A Fast Spatial Multiple Imputation Procedure for Imprecise Armed Conflict Events*

Mihai Croicu[1] and Håvard Hegre[1]

[1]Department of Peace and Conflict Research, Uppsala University

March 26, 2018

Abstract

The proliferation of large, geographically disaggregated data on armed conflict, protest and other forms of political violence has led to a substantial development of research avenues. It has also brought with it numerous problems, both in terms of management of data quality and assurance of consistency. Dataset authors are generally aware of shortcomings in the sources they use, and alert users to such limitations. Still, researchers have been slow to provide solutions to these problems, even where they are clearly stated by dataset authors, in many cases due to the computational costliness of the canonical solutions. One such major but mostly ignored problem is the presence of large amounts of "known geographically imprecise" (KGI) or "known temporally imprecise" (KTI) observations in important datasets. In the spatial domain (KGI), dataset authors typically alert to geographical imprecision for 20–40% of observations . We introduce a simple, multiple-imputation based technique to address this issue. We show that the resulting imputations are reasonabley precise, and that using the imputed data considerably improves out-of-sample predictive performance relative to excluding the imprecise observations.

*The authors would like to thank Stefan Döring, Hanne Fjelde, Sophia Hatz, Kristian Petrova, David Randahl, Ida Rudolfsen, Eric Skoog, and Nina von Uexkull for helpful comments. The research was funded by the European Research Council, project H2020-ERC-2015-AdG 694640 (ViEWS). The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). For more information on the project see http://www.pcr.uu.se/research/views/.

1 The problem: Known geographic and temporal imprecision

In recent years, peace and conflict research increasingly use disaggregated (event) data on organized violence. Datasets such as UCDP-GED (Croicu and Sundberg, 2015), ACLED (Raleigh et al., 2010) or SCAD (Salehyan et al., 2012) are rapidly becoming gold standards in multiple areas of inquiry (Cunningham, Gleditsch, and Salehyan, 2016).

As a result, substantial work has been carried out identifying and evaluating the potential biases and missingness-related problems such data pose (Weidmann, 2014; Weidmann, 2016; Croicu and Kreutz, 2017). Indeed, the authors of such datasets themselves have devised sometimes highly complex, multi-dimensional systems (such as precision scores) to identify temporal, spatial and identification-related clarity problems for each individual observation (Croicu and Sundberg, 2013). For instance, many datasets would note that original source for a given event only records very inexactly its spatial location. The coders then typically report a precise placeholder location along with a code for the level of uncertainty associated with it (Croicu and Sundberg, 2013, pp. 35-38).

To date, no easy-to-implement, practical solution has been proposed to make proper use of the events that are flagged by the dataset creators as suffering from known explicit spatial imprecision problems. Hence, researchers often use the place-holder location assigned to it as precise location information despite being aware this is incorrect. This may make sense when it is important to capture another dimension of interest (e.g. have a more complete set of fatality estimates, a more complete count of protest events etc). However, they then introduce what we will call 'known geographic imprecision' (henceforth, KGI). Similar issues exist in the temporal domain. We will refer to this as 'known temporal imprecision' (KTI). Indeed, the overwhelming majority of research done using such data (von Uexkull, 2014; Raleigh and Hegre, 2009; Raleigh, Choi, and Kniveton, 2015; von Uexkull et al., 2016; Croicu and Kreutz, 2017; Pierskalla and Hollenbach, 2013; Fjelde and Hultman, 2014) use the data 'as-is', treating such observations in the same way as precisely observed observations. Alternatively, they use listwise deletion for those observations that do not pass an arbitrary threshold of completeness with ensuing problems related to data missingness.

Much of the KGI is of a non-random nature, correlating with many of the commonly used predictors of armed conflict (Croicu and Kreutz, 2017; Weidmann, 2016). Indeed, in the case of `UCDP-GED`, around 15% of all observations have very high KGI (see Section 2). These records only have the largest administrative unit or country known – and thus their corresponding spatial reference as assigned by the coding staff represents an area up to more than a thousand kilometers in radius. The non-random (MNAR) nature of this missingness is self-evident when looking at the data. The amount of observations listed as KGI per country in the geospatial domain ranges from 0 to over 50% of total per-country observations.

Furthermore, common coding standards compound the problem. Given this, the assigned point (typically the centroid of the administrative division or country (Croicu and Sundberg, 2013) regularly falls extremely far away from the area where known conflict activity has taken place.[1]

[1]The average distance between the assigned location for 'only country known' events and the mean actual location of fighting (the centroid of the conflict zone) is 293 km. The problem is especially acute if measuring conflict intensity

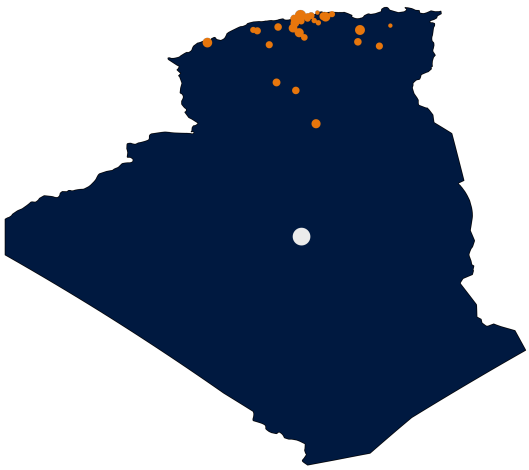

Figure 1. The Known Geographical Imprecision problem. The map shows the reported geographical locations for all state-based conflicts in Algeria in 1996 in UCDP GED 17.1. Yellow dots are the precisely located data points. The white, larger dot is the *assigned* location of data points for which insufficient information was available to code more precisely than 'within the country'. The size of each point is the logarithm of intensity of the fighting (recorded fatality estimates summed over all events) in a given location.

Figure 1 illustrates the problem with a real-life example. In Algeria in 1996, UCDP was unable to identify from its sources the location of 11% of events (comprising 43% of reported fatalities) of the observations more precisely than occurring 'within the country'. These points were assigned the centroid of the Algeria polygon as their location and marked with the precision score 6 corresponding to 'within the country'. The assigned locations (in the middle of the polygon, in white) are almost 1000 km from the cluster of known locations, and more than 400 km from the most proximate one. As there are multiple observations that are KGI, employing both of the most-widely used strategies are problematic. Keeping these with their assigned location in the data set will result in that remote location being the most intense in the country, attenuating any location-specific variable that is associated with the real event. Since the number of KGI observation for the polygon is substantially above the average globally, list-wise deletion will bias the result towards showing the polygon more peaceful than it actually was.

Thus, in this paper, we propose a relatively simple to use, computationally cheap method to address the known geographic imprecision problem through multiple imputation. We achieve this by drawing on recent geographically-precise events within the same conflict and the administrative division or country the KGI is assigned to. We use these event locations to compute a normalized score of latent conflict propensity for all spatio-temporal points within the region. After trans-

and not performing listwise deletion (e.g. Pierskalla and Hollenbach, 2013), as all the spatially imprecise points in a given geographic feature will be clustered to a single location, in many cases making that arbitrary point much more violent than the surrounding real, spatially precise locations. Exactly one third of the most intense 100 location-months in UCDP GED in fact known geographically imprecise locations, not the actual location of fighting.

forming this to a probability distribution, we sample a set of k imputations drawn from the set of spatio-temporal points. We estimate the performance of the method through 5-fold out-of-sample validation as well as through the analysis of predictive performance of the imputed datasets against the baseline (non-imputed) datasets.

2 Data

We apply and test the method on UCDP-GED, a dataset consisting of 135,181 observations of fighting resulting in at least one fatality. The data resolves down to individual location and day level, is global and covers the 1989 to 2016 period (Croicu and Sundberg, 2013). Out of these, 20,993 are KGI which we define as those events/observations flagged by the dataset maintainers as having only administrative feature known or only country known. [2]

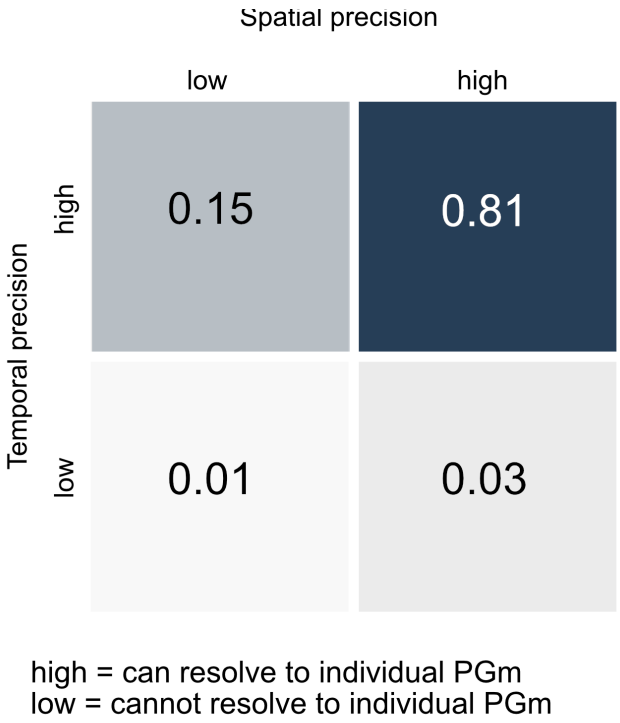

Figure 2. An illustration of known geographic imprecision (KGI) proportions in UCDP GED 17.1.

Each observation (both KGI and complete) is associated with a pair of armed actors fighting each other in a single UCDP armed conflict, as well as with a pair of geographic coordinates (latitude/longitude) of a location (Croicu and Sundberg, 2013). For geographically precise points

[2] In UCDP GED parlance, these are recorded as having geo-precision scores labeled `where_prec` 4 or 6. A further 25% of all observations are labeled as having a lower level of KGI, with smaller-level administrative divisions known. We treat these as complete data since the size of spatial uncertainty these point contain is found to be similar to that contained by a data-complete point (Weidmann, 2014), however the method presented below is similarly usable to impute these if so required.

that represents the location of fighting. For KGI points it is the centroid of the (best resolved) administrative division or country in which fighting has taken place.

The overall proportion of KGI data points to total data points varies significantly by country; for countries experiencing at least 25 UCDP GED events, 16% of all observations are fully KGI with a standard deviation of 12%. At one extreme, Spain and Brazil have zero KGI observations; at the other, Cambodia and Armenia have a proportion of over 50% KGI data points. Given this, list-wise deletion artificially underestimates both onsets and incidence in low-information areas and artificially overestimates them in high-information areas. Further, so does the space each KGI point represents (as each such point represents the surface of a country or administrative unit) – ranging from as little as 50 square km to as much as a few million square km.

3 Canonical solutions

There are two known canonical solutions to the problem – regression based multiple-imputation techniques and spatial point processes.

Examples of the first approach are MICE (Multivariate Imputations by Chained Equations, Buuren and Groothuis-Oudshoorn, 2011) and Amelia II (Honaker et al., 2011). These use regression techniques trained on known (non-missing co-variates) to determine a likely set of values for every missing data point (Buuren and Groothuis-Oudshoorn, 2011). The technique is problematic in our context. On the one hand, there is no agreed-upon standard set of predictors for armed conflict that would result in a non-biased dataset. On the other, conflict data are more often than not used as a response variable. This approach would certainly result in over-fitting, since the variables used for imputation (and thus determine the 'observed' response) are also used for prediction or explanation.

The second approach is exemplified by techniques such as kriging (Stein, 2012). It considers the known set of points as the product of an unknown but stable random Gaussian process operating in an $\mathbb{R}^2$ space, with the observed points as one realisation of the process. This allows the estimation of the best-fitting two-dimensional function that most closely fits the local auto-correlation with the help of a hyper-function describing the general shape of the auto-correlation (Stein, 2012). A major problem with this technique is the assumption of stability – i.e. that the unobserved spatial autocorrelation function takes a parametric form. This assumption is much more tenable with natural processes such as the identification of ore or petroleum deposits (where this technique is most used) than with armed conflict events that are interrelated in highly complex processes that determine entire armed conflicts.

Both approaches also suffer from high computational complexity. All point-process techniques follow a third-power law with the number of points used for training ($O(n^3)$ computational complexity). Hence, they are currently unfeasible for more than 5-10,000 observations (less than 10% of the actual size of GED or ACLED) without requiring even stricter assumptions regarding the stability of the process. These assumptions cannot be made with the kinds of problems we have

(Cressie and Johannesson, 2008). Regression-based models typically require multiple rounds of imputation until stability is reached. As most local conflict models are both temporally and spatially auto-correlative, they typically require a fairly complex lagged-dependency modelling which in turn requires the recalculation of both the left- and right-hand side of the equations at each round (Hegre et al., 2017; Colaresi and Mahmood, 2017; Brandt, Freeman, and Schrodt, 2014; Gleditch and Ward, 2012; Gleditsch and Ward, 2000). In sum, both approaches require extremely computationally expensive routines. For the vast majority of users, these computational costs are prohibitive, and in some cases, completely unfeasible.

4 A proposed practical solution

Instead, we propose a much simpler methodology, loosely inspired by the inverse distance weighting solution (Lu and Wong, 2008) widely used in geo-statistics and the Matched Wake cylinders introduced by Schutte and Donnay (2014).

The essence of the algorithm is to use whatever information is provided by other events in the same conflict for which precise geographical information exists, in order determine a plausible spatial distribution of conflict events in the administative region the KGI is located within. In Algeria, for instance (Figure 1), where we know the KGI events were part of the conflict between the Algerian government and AIS/FIS (see `http://ucdp.uu.se/#/conflict/386`), it is reasonable to assume that the real locations of these events were in the north of the country along with all the events we know where happened. We use the known locations for the subset of events close in time to the KGI event to create a normalized score of latent conflict propensity for all potential spatio-temporal points – a probability distribution over the population of locations within the administrative region of the KGI. In the Algeria example, this population is the entire country. When UCDP is able to assign an event to a first-order administrative region, the locations within this constitutes the population of potential points. This normalized score is then used to sample a set of k imputations drawn from the distribution of the latent variable over the relevant set of spatio-temporal points. :

4.1 Restricting assumptions

The resolution of these spatio-temporal points could theoretically be made infinitely detailed. In order to make this tractable we make a series of assumptions to reduce these to a finite set and to reduce computational complexity:

1. We restrict our application to estimate the latent conflict activity aggregated to a relatively coarse geographical grid, the 0.5×0.5 PRIO-GRID fishnet (Tollefsen, Strand, and Buhaug, 2012). These cells are about 50×50 kilometers in area. Even 'precisely coded' conflict event locations are typically uncertain spatially to a 50–75 km radius(Weidmann, 2014). Hence, imputation onto this grid implies a marginal reduction in real precision at the same time as it significantly reduces computational complexity. This choice is also useful since PRIO-GRID is

wide-spread in studies that make use of high-resolution, subnationally disaggregated conflict data.

2. Due to data-point sparseness, if any temporal imputation is required, the calendar month is the highest resolution supported in our imputations. Since the amount of KTI is very limited, with only 4% of the data suffering from KTI, and about 1% having more than two choices for imputation, we input KTI events by sampling from the uniform distribution over the possible month range.

3. We assume that the KGI data-points share the same data-generation process of the complete-information points within the same conflict, i.e. that KGI points originate from similar reporting as complete data points and that KGI points do not describe fundamentally different conflict patterns or behaviors as the ones captured in precise points.[3]

If the computational resources are available, the method we propose could of course be applied to problems with a finer spatial or temporal resolution. However, the method does require some aggregation over time and space.

4.2 The algorithm

Step 1 of the algorithm is to establish the population of locations into which the KGI point may be imputed. This step is facilitated by the following:

1. Each KGI point is represented by a geographical coordinate pair identifying the centroid of the administrative area where it occurred

2. Each event in the UCDP-GED dataset is recorded as belonging to a given conflict. This is a common property of all UCDP-GED points (Croicu and Sundberg, 2013); other datasets make similar connections (ACLED attaches actor information to each data point, SCAD attaches conflict information etc. to each point), allowing for generalization of our method.

3. The set of possible location is simplified to the relatively coarse PRIO-GRID cell per assumption 1 above

[3]In essence, this means that we assume KGI points not to represent fighting in completely different areas than spatially precise points. This point may be contentious; however, a manual investigation of the underlying reporting in 500 random geographically precise data-points and 500 KGI data points data points does not indicate two different DGPs for these events due to e.g. spatial remoteness of KGI data points. KGI originates, for the most part, from three issues: 1. the impossibility of the coder to identify the exact village or town by name in the gazetteer (frequently due to spelling issues); 2. the focus of the reporter on non-geographic information about fighting to the detriment of geography in the limited space dedicated to reporting etc. 3. the reporter describing multiple separate incidents of fighting, taking place in a given general area ("summaries"). It is fairly rare (under 10% of the sample of KGI points) that remoteness of fighting is the stated reason for KGI.

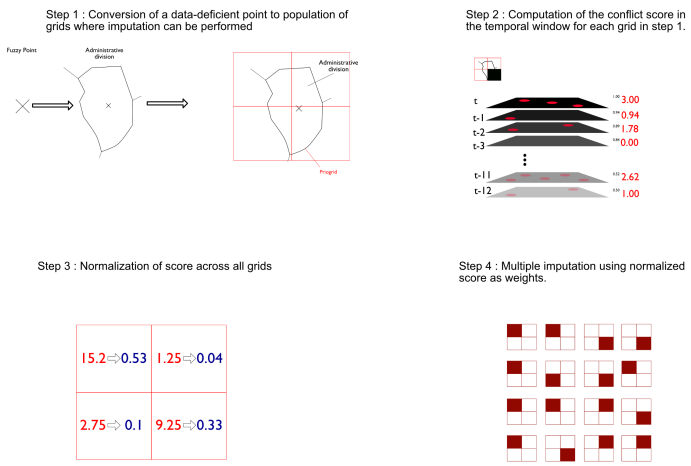

Figure 3. A graphical illustration of the algorithm using a simple simulated dataset of one KGI event and 24 precise events.

Given this, the problem in step 1 is reduced to finding, for each KGI event e, the set of PRIO-GRID cells that intersect with the administrative division or country that the KGI point represents.[4] We will use ADM_e as the notation for the administrative area recorded for each KGI point, APG_e as the notation for the population of PRIO-GRID cells associated with this administrative unit, and pg_i for each individual cell within this set:

$$pg \in APG_e$$

$$APG_e \subset PRIO - GRID$$

$$APG_e = ADM_e \cap PRIO - GRID$$

Technically, this is constructed using two separate spatial intersections; first one between the KGI point itself and a complete set of administrative divisions (for those KGI points describing administrative divisions) respectively countries (for those KGI points describing countries) in order to identify the correct ADM_e and then one between the ADM_e and the complete PRIO-GRID dataset to extract APG_e. In Figure 3, this process is shown in panel 1. The depicted contour of the administrative division is ADM_e, the collection of four intersecting PRIO-GRIDs APG_e and the four intersecting PRIO-GRIDs the four individual pg_i possible ($pg_{1,2,3,4}$).

The algorithm is able to process an arbitrary number of separate, time-variant datasets on administrative divisions (as borders do change). In practice, time-variant data was only used for countries in the form of the `Cshapes` version 0.6 dataset (Weidmann, Dorussen, and Gleditsch,

[4]PRIO-GRID cells situated at the border between to administrative entities are treated the same as non-border cell – since the imputation will be trained on precise events located in the same administrative unit and cell, the count of such events should already be proportional to the surface area of the cell in the administrative unit; restricting such cells at this point would double-count the border.

2010). For administrative divisions, the time-invariant `GADM` version 3 (GADM, 2012) data set was used.

Step 2 involves calculating the score representing the latent conflict activity. For each $pg_i \in APG_e$, this score is constructed from the counts of all geographically *precise* events that occur within the same conflict as e, restricted to a given historical window.

Selecting an appropriate historical window is tricky. The first obvious alternative is to use the same time window (i.e, the same month in our application) as the KGI event e and vary the window with each event.[5] This alternative is tantamount to using simultaneously-occurring, data-complete events as a basis for the estimation of the most probable locations of KGI events. However, this is not feasible given the rarity of conflict events. Even using the coarsened assumptions above (a spatial resolution of one PRIO-GRID and a temporal resolution of a month), this will result in a median count of 0 of complete-information events per potential $pg \in APG_e$ across all 20833 KGI events e. This obviously results in severe under-fitting (as any historical event in pg will lead to practically all imputations being made there).

Instead, we define a larger window – a fixed window of one year up to and including the month of e. Initial testing suggests this yields good results. Ttheoretically it also makes sense – UCDP's entire data catalogue, including the core definition of conflict, is built around the calendar year (Croicu and Sundberg, 2013). One problem with the wide window is the potential of overfitting. In effect, information that was reflecting the spatial extent of the conflict up to a year earlier is imposed on a much later point in time. This has the potential of, for example, contaminating test samples, as signal from the training period may be carried forward into the future (and thus into the test dataset), thus erroneously leading to noise in the future being interpreted as signal to the evaluation routines. This is, of course, only valid if the signal sent from the past has 'expired', i.e. if fighting has relocated over the period. Fortunately, conflict areas are relatively static, at least on a year-to-year basis (Beardsley, Gleditsch, and Lo, 2015). The signal provided by a data point a few months back, then, substantially overpowers the potential noise from a hypothetical, systematic shift in conflict locations.

To further mitigate this danger, we emulate learning and memory processes by applying an information decay function ($2^{-\Delta/12}$) to observations before the month of e, where Δ is the number of months since the preceding data point. Accordingly, a [geographically precise] [complete data] event occurring 12 months before the month of e will be weighted in the computation of the score at half what an event taking place in the current month of e will. This procedure is analogous (and identical in terms of the formula employed) to the one used in dynamic step-by-step forecasting models such as those in Hegre et al., 2016; Hegre et al., 2018, thus further mitigating potential overfitting. We recognize that this problem is still only partially solved. The potential for overfitting is evaluated in an out-of-sample predictive framework in the next section.

Further, we include a flat scalar prior b_e for the latent score of event in any cell pg_i. This allows

[5] This is possible as all UCDP-GED observations supply a time interval for each observation, not just a single date.

for imputing events in a pg without known conflict activity in the given historical window, as well as for imputation to occur where no other activity except e has happened in the corresponding APG_e. Assuming a high b_e implies increasing the spatial randomness of imputation across the APG_e; a lower b_e reduces it (increasing clustering around areas with complete information events). b_e is currently set to 0.1 for all e, but can be customized per event (e.g. a natural language processing could increase b_e for those events where the location of the event was described as diffuse or distant).

In essence, thus, the latent score s_e for each $pg \in APG_e$ for a given e is computed as:

$$s_{e,pg} = b_e + \sum_{\tau=0}^{12} 2^{-\tau/12} C(pg_{m-\tau})$$

where $C(pg, t)$ is a count function counting the number of data-complete points in `GED` in the same conflict as e for a given calendar-month t; m is the calendar month of e. In Figure 3, this is represented in panel 2, where the count of precise points in each grid in the past 12 months is adjusted using the decay function to create an additive score (e.g. for month $m - 0$, three precise points in a given PRIO-GRID cell add 3.00 to the score, as each point is valued as 1.0; while five points in month $m - 11$ contribute only 2.62 as each point is valued 2.62.

Step 3 In the third step, we create a probability matrix P_e containing probabilities $p_{e,pg}$ for each $pg \in APG_e$ by normalizing the $s_{e,pg}$ scores across the entire APG_e matrix, rescaling them so that they sum up to one.

$$p_{e,pg} = \frac{s_{e,pg}}{\sum_{i \in APG_e} s_i}$$

This is represented in panel 3 in Figure 3. The weight matrix is similar to a probability distribution function for the probability that a given event e occured in cell pg_i.

Step 4 Finally, m imputations $\tilde{Y_e} \in APG_e$ are sampled using P_e as the weights matrix, as represented in panel 4 in Figure 3.

Recall that imputations are done in terms of discrete PRIO-GRID cells, not the exact geographical location. As noted above, we believe the real decrease in precision is marginal.[6]

If richer data are available, the APG_e matrix can be built with a finer definition, or a hybrid method using the canonical methods above, but built for each pg can be employed.

4.3 Implementation details

The process itself is implemented in `Python` and `PostgreSQL/PostGIS`, and is highly computationally effective. Computational costs depend on the number of e to be imputed, and the size of the APG_e. The overall complexity is then of order $O(e * size(APG_e^2))$ with parallelization possible for each KGI event e. Imputation of the entire UCDP set of 20,993 KGI points takes under 1 core-hours to run.

[6] If an estimated pair of geograhpical coordinates are required, users may simply do a random draw from the bivariate uniform distribution across the geographical extent of the PRIO-GRID cell.

In comparison, an adaptation of MICE round using the best performing model in the Dynasim one-step ahead toolkit available at time of running (Hegre, Hoyles, and Nordkvelle, 2018; Hegre et al., 2018) and 5 iterations,[7] also trained and imputed on PRIO-GRID-month level,[8] took over 80 core-hours to execute, and performed worse (see below). Iterating the right-hand side of the equation, given the highly complex spatio-temporal lagging required, was estimated to take between 1000 and 1500 core-hours, completely unfeasible for most users.

5 Evaluation

This imputation algorithm performs very well despite its simplicity. We here present a tripartite evaluation. First, we assess how well it imputes in an experiment where we artificially impose KGI on events for which we know the exact location. Second, we demonstrate how imputation improves standard statistical estimation procedures. Third, we show how the procedure improves forecasting of conflict events.

5.1 Evaluating precision of imputations

To evaluate how well it fills in values for the UCDP-GED dataset, we perform a k-fold out-of-sample validation (Kohavi, 1995) setting $k = 5$. This choice is commonly used in similar application and has the additional advantage that it yields evaluation folds with a similar proportion of KGI observations as is empirically the case in the UCDP-GED dataset.

The out-of-sample evaluation is based solely on the UCDP-GED events with complete geographical information in order to have true actuals to compare with. We split this 'sample' randomly into five equal-size partitions. In each of the 5 iterations, we turn one partition into a test set of observations that are KGI. This is done by removing the information on the actual location of the event while retaining information on the administrative division (ADM_e) it occurred within, emulating the uncertainty that the UCDP records for such events.[9] The multiple imputation routine is then trained on the remaining four partitions to impute the geographical location (the training set). The imputation result is then compared to the known (real) values in the test set to estimate how well the algorithm performed. The evaluation metric is the average over all points in the test set of the

[7] The actual model used was a general additive model, using the following equation, in the `R` formula language: $conflict.dummy \sim ged.dummy.tlag1 + s(dist.tlag1) + s(dist.tlag2.6) + s(dist.tlag7.11) + proxevent + proxevent.lag1 + proxevent.lag2 + ln(pop) + ln(ttime) + mountains.mean + ln(capdist) + ln(imrmean) + gcp.li.mer + diamsec.s + gem.s + goldplacer.s + goldsurface.s + petroleum.s + ln(diamsec) + ln(gem) + ln(gold1) + ln(gold2) + ln(petroleum) + agri.ih + barren.ih + forest.ih + savanna.ih + shrub.ih + pasture.ih$, where s(x) is smoothed, non parametric form of x, dist the distance to the nearest event, proxevent the time since nearest event, tlag.m.[n] the temporal lag at month m or the additive temporal lag between m and n. For a full definition of the data, see the citation. This equation was selected as it had the highest predictive performance for conflict of all the 'N-step ahead' models in the ViEWS framework as of July 2017 (Hegre et al., 2018).

[8] This training used 100% of all positive (conflictual PRIO-GRID-month cells) and 10% of the negative cells (for balance and fairness of comparison), and only iterating the left-hand side of the equation

[9] This is a simple procedure as it only involves a change of precision score; since any location within the polygon is treated the same by the algorithm during Step 1, there is no practical need to actually identify the centroid of the feature where the complete point is located.

distance between the actual geographical location and the imputed location. This procedure was repeated for each of the five partitions, so that each data point has been imputed at least once.

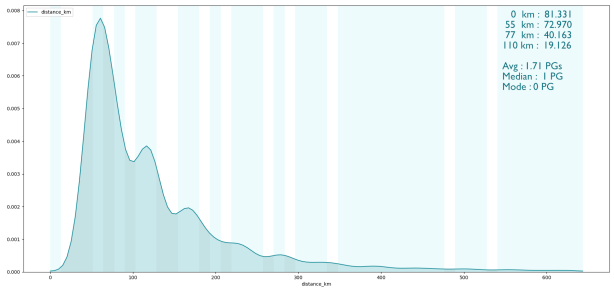

Figure 4. Kernel density plot of the spatial error between actual location (grid) of event and inputted location of event in out of sample 5-fold evaluation at 20% KGI. 15 imputations for each such point were done. Blue lines are the errors corresponding to the the actual imputed values.

	Fold				
	1	2	3	4	5
Number of imputed *pgms*	302666	302890	302666	302778	302876
Mode of distance (actual–imputed)	0	0	0	0	0
Median distance (actual–imputed)	1.0	1.0	1.0	1.0	1.0
Mean distance (actual–imputed)	1.71	1.68	1.7	1.69	1.71
Count imputed in correct pg_i	81331	83027	82225	83576	82177
Count imputed 1 grid away	72970	72970	72143	71664	72834
Count imputed 1.41 grids away (diagonally)	40163	40153	40095	40123	39937
Count imputed 2 grids away	19126	18708	19088	18933	19565

Table 1. Comparing actual and imputed locations of KGI observations, by fold. Since imputation is done on a grid, all distances are expressed in grid units. One grid is .5 square degrees WGS84, approximately 55 by 55 km at the equator. For each KGI point e 15 imputations were generated.

Results are presented in Table 1 and Figure 4. In this application, the procedure imputes into the correct cell in a plurality of cases, with a median error of 1. In just above 50% of the cases, the imputed location was either correct or into in the cell immediately adjacent to the correct one). In terms of variance, the approach also performs well – the counts of distances between the real location of the points in the test set and the imputed location of those points resembles a Poisson distribution with a low λ parameter (since we input on a discrete grid, the distribution is discrete), which indicates the method operates satisfactorily and in accordance to theoretical expectations (Raghunathan et al., 2001).

In contrast, the MICE adaptation using the GAM model described above performed substantially worse. We tested a single 20% fold using the same evaluation routine, varying only the left side of the equation during the 5 iterations of the model. The MICE adaption produced the following distribution of errors across the 15 imputations: The mode of errors were 1, the median 2 and the mean 2.55. No kriging specification (using PyKrige) could be made to converge.

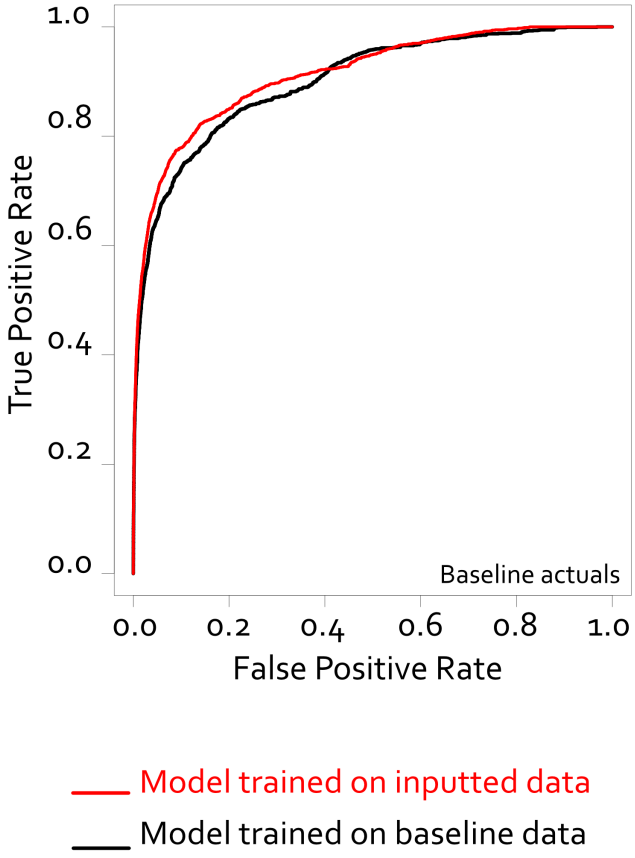

Figure 5. ROC plot of an inputted and non-inputted model (otherwise identical) predicting a non-inputted test 1-36 months in the future of the training set.

5.2 Contribution to statistical modeling

Table 2 shows the results from estimating a statistical model on 1989–2013 data on the imputed data and on the non-imputed data only containing complete information observations (in effect, assuming that the events did not happen). The model is an example of models underlying the ViEWS Dynasim component (Hegre, Hoyles, and Nordkvelle, 2018, see more on that model below).

The four left-most columns in Table 2 shows the average parameter estimates across the five imputations and the mean standard errors corrected using the Rubin rule (Rubin, 1976). The four columns to the right show the corresponding estimates when we exclude the geographically imprecise events.

The estimates for the imputed datasets are roughly similar to those for the non-imputed ones. Statistical significance levels are roughly similar, and the pseudo-R squares for the two sets of models do not differ much. There are no indications that our imputation procedure bias results or standard errors.

name	*b_imp*	*s_imp*	*z_imp*	*p_imp*	*b_nonimp*	*se_nonimp*	*z_nonimp*	*p_nonimp*
agri_ih_li	-0.0020	0.0011	-1.8820	0.060	-0.0022	0.0010	-2.2000	0.028
barren_ih_li	-0.0070	0.0010	-6.8709	0.000	-0.0087	0.0010	-8.7000	0.000
decay_12_cw_ged_dummy_ns_0	0.5436	0.0791	6.8746	0.000	0.5959	0.0820	7.2671	0.000
decay_12_cw_ged_dummy_os_0	1.1237	0.0474	23.7295	0.000	1.1899	0.0510	23.3314	0.000
decay_12_cw_ged_dummy_sb_0	4.0129	0.0492	81.5214	0.000	4.2060	0.0530	79.3585	0.000
excluded_li	0.2688	0.0202	13.3112	0.000	0.2624	0.0230	11.4087	0.000
forest_ih_li	-0.0046	0.0010	-4.4866	0.000	-0.0043	0.0010	-4.3000	0.000
gcp_li_mer	0.0049	0.0142	0.3448	0.730	0.0184	0.0150	1.2267	0.220
imr_mean	0.0007	0.0000	16.6450	0.000	0.0008	0.0000	16.3265	0.000
Intercept	-8.6633	0.2724	-31.7977	0.000	-8.5551	0.2920	-29.2983	0.000
l1_ged_dummy_sb	0.2961	0.0390	7.5989	0.000	0.2729	0.0400	6.8225	0.000
l2_ged_dummy_sb	0.1632	0.0390	4.1832	0.000	0.1146	0.0400	2.8650	0.004
l3_ged_dummy_sb	0.1707	0.0376	4.5333	0.000	0.0975	0.0410	2.3780	0.017
l4_ged_dummy_sb	0.2176	0.0413	5.2620	0.000	0.1352	0.0410	3.2976	0.001
l5_ged_dummy_sb	0.1973	0.0405	4.8733	0.000	0.1994	0.0420	4.7476	0.000
l6_ged_dummy_sb	0.0843	0.0417	2.0225	0.043	0.0433	0.0440	0.9841	0.325
l7_ged_dummy_sb	0.2347	0.0443	5.3039	0.000	0.1942	0.0440	4.4136	0.000
l8_ged_dummy_sb	0.2937	0.0426	6.9015	0.000	0.2865	0.0440	6.5114	0.000
l9_ged_dummy_sb	0.3174	0.0480	6.6068	0.000	0.3455	0.0440	7.8523	0.000
ln_bdist3	-0.0664	0.0101	-6.5675	0.000	-0.0729	0.0120	-6.0750	0.000
ln_capdist	0.0751	0.0162	4.6242	0.000	0.0645	0.0170	3.7941	0.000
ln_dist_diamsec	0.2911	0.0240	12.1235	0.000	0.3007	0.0260	11.5654	0.000
ln_dist_petroleum	-0.3108	0.0213	-14.6103	0.000	-0.2703	0.0250	-10.8120	0.000
ln_pop	0.1787	0.0134	13.3230	0.000	0.1827	0.0150	12.1800	0.000
ln_ttime	-0.1416	0.0259	-5.4642	0.000	-0.2210	0.0290	-7.6207	0.000
mountains_mean	0.0389	0.0430	0.9029	0.367	0.0554	0.0460	1.2043	0.228
pasture_ih_li	0.0010	0.0010	1.0308	0.303	0.0008	0.0010	0.8000	0.424
q_1_1_l1_ged_dummy_ns	-0.1646	0.0808	-2.0365	0.042	-0.2452	0.0860	-2.8512	0.004
q_1_1_l1_ged_dummy_os	0.0644	0.0246	2.6218	0.009	0.0330	0.0250	1.3200	0.187
q_1_1_l1_ged_dummy_sb	0.2878	0.0240	12.0112	0.000	0.2936	0.0210	13.9810	0.000
q_1_1_l2_ged_dummy_sb	0.1266	0.0218	5.8108	0.000	0.1457	0.0220	6.6227	0.000
q_1_1_l3_ged_dummy_sb	0.0965	0.0292	3.3037	0.001	0.0953	0.0210	4.5381	0.000
savanna_ih_li	-0.0053	0.0010	-5.2242	0.000	-0.0046	0.0010	-4.6000	0.000
shrub_ih_li	0.0059	0.0011	5.5626	0.000	0.0079	0.0010	7.9000	0.000
urban_ih_li	0.0448	0.0122	3.6855	0.000	0.0397	0.0120	3.3083	0.001

Table 2. Parameter estimates; model estimate on imputed and non-imputed data

Unfortunately, we do not currently have access to results from a similar models using non-imputed data *including* the KGI points. We expect that these results would compare unfavorably to both models presented in Table 2.

5.3 Contribution to predictive performance

We then test whether the imputation strategy improves performance in a predictive framework. We choose the hardest test possible, the Dynasim framework, a multi-step simulation predicting month-by-month 36 months in advance (Hegre et al., 2016), accounting for both model and data uncertainty. We train on 1989–2013 data and predict the risk of conflict for 2014–2016.

To also account for the possibility of overfitting discussed above, we further strengthen our test routine. We train Dynasim on both the imputed and the non-imputed datasets only containing complete information observations (in effect, assuming that the events did not happen). We then predict only the complete information events in the test period (2014–16), so that there can not be any biasing involved (i.e. no unfair advantage for the model in "shadows" of past (training) noise seeping into the future (predicting) data). In effect, if the inputted dataset predicts the complete information events (true, full signal) better, the imputation algorithm adds true predictive power to the model and thus improves the noise/signal ratio in the dataset.

ROC curves for Dynasim are presented in Figure **??**. The area under the curve improves from 0.882 in the baseline model to 0.904 in the imputed model, showing both overall improvement of the model when using imputation as well as a lack of over-fit by this imputation strategy. The relatively small improvement (as well as the very high area under the curve (AUC)) can be explained as no cost function was involved and the test data is extremely unbalanced (98% of the data are non-conflict observations, only 2% are conflict observations) – with more balanced data, a lower AUC for both models can be expected as well as a much higher improvement in performance when using inputted data (as the performance increase comes from a better prediction of the small class (conflict) not from an improvement in predicting zeroes).

6 Conclusions

We have presented a simple and fast spatial multiple imputation procedure that produces a distribution of plausible locations for armed conflict events rather than simple imposing the centroid of a country or an administrative unit as a convenient, but clearly biased placeholder location. The procedure performs well. Out-of-sample evaluation indicates that the procedure suggests locations that is not more than 75km off in more than half of the cases. The procedure does not seem to be systematically biased. The proecdure is enormously useful for forecasting purposes – out-of-sample evaluation of predictive performance indicates that forecasts based on imputed data greatly improves on those omitting the imprecise observations.

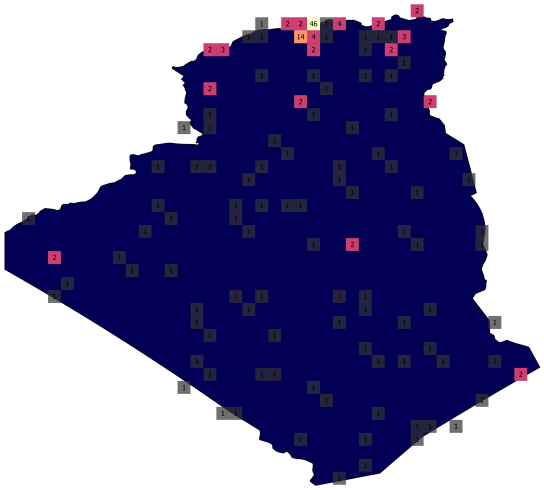

Figure 6. The result of the imputation algorithm (15-imputations for each data-defficient point) for the example given in the introduction.

References

Beardsley, Kyle, Kristian Skrede Gleditsch, and Nigel Lo (2015). "Roving bandits? The geographical evolution of African armed conflicts". In: *International Studies Quarterly* 59.3, pp. 503–516.

Brandt, Patrick T, John R Freeman, and Philip A Schrodt (2014). "Evaluating forecasts of political conflict dynamics". In: *International Journal of Forecasting* 30.4, pp. 944–962.

Buuren, Stef van and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1–67. URL: `http://www.jstatsoft.org/v45/i03/`.

Colaresi, Michael and Zuhaib Mahmood (2017). "Do the Robot: Lessons from Machine Learning to Improve Conflict Forecasting". In: *Journal of Peace Research* 54.2, pp. 193–214.

Cressie, Noel and Gardar Johannesson (2008). "Fixed rank kriging for very large spatial data sets". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 209–226.

Croicu, Mihai and Joakim Kreutz (2017). "Communication Technology and Reports on Political Violence: Cross-National Evidence Using African Events Data". In: *Political Research Quarterly* 70.1, pp. 19–31.

Croicu, Mihai and Ralph Sundberg (2013). *UCDP Georeferenced Event Dataset Codebook Version 4.0*. Typescript, Uppsala Conflict Data Program. URL: `http://www.pcr.uu.se/research/ucdp/datasets/ucdp_ged/`.

— (2015). "UCDP Georeferenced Event Dataset Codebook Version 4.0". In: *Journal of Peace Research* 50.4, pp. 523–532. URL: `http://www.pcr.uu.se/research/ucdp/datasets/ucdp_ged/`.

Cunningham, David E, Kristian Skrede Gleditsch, and Idean Salehyan (2016). "Trends in Civil War Data". In: *What Do We Know about Civil Wars?* Ed. by T. David Mason and Sara McLaughlin Mitchell. Rowman & Littlefield, pp. 247–260.

Fjelde, Hanne and Lisa Hultman (2014). "Weakening the Enemy: A Disaggregated Study of Violence against Civilians in Africa". In: *Journal of Conflict Resolution* 58.7, pp. 1230–1257.

GADM (2012). *Geoadministrative Areas Database, 2.0*. URL: http://www.gadm.org/.

Gleditch, Kristian S. and Michael D. Ward (2012). "Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes". In: *Journal of Peace Research* 50.1, pp. 17–31.

Gleditsch, Kristian S. and Michael D. Ward (2000). "War and Peace in Space and Time: The Role of Democratization". In: *International Studies Quarterly* 44.1, pp. 1–29.

Hegre, Håvard, Frederick Hoyles, and Jonas Nordkvelle (2018). *Dynasim – A fast and simple simulator of endogenous (simultaneous) panel data regression models*. Typescript, Uppsala University.

Hegre, Håvard et al. (2016). "Forecasting civil conflict along the shared socioeconomic pathways". In: *Environmental Research Letters* 11.5, p. 054002. DOI: 10.188/1748-9326/11/5/054002.

Hegre, Håvard et al. (2017). "Introduction: Forecasting in peace research". In: *Journal of Peace Research* 54.2, pp. 113–124.

Hegre, Håvard et al. (2018). *ViEWS: A political Violence Early Warning System*. Typescript, Uppsala University.

Honaker, James et al. (2011). "Amelia II: A program for missing data". In: *Journal of statistical software* 45.7, pp. 1–47. URL: http://artax.karlin.mff.cuni.cz/~hans/src/doc/r-cran-amelia/amelia.pdf (visited on 05/15/2016).

Kohavi, Ron (1995). "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 1137–1143. URL: http://dl.acm.org/citation.cfm?id=1643031.1643047.

Lu, George Y and David W Wong (2008). "An adaptive inverse-distance weighting spatial interpolation technique". In: *Computers & Geosciences* 34.9, pp. 1044–1055.

Pierskalla, Jan H. and Florian M. Hollenbach (2013). "Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa". In: *American Political Science Review* 107.2, pp. 207–224.

Raghunathan, Trivellore E et al. (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models". In: *Survey Methodology* 27.1, pp. 85–96.

Raleigh, Clionadh, Hyun Jin Choi, and Dominic Kniveton (2015). "The devil is in the details: An investigation of the relationship between conflict, food prices and climate across Africa". In: *Global Environmental Change* 32, pp. 187–199.

Raleigh, Clionadh and Håvard Hegre (2009). "Population, Size, and Civil War. A Geographically Disaggregated Analysis". In: *Political Geography* 28.4, pp. 224–238.

Raleigh, Clionadh et al. (2010). "Introducing ACLED: An Armed Conflict Location and Event Dataset". In: *Journal of Peace Research* 47, In press.

Rubin, D.B. (1976). "Inference and missing data". In: *Biometrika* 63, pp. 581–592.

Salehyan, Idean et al. (2012). "Social Conflict in Africa: A New Database". In: *International Interactions* 38.4, pp. 503–511. DOI: 10.1080/03050629.2012.697426. URL: http://www.tandfonline.com/doi/abs/10.1080/03050629.2012.697426.

Schutte, Sebastian and Karsten Donnay (2014). "Matched wake analysis: finding causal relationships in spatiotemporal event data". In: *Political Geography* 41, pp. 1–10.

Stein, Michael L (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug (2012). "PRIO-GRID: A unified spatial data structure". In: *Journal of Peace Research* 49.2, pp. 363–374. DOI: 10.1177/0022343311431287.

von Uexkull, Nina (2014). "Sustained drought, vulnerability and civil conflict in Sub-Saharan Africa". In: *Political Geography* 43, pp. 16–26.

von Uexkull, Nina et al. (2016). "Civil conflict sensitivity to growing-season drought". In: *Proceedings of the National Academy of Sciences* 113.44, pp. 12391–12396.

Weidmann, Nils (2016). "A closer look at reporting bias in conflict event data". In: *American Journal of Political Science* 60.1, pp. 206–218.

Weidmann, Nils B. (2014). "On the Accuracy of Media-based Conflict Event Data". In: *Journal of Conflict Resolution* Online first, DOI: 10.1177/0022002714530431, pp. 1–21.

Weidmann, Nils B, Hans Dorussen, and Kristian Skrede Gleditsch (2010). "The Geography of the International System: The CShapes Dataset". In: *International Interactions* 36.1, pp. 86–106.