current study. Moreover, the rate of deterioration in the ARFIMA model’s performance compared to that of the naïve model was significantly smaller on all three levels of administrative units than in the original study.

Furthermore, with regard to the difference in scales mentioned above, it should be noted that, because the MAE is based on error size, which, in turn, is determined by the subtraction of the observed from the predicted in absolute values, this finding may suggest that Yonamine’s (2013) study contained a higher number of cases per month-district units. That would be in line with a common observation in the literature that his data source, GDEL, may contain a considerable number of duplicated events and thus suffers from a high number of false positives. This may also help to explain the small difference between the results obtained by the current study at the district and province levels—reflected in the number of administrative unit month for which the naïve model outperformed the ARFIMA model—which were very close: 43 versus 5 and 42 versus 6, respectively. The similarity between these results was also reflected in the much smaller difference in scale of the results compared to the original study.

The previous two sections of the analysis helped to assess the usability of Yonamine’s (2013) analytical strategy with ICEWS data. Having done so, the approach was employed on the entire ICEWS dataset timeframe—January 1995 through August 2021. The significantly longer period added yet another dimension for evaluating Yonamine’s (2013) approach when using the ICEWS data; however, it required the identification of a suitable cutting point for the data, dividing it into an ITS and TS. Three cutting points for sets of ARFIMA and naïve predictions at the country level were therefore determined: $\frac{2}{3}$ ITS— $\frac{1}{3}$ TS, $\frac{1}{2}$ ITS— $\frac{1}{2}$ (TS), and $\frac{3}{4}$ ITS— $\frac{1}{4}$ TS. The country level was selected under the assumption that it was at that level that the ARFIMA model’s performance is the worst. Therefore, any model that performed better at this level would do even better at the other levels. The results obtained by this step are presented in Table 3.

Due to the very close “accuracy rate” measurements for the first two upper sets, a further comparison was conducted at the district level. As mentioned above, this level is assumed to yield the best results in terms of the ARFIMA model outperforming the naïve one. The results of the comparison are presented in Table 4.

Table 3. A comparison of predictions at the country level

	AE < NE ²⁷			Accuracy rate ²⁸ (%)	Sum of MAE ²⁹		Reduction in MAE size (%)
	TRUE	FALSE	TOTAL		AE	NE	
ITS:50%, TS:50%	91	69	160	56.87	16,152	17,522	7.82
ITS:66%, TS:33%	62	47	109	56.88	10,794	11,814	8.64
ITS:75%, TS: 25%	44	36	80	55	7,937	8,574	7.43

Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).
 Abbreviations: AE, ARFIMA error; ITS, initial training set; MAE, mean absolute error; NE, naïve error; TS, test set.

Table 4. A comparison of predictions at the district level

	AE < NE ³⁰			Accuracy rate ³³ (%)	Sum of MAE ³²		Reduction in MAE size (%)
	TRUE	FALSE	TOTAL		AE	NE	
ITS:50%, TS:50%	136	24	160	85	165.71	197.69	16.18
ITS:66%, TS:33%	87	22	109	79.81	105.69	122.55	13.76

Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).
 Abbreviations: AE, ARFIMA error; ITS, initial training set; MAE, mean absolute error; NE, naïve error; TS, test set.

The results of this comparison indicate that, at the district level, the performance of the ARFIMA model was better in the upper set of the two models than the lower one for both the accuracy rate and proportion of reduction in MAE size. Therefore, the next step of the analysis was to focus on a comparison of the results of the ARFIMA and naïve models for the division of the dataset into equal sized ITS and TS. The results for all three levels of administrative units are presented in Table 5.

Table 5 contains somewhat surprising results. For the original timeframe, the use of the ARFIMA model on ICEWS data resulted in a contradiction with Yonamine’s (2013) results in terms of the reduction in MAE size. As shown in Table 2, the performance of the ARFIMA model yielded better results at the province level than at the district level, while in the original study, the latter provided the best results. However, using the division of the entire dataset into two equal sized sets, the ARFIMA model’s outperformance of the naïve model met the original study’s expectations as measured by reduction in MAE size: the results at the district level were the best and the outperformance rate deteriorated the higher the level of aggregation. The same can also be seen in Figures 13–15, which show a comparison of the accuracy rate of both ARFIMA and naïve models as measured in MAE across the three administrative units. Figure 13 provides an overview of the results at the district level.

Figure 13 shows the red line, representing the MAE results for the ARFIMA model, as consistently below the green line, which represents the MAE of the naïve model, almost throughout the entire duration of the TS. It provides a visual demonstration of the 85% accuracy rate mentioned in Table 5. Similarly, the not much worse results obtained for the province level are shown in Figure 14.

²⁷ Measured as month units for which AE < NE is either TRUE or FALSE.

²⁸ This parameter was calculated using the following equation: $\frac{AE \times 100}{n}$.

²⁹ This parameter was calculated using the following equation: $100 - \frac{AE \times 100}{NE}$.

³⁰ Measured as month units for which AE < NE is either TRUE or FALSE.

³¹ This parameter was calculated using the following equation: $\frac{AE \times 100}{n}$ with n = number of districts used in the calculation.

³² This parameter was calculated using the following equation: $100 - \frac{AE \times 100}{NE}$.

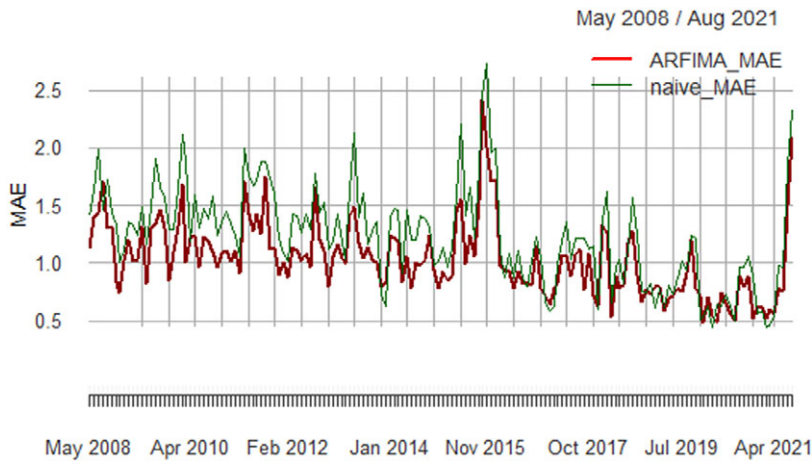


Figure 13. An accuracy assessment with a cutting point at $\frac{1}{2}$ of the full timeframe—district level. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

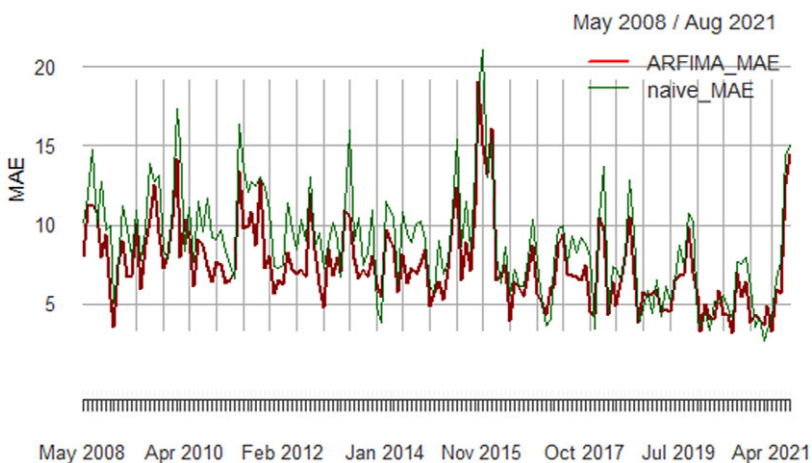


Figure 14. Accuracy assessment with cutting point $\frac{1}{2}$ of the full timeframe—province level. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

Finally, the significant deterioration in the ARFIMA model's outperformance of the naïve model is visually noticeable in Figure 15.

As mentioned in the discussion of the second section, the difference in results between the original and current studies probably originates from their respective reliance upon GDELT and ICEWS, respectively. This also led to the smaller number of cases per administrative unit months characteristic to ICEWS in the current study. Therefore, it is reasonable to assume that the aggregation of the district-level data into the province level in the current study resulted in better performance than displayed at the district level in the original timeframe, before the deterioration at the (too-) aggregated country level. It does, however, seem that the longer duration of the last section provided the ARFIMA model with enough data at the ITS to yield better performance in spite of the smaller number of cases per administrative unit month in the ICEWS data.

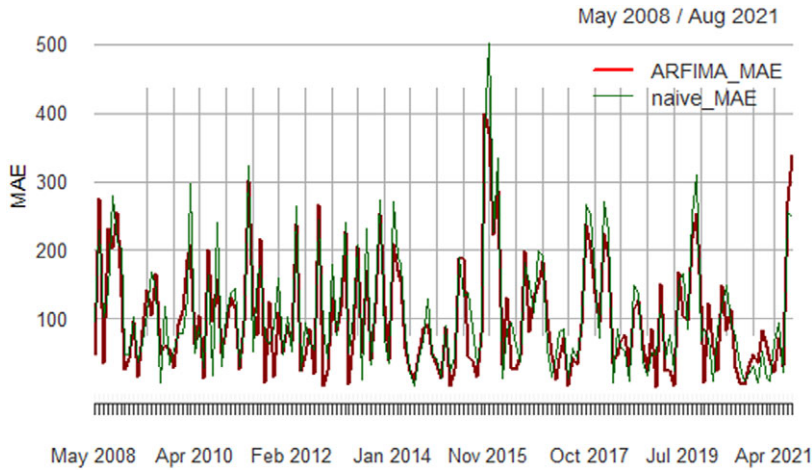


Figure 15. Accuracy assessment with cutting point 1/2 of the full timeframe—country level. Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).

Table 5. A comparison of predictions with a cutting point at 1/2 full timeframe

	AE < NE ³³			Accuracy rate ³⁴ (%)	Sum of MAE ³⁵		Reduction in MAE size (%)
	TRUE	FALSE	TOTAL		AE	NE	
District level	136	24	160	85	165.71	197.69	16.18
Province level	131	29	160	81.87	1,201.11	1,421.20	15.49
Country level	91	69	160	56.87	16,152	17,522	7.82

Data source: Lockheed Martin Advanced Technology Laboratories (ATL) (2021).
 Abbreviation: AE, ARFIMA error; MAE, mean absolute error; NE, naïve error.

Conclusions

The present article sets out to replicate the first study that used automated machine coding event data to predict violent events at the subnational level. The original study can be seen as an early example of the current trend in the literature to advance short-term, localized predictions. After replicating the original study using a more reliable data source—ICEWS rather than GDELT—followed by a comparison of its results with those of the original study, the original study’s analytical strategy was implemented on a much longer timeframe. The latter analysis makes several contributions to the literature. Firstly, it demonstrated that Yonamine’s (2013) analytical strategy is feasible and beneficial for short-term, localized predictions using ICEWS—the current common standard. However, some precautions should be taken.

The ICEWS’s smaller number of events per administrative unit month requires either reliance on longer periods than Yonamine (2013) did or on aggregation at a higher administrative unit level in order to maximize ARFIMA’s performance. The current study, however, also revealed that the deterioration in ARFIMA’s outperformance compared to the naïve model is much smaller at the country level than it is in the original, GDELT-based study. This finding may indicate that country-level predictions using univariate data may be sounder due to such data limitations.

³³ Measured as month units for which AE < NE is either TRUE or FALSE.

³⁴ This parameter is calculated using the following equation: $\frac{AE \times 100}{n}$

³⁵ This parameter is calculated using the following equation: $100 - \frac{AE \times 100}{NE}$

Second, the current study provides support for Yonamine's (2013) original conclusion that univariate time-series data could provide a "good enough" forecasting tool for policy actors. In practice, the implementation of his analytical strategy only requires a reliable, timely data source and enough computing power. Given the current almost weekly release of ICEWS data and ubiquitous availability of either cloud or strong personal machines, both requirements could be easily met. However, Yonamine's (2013) original study did not make replication an easy task. While his personal Github account contained some pieces of code that provided a little insight into his approach, it was never referenced in his article or near completion. Thus, the third contribution of the current study lies in providing detailed code and data that made reproducibility possible.

Fourth, the extensive, successful use of the naïve model in the current study as a benchmark for prediction accuracy, together with relevant visualization to ease interpretation, may be seen as support for the abovementioned call to use such a simple algorithm as a benchmark. In other words, the current study may help to establish this as a best practice. Fifth, a comparison of ARFIMA with other suitable models for univariate time-series data predictions could be advantageous, in particular, for policy actors who look for feasible, "good enough" forecasting tools. In this respect, the popular long short-term memory (LSTM) algorithm may be a desirable comparison (Chen et al., 2020). Finally, the use of MAE as a measurement of error should be compared to other possible types of measurements.

Naturally, further work is required to evaluate these conclusions. One important direction for further study is related to the focus of the current study, due to its replicatory objective, on Afghanistan. Using similar research designs on other countries would help to explore the reliability of the present study's findings. Similarly, there is a need to compare the analysis of the ICEWS data, constituting the backbone of the article, to one derived from another data source. In spite of the author's efforts, no other non-event data source was found to enable such a validation exercise. The data collected by the International NGO Safety Organisation (INSO) concerning incidents in Afghanistan, which is not accessible to the public, may be a relevant data source for this task.³⁶ By way of overcoming this limitation, organizations considering to use the suggested approach for the short-term, localized prediction of violent events may opt to "quantify" their own often rich qualitative data concerning localized violent events.³⁷ This could provide an unparalleled resource—relying on the intimate knowledge of local staff and their relationship with local populations—either to be used for prediction per se or as a validation measure.³⁸

This brings up the crucial issue of specific and general limitations to the adoption of the presented method by policy makers and actors. First and foremost, the 1-month-ahead predictions, demonstrated by Yonamine (2013) and the current article, may not be as directly useful to policymakers as was hoped by the original article. While policy actors are often interested in mid-to-long term forecasting, in-country humanitarian actors are often occupied with even shorter timeframes than 1 month ahead. The use of the suggested method may be best introduced as complementary to existing methods of prediction used by policymakers and humanitarian actors. One area where it may be significantly beneficial is as a "remote monitoring" tool:

"Remote monitoring describes the monitoring of (a) context evolution, (b) implementation of programs and its effects, (c) performance and compliance of partner organisations in areas where physical access to project sites, affected populations and/or partner organisations is restricted or not possible." (Sida and Oakley, 2019, p. 2)

³⁶ The civilian death figures collected by the United Nations Assistance Mission in Afghanistan (UNAMA) could perhaps have provided a rough baseline for comparison. However, they seem to be disseminated only as a periodical textual document rather than a dataset.

³⁷ Constituting a de-facto field work, this raises a complicated set of ethical and moral dilemmas. See Malejacq and Mukhopadhyay (2016).

³⁸ Similar recent calls for mixed-method approaches for data collection and analysis were made in other fields. For an excellent, insightful introduction to the topic see Skarbek (2020).

As long as a “good enough” data source (e.g., ICEWS) is available, the suggested approach in the current article can clearly contribute to the “context evolution” dimension of “remote monitoring.” However, the interpretation of results, that is, predictions, and, no less important, the drivers or mechanisms causing them, requires forecasters to work hand-in-hand with subject matter experts who have intimate familiarity with the region in question. This may be even more important with parsimonious methods, like the one suggested in the present article, which lack additional information that may provide clues concerning these drivers. In regard to the latter point, while growing attention is paid to the spatial dimension of violence in the literature (Rokem et al., 2018), the potential contribution of the current article is highly limited in this regard due to its reliance on univariate data. The lack of additional information limits its ability to contribute to further understanding of the spatial mechanisms or drivers of violence in Afghanistan. However, the spatial data visualization step described above may still help analysts gain insights into the temporal and spatial dynamics of violent events in a given polity.

On a more general level, the way forward to improve conflict prediction will be to more explicitly incorporate insight from theories of armed conflict. Further work on this could build on the recent discussion between Blair and Sambanis (2020; 2021) and Beger et al. (2021). Moreover, this issue and the considerations above touch on the wider question of the relevance and utility of conflict prediction for policy makers and actors. As was discussed in the above paragraphs, it is the firm belief behind the current article that such methods could not only provide accurate predictions but could be highly beneficial to the operations of humanitarian actors in the form of either adjusting “on ground” activity to predicted violence or “distant monitoring.” As was also argued, there is a crucial need for collaboration between the forecasters and subject matter experts in order to give the predictions policy- or action-relevance. This is particularly important with parsimonious statistical models, like ARFIMA, which do not draw upon additional information than the one (or few) variable included. The main advantage of these models—enabling the prediction of violent events in conflict zones where collecting rich contextual information is dangerous and hard—is also what limits their ability to shed light on the why questions.

Understanding this point—the advantage and limitations of univariate models—touches on a fundamental issue that is paramount to the adoption of data science methods by humanitarian actors: the relative lack of statistical literacy and familiarity with data science among humanitarian as well as security and intelligence actors on the one hand and the lack of contextual and social-science knowledge by data scientists on the other.

The former rely and are used to highly contextual, qualitatively produced assessments and prediction reports. These (ideally) draw upon subject matter expertise and social sciences theories, stemming, for example, from development or area studies. Many of them may not be familiar with the statistical learning-based data science methods used for producing quantitative, localized, near-term predictions. A lack of statistical literacy among humanitarian actors could hamper both trust in the analysis and dialogue with the analysts. In the short term, collaboration of subject matter experts (usually qualitatively trained) in the production and interpretation of the quantitatively driven prediction reports is a crucial step to overcome that. Well-written, visualized reports that take the context in which the data need to be interpreted into account could circumvent the lack of knowledge on both sides and prevent misunderstandings. On the longer term, the training of data analysts with relevant social sciences and area studies or the training of humanitarian actors in data science could bridge this gap and facilitate the adoption of these methods by humanitarian actors.

As global instability grows, available humanitarian resources decline and timely, safe access to crisis regions decreases, amidst growing availability of real-time data from most corners of the globe, there may not be a realistic choice to humanitarian and security actors but to adopt data science and forecasting methods. Achieving such mutual understanding and trust as soon as possible is mandatory for efficient and proactive provision of conflict mitigation and humanitarian assistance.

Acknowledgments. This article is based on MA thesis written under the supervision of Daina Chiba as part of the MA track of the Essex Summer School in Social Science Data Analysis. The author is grateful to him for superb supervision as well as for feedback on earlier versions of the article. Similarly, the author is grateful for the feedback and advice of Kamran Bokhari, Andreas Beger, and

Vanessa Hubl on specific issues, as well as to the editors and four anonymous reviewers. The author specially thanks Claes Wallenius for assigning the time to finish the article during his work at the Swedish Defence University.

Funding Statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing Interests. The authors declare no competing interests exist.

Author Contributions. Conceptualization: T.L.; Data curation: T.L.; Formal analysis: T.L.; Investigation: T.L.; Methodology: T.L.; Project administration: T.L.; Software: T.L.; Validation: T.L.; Visualization: T.L.; Writing—original draft: T.L.; Writing—review and editing: T.L.

Data Availability Statement. The replication code of the article can be found on GitHub at <https://github.com/tamirlibel/LessonUnreplicated>.

References

- Bara C** (2020) Forecasting civil war and political violence. In Wenger A, Jasper U and Dunn Caveltly M (eds), *Probing and Governing the Future: The Politics and Science of Prevision*. London and New York: Routledge, pp. 177–193.
- Barfield T** (2016) Afghanistan’s arduous search for stability. *Current History* 115(780), 136–143.
- Bazzi S, Blair RA, Blattman C, Dube O, Gudgeon M and Peck R** (2019) The promise and pitfalls of conflict prediction: Evidence from Colombia and Indonesia. *The Review of Economics and Statistics* 104(4), 1–45.
- Beger A** (2021) icews: Get ICEWS event data. Available at <https://www.andybeger.com/icews/>; <https://github.com/andybega/icews>. Accessed on: 29/11/2021
- Beger A, Morgan RK and Ward MD** (2021) Reassessing the role of theory and machine learning in forecasting civil conflict. *Journal of Conflict Resolution* 65(7–8), 1405–1426.
- Blair RA and Sambanis N** (2020) Forecasting civil wars: Theory and structure in an age of “big data” and machine learning. *Journal of Conflict Resolution* 64(10), 1885–1915.
- Blair RA and Sambanis N** (2021) Is theory useful for conflict prediction? A response to Beger, Morgan, and Ward. *Journal of Conflict Resolution* 65(7–8), 1427–1453.
- Boiney J and Foster D** (2013) *Progress and Promise: Research and Engineering for Human Sociocultural Behavior Capability in the US Department of Defense*. Mclean, VA: Mitre Corp.
- Box-Steffensmeier JM and Tomlinson AR** (2000) Fractional integration methods in political science. *Electoral Studies* 19(1), 63–76.
- Cederman LE and Weidmann NB** (2017) Predicting armed conflict: Time to adjust our expectations? *Science* 355(6324), 474–476.
- Chadefaux T** (2017) Conflict forecasting and its limits. *Data Science* 1(1–2), 7–17.
- Chai T and Draxler RR** (2014) Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7(3), 1247–1250.
- Chiba D and Gleditsch KS** (2017) The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research* 54(2), 275–297.
- Colaresi M and Mahmood Z** (2017) Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research* 54(2), 193–214.
- Cranmer SJ and Desmarais BA** (2017) What can we learn from predictive modeling? *Political Analysis* 25(2), 145–166.
- Dowding K and Miller C** (2019) On prediction in political science. *European Journal of Political Research* 58(3), 1001–1018.
- Gerner DJ, Schrodt PA, Yilmaz O and Abu-Jabr R** (2002) *Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions*. New Orleans: International Studies Association.
- Global Administrative Areas** (2012). GADM database of Global Administrative Areas, version 2.0. [online] URL: www.gadm.org. Accessed on 26/10/2021
- Graves T, Gramacy R, Watkins N and Franzke C** (2017) A brief history of long memory: Hurst, Mandelbrot and the road to ARFIMA, 1951–1980. *Entropy* 19(9), 437–458.
- Hyndman RJ and Khandakar Y** (2008) Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 27(3), 1–22.
- Kotzé E, Senekal B and Daelemans W** (2020) Exploring the classification of security events using sparse and dense representation of text. In *2020 International SAUPEC/RobMech/PRASA Conference, Cape Town, South Africa, 29–31 January 2020*. New York: IEEE, pp. 1–6.
- Liu K, Chen Y and Zhang X** (2017) An evaluation of ARFIMA (autoregressive fractional integral moving average) programs. *Axioms* 6(2), 16–32.
- Lockheed Martin Advanced Technology Laboratories (ATL)** (2021) Integrated Crisis Early Warning System (ICEWS) Dataverse. Available at <https://dataverse.harvard.edu/dataverse/icews>. Accessed on: 29/11/2021
- Malejacq R and Mukhopadhyay D** (2016) The ‘tribal politics’ of field research: A reflection on power and partiality in 21st-century warzones. *Perspectives on Politics* 14(4), 1011–1028.

- Murtazashvili JB** (2022) The collapse of Afghanistan. *Journal of Democracy* 33(1), 40–54.
- Peng, C., Jatowt, A., & Yoshikawa, M.** (2020, March 30–April 3). *Conflict or cooperation?: predicting future tendency of international relations*. 35th Annual ACM Symposium on Applied Computing, Brno, Czech Republic. <https://doi.org/10.1145/3341105.3373929>
- Reisen V, Abraham B and Lopes S** (2001) Estimation of parameters in ARFIMA processes: A simulation study. *Communications in Statistics – Simulation and Computation* 30(4), 787–803.
- Rokem J, Weiss CM and Miodownik D** (2018) Geographies of violence in Jerusalem: The spatial logic of urban intergroup conflict. *Political Geography* 66, 88–97.
- Ruttig T** (2018) The Afghanistan Election Conundrum (12): Good news and bad news about district numbers. Afghanistan Analysts Network. Available at www.afghanistan-analysts.org/en/reports/political-landscape/the-afghanistan-election-conundrum-12-good-news-and-bad-news-about-district-numbers/ (accessed 7 September 2021).
- Schrodt PA and Analytics P** (2015) Event data in forecasting models: Where does it come from, what can it do. Unpublished manuscript. Available at <https://parusanalytics.com/eventdata/papers.dir/Schrodt.PRIO15.EventData.v1.1.pdf>.
- Shmueli G** (2010) To explain or to predict? *Statistical Science* 25(3), 289–310.
- Sida L and Oakley L** (2019) *Remote Monitoring in SDC: Challenges and Opportunities. Briefing Note: SDC-IDS Collaboration on Poverty, Politics and Participatory Methodologies*. Available at https://www.shareweb.ch/site/Poverty-Wellbeing/resources/Documents/SDC-IDS%20BriefingNote%2010_LSida_and_LOakley.pdf. Accessed on 29/11/2021
- Skarbek D** (2020) Qualitative research methods for institutional analysis. *Journal of Institutional Economics* 16(4), 409–422.
- Veenstra JQ** (2012) *Persistence and Anti-Persistence: Theory and Software*. Unpublished PhD thesis. Canada: Western University.
- Ward MD, Beger A, Cutler J, Dickenson M, Dorff C and Radford B** (2013) Comparing GDELT and ICEWS event data. *Analysis* 21(1), 267–297.
- Ward MD, Greenhill BD and Bakke KM** (2010) The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47(4), 363–375.
- Willmott CJ and Matsuura K** (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30(1), 79–82.
- Yonamine JE** (2013) *Predicting future levels of violence in Afghanistan districts using GDELT*. Unpublished manuscript. Available at <http://data.gdelproject.org/documentation/Predicting-Future-Levels-of-Violence-in-Afghanistan-Districts-using-GDELT.pdf> (accessed 14 October 2021).
- Zammit-Mangion A, Dewar M, Kadirkamanathan V and Sanguinetti G** (2012) Point process modelling of the Afghan war diary. *Proceedings of the National Academy of Sciences* 109(31), 12414–12419.