Regions at Risk
Predicting Conflict Zones in African Insurgencies

Sebastian Schutte*

Zukunftskolleg / Department of Politics and Management, University of Konstanz

sebastian.schutte@uni-konstanz.de

WORD COUNT: 8572

Abstract

A method for predicting conflict zones in civil wars based on point process models is presented in this paper. Instead of testing the validity of specific theoretical conjectures about the determinants of violence in a causal framework, this paper builds on classic literature and a wide body of recent studies to predict conflict zones based on a series of geographic conditions. Using an innovative cross-validation design, the study shows that the quantitative research program on the micro-foundations of violence in civil conflict has crafted generalizable insights permitting out-of-sample predictions of conflict zones. The study region is delimited to 10 countries in Sub-Saharan Africa that experienced full-blown insurgencies in the post-Cold War era.

*I would like to thank Rolf Turner for having saved the day twice when I got lost in the intricacies of the `spatstat` package for the `R` programming language. Yuri Zhukov provided very helpful feedback as a discussant for this paper at the Peace Science Society Workshop "Disaggregation in Terrorism Studies", Philadelphia PA, October 9, 2014.

1 Introduction

In June 2014, the civil war in Iraq reached a turning point when the "Islamic State of Iraq and the Levant" (ISIL) group captured seven major cities in the northern part of the country.[1] The Kurdish-dominated areas in Iraq and Syria have been traditionally calmer in both war-torn countries and neither international organizations nor governments had seen this escalation coming. This episode demonstrates that regions at risk in ongoing conflicts are hard to identify even under the watchful eye of the international community. With the recent uprisings of the Arab Spring, the ongoing violence in Iraq and Afghanistan, and numerous conflicts in central Africa, ISIL's advances will not be the last geographic expansion of conflict with disastrous humanitarian consequences.

The question therefore springs to mind whether and to what extent the scholarly research program on irregular conflicts can help us to predict major conflict zones in civil wars in advance. Recent empirical research on the spatial determinants of violence in civil conflict has generated substantial insights. Theoretically, the failure of states to control their remote periphery has been repeatedly used as an explanation for political violence (Herbst, 2000; Fearon and Laitin, 2003; Herbst, 2004; Scott, 2009; Buhaug et al., 2009). Drawing on these insights, a series of studies has combined Geographic Information Systems and multivariate regression designs to test related hypotheses (Buhaug and Gates, 2002; Buhaug and Rød, 2006; Buhaug et al., 2009). Moreover, properties of irregular conflicts have also been modeled in disaggregated computational studies drawing on geographic information (see Bhavnani et al., 2008; Weidmann and Salehyan, 2012; Bhavnani et al., 2013).

Despite this progress in combining theoretical and quantitative insights, the external validity and in particular the predictive capabilities of this research program remain understudied. On the country level, quantitative predictions of political instability have made substantial progress in recent years (see Goldstone et al., 2010; Ward et al., 2013). Beyond their practical utility of informing relief organizations and policy decision, predictions offer scientific benefits as they directly communicate the degree to which a studied phenomenon is understood (Ward et al., 2010; Schrodt, 2014).

Yet another advantage of predictions is that they are not restricted to the empirical sample: predicting locations of violent conflict beyond the sample that the model was fit-

[1] see `http://www.stratfor.com/image/islamic-state-timeline`, last retrieved on Sep. 9, 2014.

ted on reveals to what extent the data generating mechanism and relevant variables were correctly identified. Based on these considerations, this paper tests to what extent geographic covariates of irregular warfare that have been identified in previous work improve predictions of conflict zones. To evaluate these predictions, I rely both on a quantitative and a qualitative metric. Spatial predictions of conflict intensities are compared to empirically observed intensities to calculate error scores. Moreover, high intensity conflict regions yielding more than 50 percent of the maximal conflict intensity are compared visually to empirically observed hot spots.

The merit of this exercise is twofold: first, a direct comparison between the predictions of models and random baselines serves as a reality check for the geographic research program, clearly communicating to what extent informed predictions outperform random guesses. Second, the applied methodology generates easily communicable predictions that could be utilized beyond basic research to aid planning of humanitarian relief operations, for example.[2]

To demonstrate the predictive capabilities of the associated variables, point process models are used to predict instances of lethal violence in ten recent insurgencies in Sub-Saharan Africa. Predictions of these models are compared to the empirical record using an innovative cross-validation design. The results indicate that central variables of the geo-quantitative research program lead to drastically improved out-of-sample predictions in comparison to uniform baselines.

2 Literature review

The connection between geography and war has long been considered important. Already the classic literature of revolutionary warfare and counterinsurgency devoted attention to the topic (see McColl, 1969; Guevara, 1961, 10; Mao [1938] 1967).

Contemporary research has stressed the primarily local determinants of fighting in civil conflicts (Buhaug and Gates, 2002; Buhaug and Rød, 2006; O'Loughlin and Witmer, 2010; Buhaug et al., 2011; Rustad et al., 2011). While interstate armies generally move and fight under central command, irregular conflicts are frequently fought out between local militias and rebel supporters. Instead of strategic decisions of where to send mechanized armies to fight, local encounters between irregular fighters and military units

[2]To this end, I have extrapolated likely conflict zones for Africa and the Greater Middle East as explained in the supplementary information.

determine much of the violence in civil conflicts (Kalyvas, 2005).[3]

A series of publications has been devoted explicitly to coding and explaining the location, size, and extent of primary conflict zones in armed conflict. Buhaug (2010) applies a distance-decay model from the study of interstate war to internal conflicts and finds that the relative military strength of the belligerents is a strong predictor for the location of primary conflict zones. Drawing on a bargaining perspective, Butcher (2014) analyzes the location of conflict zones and concludes that multilateral sub-national conflicts tend to occur more in the periphery. Braithwaite (2010) analyzes under which conditions hot spots of international armed conflict are likely to emerge. Hallberg (2012) contributes a geo-referenced data set on primary conflict zones in civil wars since 1989. The recent turn towards more disaggregated empirical studies has led to an increased interest in data on single conflict events (Raleigh and Hegre, 2005; Sundberg et al., 2011) and geo-referenced data on local determinants of conflict intensity. Consequently, geographic and local socio-economic conditions have moved into the focus of empirical studies. Drawing on spatially disaggregated data on wealth and data on conflict events, Hegre et al. (2009) found that violence tends to cluster in more wealthy regions, possibly because rebels prioritize them in their attacks. Raleigh and Hegre (2009) also found that population concentrations generally see higher levels of fighting. While the statistical association is strong, it remains unclear how this effect comes about. A relatively constant per-capita rate of violence as well as strategic targeting of civilian concentrations spring to mind as possible explanations.

While the emphasis on local determinants of conflict is justified both theoretically and empirically, the diffusion of irregular civil conflict over time has also been studied. <AUTHOR> investigated different diffusion scenarios for violence in civil wars, comparing instances of empirical diffusion against random baseline scenarios. Zhukov (2012) used road-network information in a refined empirical analysis and found that violence in the north Caucasus tended to relocate over time along roads. Both studies point to the fact that a substantive number of civil war events result from previous fighting in neighboring regions rather than being solely caused by local conditions. Beyond spatial expansion, reaction to specific instances of violence has also been analyzed (Lyall, 2009; Kocher et al., 2011; Braithwaite and Johnson, 2012, <AUTHOR>). Again, reactive patterns in

[3] Especially the conflicts in Sub-Saharan Africa since 1990 have seen major involvements of irregular forces. This also applies to the series of clashes referred to as “Africa’s World War” between 1998 and 2003 (see Prunier, 2009).

conflict event data underline that the conflict history as well as local socioeconomic and geographic conditions jointly affect levels of violence in civil wars.

In summary, the presented literature on the determinants of violence in civil conflicts suggests an interaction of multiple factors. Strategically, the military capabilities of the actors as well as terrain conditions and infrastructure play an important role for the locations of major battle zones. On a tactical level, violence tends to cluster as actors fight repeatedly over specific locations, but it also diffuses into previously unaffected regions. Finally, the types of violence applied by actors in the field crucially affects subsequent levels of violence. While these insights are important for testing and building theories of the dynamics of violence in irregular conflict, the question of whether or not they translate into generalizable and ultimately actionable knowledge remains unanswered. Regression studies and matching designs are generally used to test whether or not specific variables have a causal effect in line with theoretical considerations. The estimated effects, however, relate solely to hypothetical all-else-being-equal, or "*cetris paribus*" scenarios.

Despite the obvious merit of inferential designs, the added scientific and practical value of predictions has been pointed out in recent publications (Ward et al., 2010; Goldstone et al., 2010; Gleditsch and Ward, 2012; Schrodt, 2014). Tangible predictions about the location and timing of violence are a concern of policy makers and relief organizations. Consequently, political scientists have embarked on generating predictions for which countries are likely to experience civil wars (Weidmann and Ward, 2010; Goldstone et al., 2010; Ward et al., 2010, 2013) and which regions are most prone to violence in Afghanistan (Zammit-Mangion et al., 2012; Yonamine, 2013). Advancing this line of research, this paper is a first attempt to predict major conflict zones *across* civil conflicts. The performance of these predictions is assessed in comparison to random baselines. In essence, this paper communicates how much predictive power the quantitative research program on the micro-dynamics of civil wars has gained in comparison to agnostic guessing about where violence will occur. This exercise serves as an important reality-check for our ability to predict sub-national conflict intensity based on central variables identified in the literature.

Of course, assessing the predictive performance of these variables requires a suitable empirical setup. The paper proceeds as follows: in the next section, I will discuss the case selection for this study. After that, I will identify central variables for the prediction of variation in conflict intensity from the above-referenced literature. A generic setup for predicting violence based on these variables is presented in the subsequent section, loosely based on Zammit-Mangion et al. (2012). Finally, I will test whether and to what extent these variables produce improved out-of-sample predictions in comparison to agnostic baselines.

3 Scope and case selection

Because the literature on revolutionary warfare and counterinsurgency studies have been most vocal in proposing a direct link between rebel presence and terrain conditions, I decided to narrow down the empirical analysis to insurgencies, i.e. conflicts in which the rebels are not recognized as belligerents and heavily rely on civilian assistance to wage a guerrilla war against the state (Galula, 1964; CIA, 2009, 2). However, not all civil or irregular conflicts are insurgencies. Kalyvas and Balcells (2010) report a declining trend for this type of conflict and an increase in wars that blend elements of conventional fighting with irregular rebellions. Moreover, fighting in quasi-conventional civil wars, such as in Yugoslavia in the early 1990s, might be better predicted by ethnic boundaries than terrain conditions. Despite the overall decline in the frequency of insurgencies, they still constitute the most frequent type of armed conflict in the post-World War II period. Conflict events from the "Geo-referenced Event Dataset" (GED) (Sundberg et al., 2011) cover lethal clashes that occurred between 1990 and 2010 from 42 African countries. Drawing on a separate dataset by Lyall and Wilson (2009), I identified 11 cases of insurgency that are covered in GED.[4] I decided to exclude Djibouti (1991-2001) because the country is too small for meaningful geographic analysis given the resolution of the covariates. Table 1 shows the remaining cases that were used for the analysis.

4 Spatial determinants of fighting

The localized nature of fighting in civil conflicts provides a suitable starting point for predictive modeling. Recent studies have utilized digital information on geographic conditions and conflict events to reveal a series of robust statistical relationships. I will therefore introduce conflict event datasets and data on the spatial determinants of violence that have been identified by previous studies to systematically test to what extent predictions of conflict intensity can be improved by each variable.

[4]In the GED dataset, I focused on violence by or against the state and observations that fell into periods and countries experiencing active insurgencies according to Lyall and Wilson (2009).

No.	GW number	Country	War start	War end
1	615	Algeria	1992-01-01	2002-12-31
2	490	Congo, DRC	1994-01-15	1998-12-28
3	516	Burundi	1994-04-20	2005-12-24
4	484	Republic of Congo	1997-06-05	1999-12-06
5	483	Chad	1994-01-23	1998-03-09
6	517	Rwanda	1994-02-13	1998-11-27
7	404	Guinea-Bissau	1998-06-06	1999-05-06
8	450	Liberia	2000-05-01	2003-11-21
9	437	Ivory Coast	2002-09-19	2004-11-06
10	451	Sierra Leone	1991-03-23	1999-12-19

Table 1: Overview of the cases used for the statistical analysis. Start- and end-dates correspond to the first and last observations in the GED dataset for the corrsponding conflicts.

Geographic data on armed conflict

In the past decade, several data collection efforts have been started to disaggregate civil conflicts into a series of events. These events range from skirmishes to major battles or atrocities against civilians. Both the "Armed Conflict Location and Event Dataset" (ACLED) (see Raleigh and Hegre, 2005) as well as the "Georeferenced Event Dataset" (GED) (see Sundberg et al., 2011) rely on news reports that contain information on violent events primarily in Sub-Saharan Africa. GED is based on an elaborate coding procedure that ensures reliability by cross-validating records with multiple coders (Sundberg et al., 2011). Definitions of what constitutes a conflict event vary slightly between the data sets: In ACLED, violence against civilians as well as battle outcomes such as changes in territorial control are recorded. Sporadically, ACLED also has information on initiators of violence, but information on casualties is not recorded. GED is restricted to lethal encounters between political actors and provides estimates for civilian and military casualties. Information on both outcomes of battles and initiators are missing. For this study, I used the GED dataset on lethal events in African civil conflicts between 1990 and 2010. The advantage of GED for this particular project is that lethal encounters are particularly relevant and conceptually clear. In the next sections, I will introduce covariates to predict spatial variation in lethal violence in insurgencies.

Population

Based on the notion of "population-centric warfare" (see CIA, 2009, 2), civilian population concentrations have been identified as a predictor of conflict events (Raleigh and Hegre, 2009). Insurgents seek contact with the civilian population for various reasons: to hide from incumbent forces (Salehyan and Gleditsch, 2006), to recruit additional combatants (Sheehan, 1988, 50), and to extend their geographic control over relevant parts of the country (Kalyvas, 2006, 202-207). Spatially disaggregated population counts from the Gridded Population of the World dataset (GPW) (CIESIN, 2005) were therefore included in the predictive models.

Distances to capital and border

The ultimate goal of irregular uprisings is to conquer the capital city, as was the case in Saigon in 1975, in Kabul in 1996, and in Monrovia in 2003. Defending the center is therefore a strategic imperative for the state. Repeated attempts to attack the government and incumbent counteractions make distance to the capital city a spatial predictor of higher levels of violence (see Buhaug et al., 2009; Buhaug, 2010; Toellefsen et al., 2012). Along the same lines, distance to the nearest international border that provides refuge to the rebels has been associated with levels of violence (Salehyan, 2009; Buhaug, 2010). Cases in point are the Vietcong that moved their vital supply lines partially to Laos and Cambodia and the Afghan Mujaheddin that traditionally fight superpowers from bases in the border-regions in Pakistan. Distances to capital cities and international boundaries were calculated based on Weidmann et al. (2010).

Accessibility

Remote and difficult terrain provides insurgents with the opportunity to prepare attacks and temporarily evade the fighting (Fearon and Laitin, 2003). In order to counterbalance the material superiority of the state, rebels utilize less accessible areas to prepare military operations and recruit from the local population (McColl, 1969). Terrain and soil conditions, road and railroad networks, bodies of water, and forested regions all affect the accessibility of sub-national regions. A comprehensive aggregation of these factors has been performed by Nelson (2008). Their provision of a global friction map for traveling times between all cities with more than 50,000 inhabitants in the year 2000 offers a suitable operationalization for infrastructural accessibility.

Wealth

Spatial variation in wealth has been associated with conflict events (Hegre et al., 2009). Two principal scenarios are imaginable for this variable to affect levels of violence. First, materially deprived regions could see stronger support for insurgent activities. Second, rebels might strategically target wealthier regions for private gains and/or to finance the uprising. Lootable resources in particular have been linked to intense standoffs in civil conflicts (Gilmore et al., 2005). Spatially disaggregated data on wealth (Nordhaus et al., 2006) codes disaggregated GDP data on a global scale. The derived unit is Gross Cell Product (GCP): an estimate for the market value of all goods and services in a geographic region. Cells with a maximal size of 60 nautical square miles (about 111 square kilometers) are coded in this dataset, which was also included. While the exact causal roles of these geographic factors remain disputed, general correlations between the corresponding variables and levels of violence are widely accepted.

Natural land cover

Densely forested regions can be as inaccessible as high mountain ranges. Consequently, they severely limit situation awareness and mobility for regular forces (see Crawford, 1958). In Columbia, the FARC rebels have evaded defeat for almost four decades, Ugandan LRA rebels are still at large despite regional and international attempts to stop their activities, and Vietnamese rebels waged three successful campaigns against three different global powers between 1941 and 1975. In all of these cases, dense forestation has been cited as an important enabler of guerrilla actions. I therefore included a dataset that codes the percentage of green vegetation for the year 2001 on a global scale and with a spatial resolution of $1km^2$ (Broxton et al., 2014).

While the exact causal roles of these geographic factors remain disputed, general correlations between the corresponding variables and levels of violence are widely accepted. But to what extent are these factors capable of predicting the spatial variation of the intensity of violence in civil wars? As mentioned above, this paper seeks to provide an easily communicable answer to this question. The next section details the corresponding approaches to modeling and validation.

5 Modeling approach

Several possible modeling approaches spring to mind for predicting conflict events based on the presented data. Many contemporary studies of violence in civil wars draw on econometric analyses. While the breadth of econometric methodology and the rapid rate at which it advances cannot be overlooked, the analysis of inherently spatial data introduces problems. First and most importantly, the nature of the dependent variable – conflict events distributed in space – has no obvious equivalent in the econometrician's toolbox. Researchers therefore usually aggregate event counts within spatial units such as artificial grid cells and then apply count-dependent variable models (see, for example Fjelde and Hultman, 2013; Pierskalla and Hollenbach, 2013; Basedau and Pierskalla, 2014). Unfortunately, in most cases there is no empirically or theoretically informed strategy for choosing the sizes of such cells.

Of course, statistical predictions both in- and out-of-sample also hinge (to some extent) on design decisions in the spatial aggregations. This presents a serious problem for the ambition of this paper: If the claim was made that out-of-sample predictions of conflict intensity were possible based on a grid-cell approach, this finding would partly be due to a ad hoc choice of a specific cell size. Ideally, a non-parametric technique would be used for mapping conflict events to covariate information.

To address this issue, an alternative modeling approach more frequently chosen in biology and epidemiology relies on point process models (PPM). While PPMs have been applied to conflict research before (Zammit-Mangion et al., 2012), the relative novelty of this approach requires a more in-depth discussion of their properties. The next section gives an overview of PPMs and their application to multivariate inference, as well as the chosen setup for prediction, cross-validation, and extrapolation.

5.1 Point process models

Before discussing this type of model in further detail, some terminology needs to be introduced. Spatial point patterns are generally analyzed within clearly demarcated areas. These areas are referred to as "windows" and can be either artificial geometric structures or irregularly shaped polygons. In this study, the country polygons obtained from Weidmann et al. (2010) are used as observational windows with one model being fitted per country.

Statistical models of spatial point patterns have been developed for several decades and successfully applied to various fields, such as biology, geography, and criminology. One obvious quantity of interest in spatial point patterns is their intensity, defined as the expected number of points per area in a given spatial window.

The intensity of the point process can vary continuously within the window as a function of covariates or another point pattern. While the introduction of a temporal dimension provides additional challenges, PPMs are attractive alternatives to econometric models for cross-sectional analyses of conflict events. Their main advantage is that they offer empirically driven and non-parametric solutions for selecting the area around the points for aggregating covariate information. In the next section, I will introduce some implementational details of PPMs starting with underlying assumptions. After that, I will provide a closer look at fitting these models to data.

Underlying assumptions for spatial Poisson processes

Any quantitative model must strike a balance between mathematical tractability and theoretical adequacy. Very much in favor of the first requirement for this application is the spatial Poisson process, which can serve as a suitable starting point for predictive modeling. For the spatial variant of the Poisson process, two principal sub-types must be distinguished: homogeneous and inhomogeneous processes. In the case of the homogeneous spatial Poisson process, the intensity (i.e. the number of points per area) λ is uniform for the entire observational window. Of course, modeling high- or low-intensity areas within countries requires the intensity to vary as a function of covariates.

The introduction of sub-regions within the spatial window is a way to achieve this. A heuristic method for choosing subregions for a given empirical point pattern will be discussed in the next section. For each of the subregions, covariate values can be established and used to estimate marginal effects. Points per subregion are Poisson-distributed with probability mass function for natural positive numbers X and k:

$$Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

Across regions, however, the intensity of the Poisson processes may vary and the numbers of points per subregion are independent (see Baddeley, 2008, 72ff.). Applying this formalism to the study of civil conflict does justice to the strand of literature that points to the local determinants of violence. However, it omits the well-described escalatory dynamics of violence and spatial diffusion.[5] Poisson process models nevertheless serve as a point of departure for predictive purposes and do not require ad hoc design decisions about the spatial scale of point-to-point interactions.

[5]Please refer to the supplementary information for a discussion of modeling self-exciting Point processes and corresponding computational problems.

Choosing tiles for covariate information

Modeling point intensity as a function of geographic covariates requires that the points in the empirical sample are associated with the covariate information. As mentioned before, PPMs do not rely on predefined spatial units to achieve this and instead choose suitable tiles heuristically from the point pattern. As illustrated in figure 1, this is accomplished in two steps: first a number of "dummy" points are superimposed on the empirical point pattern. They are either arranged in a grid-like structure (as shown in figure 1 on the right), or are uniformly distributed at random. In a second step, the study window is divided into tiles which are either associated with dummy points or empirical ones. The tiling algorithm is usually chosen to optimally demarcate regions that are closest to the empirical or simulated points, for example by calculating Dirichlet tiles (i.e. Voronoi diagrams; see Mitchell, 1997, 233). For each of the resulting tiles, covariate information is then aggregated. Of course, the exact tiles resulting from the tessellation are still dependent to some extent on the number of dummy points in the sample and their spatial distribution. However, the great advantage of this approach is that covariate values that are subsequently used for model fitting are obtained from areas that are closer to the empirical points than the simulated dummies. This is arguably a better approach to mapping covariate information to empirical points than an arbitrary spatial grid with empirically uninformed cell size and origin.

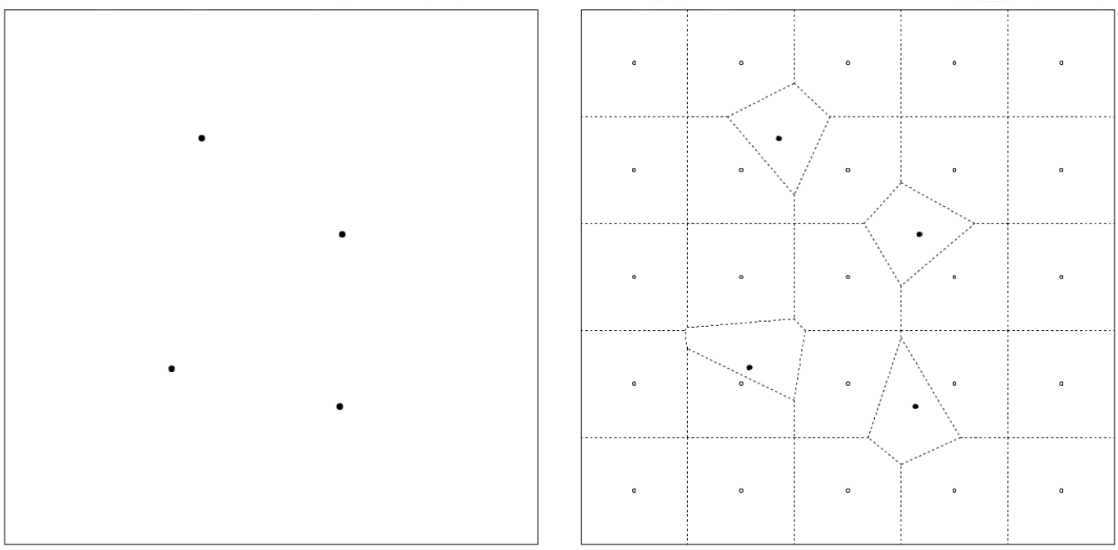


Figure 1: Illustration of a quadrature scheme based on Dirichlet tessellation (figure taken from Baddeley and Turner, 2000).

Fitting models to data

For the estimation of β-parameters, the applied tessellation techniques usually generate tiles intersecting with points in the sample and a comparable number of tiles for areas without points. A widely used approach for fitting point patterns to data relies on the Berman-Turner algorithm (see algorithm 1) which implements a maximum pseudo-likelihood approach to parameter estimation: instead of choosing parameters based on their likelihood, Berman and Turner (1992) suggest choosing them according to their conditional intensity – that is, the observed number of points per area in the tiles given the estimated intensity. Berman and Turner (1992) observe that the conditional intensity of the inhomogeneous spatial Poisson process has the same functional form as likelihood functions employed in Generalized Linear Models (GLM). This allows for a wide range of PPMs to be fitted in readily available GLM software. In detail, Berman and Turner (1992) suggest the following setup:

Algorithm 1 Berman-Turner Algorithm for Poisson process models.

1. In addition to the empirical points in the sample, generate a number of dummy points. Together with the empirical points, these are referred to as "quadrature points". Based on a tessellation scheme (Drichilet tessalation), the observational window is split up into areas associated with one quadrature point each.

2. For each quadrature area u of the spatial window W, weights are computed according to $weight_j = \frac{area(u_j)}{area(W)}$ for each $u_j \varepsilon W$.

3. For each quadrature area u, binary indicators are computed according to

   (a) $z_j = 1$ for empirical points

   (b) $z_j = 0$ for dummy points.

4. For each quadrature area u, a response variable is computed according to $y_j = z_j / weight_j$.

5. Values for the spatial covariates are obtained for each quadrature point through an intersection of the points with the underlying data $v_j = S(u_j, x)$.

6. Finally, the response variable can be estimated as $\hat{y}$ being a function of covariates v with weights *weight* in a log-linear Poisson regression.

Being able to fit models to data is a central prerequisite for assessing the predictive gain associated with individual variables. In the next section, I will introduce the remaining requirements: a general error score to compare prediction and empirical data and a cross-validation setup for out-of-sample tests.

5.2 Error score

To keep the validation approach as general as possible, I decided to simulate point patterns from the fitted prediction models and then compare them to the empirical patterns.[6] Of course, simulated point patterns vary from simulation run to simulation run as they are generated probabilistically. To establish average predictions, I simulated point patterns from the fitted models 100 times for each in-sample and out-of-sample test. A suitable method that yields a continuous non-parametric estimate for the point process was described by Diggle (1985).[7] I decided to compare simulated and empirical intensities numerically to assess the performance of different models.

While this setup might look rather cumbersome at fist glance, it essentially generates a side-by-side comparison between empirically estimated and simulated intensities. This direct comparison communicates the performance of the models in a straight-forward manner and easily generalizes to other models such as more advanced PPMs, agent-based models, or grid-cell based econometric models. Figure 2 depicts empirical and simulated conflict events, as well as corresponding intensity surfaces.

[6]The Metropolis-Hastings Algorithm otherwise familiar from Bayesian statistics is generally used to simulate point patterns from spatial probability distributions (see Baddeley, 2008).

[7]Diggle (1985) states that this method assumes that data generating mechanism to be a Cox Process (such as the spatial inhomogeneous Poisson Process) and requires a type of kernel to be specified. I used Gaussian kernels with empirically estimated bandwidth parameters in absence of any theoretical reasons to deviate from this setup.

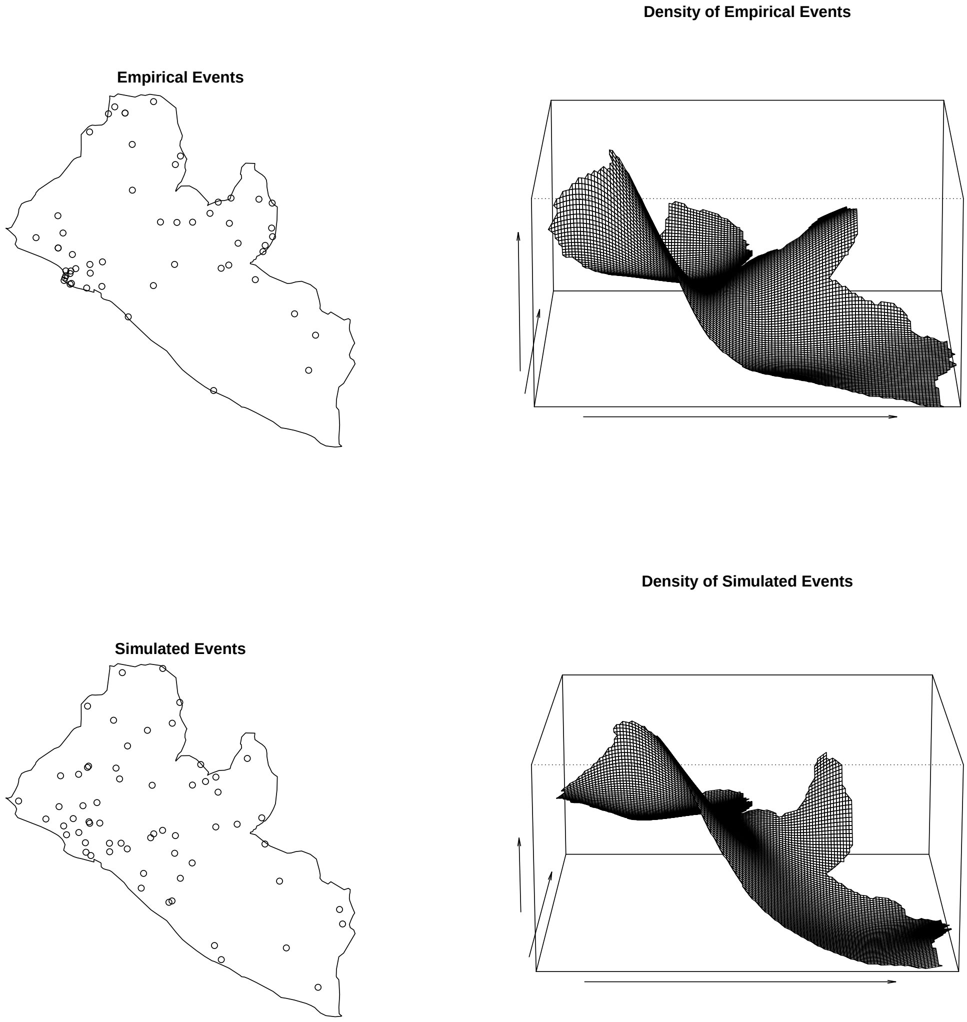


Figure 2: Example of empirical and simulated events from the Second Liberian Civil War (1999-2003). On the top left, the empirical events from the GED dataset can be seen for this conflict. On the top right, a corresponding Gaussian intensity estimation is visible. In the bottom row, simulated events and the corresponding surface are depicted.

The prediction error is computed based on the absolute differences in the densities for empirical and simulated events. Density surfaces are represented as fine-grained arrays. The mean absolute error for an array with J cells is calculated as follows:

$$MAE = \frac{\sum_1^J abs(emp_j - sim_j)}{J} \tag{2}$$

Figure 3 illustrates the comparison and the calculation of average prediction errors visually.

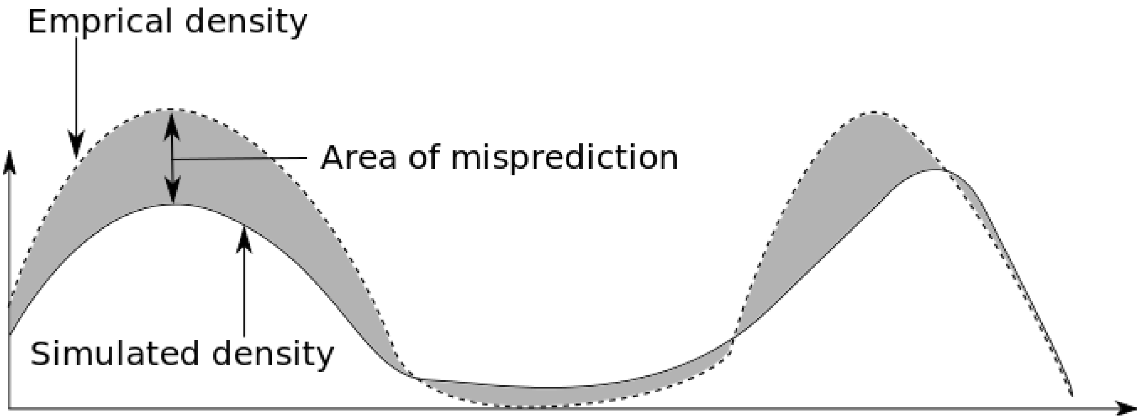


Figure 3: Depiction of the error metric used for the prediction models. The two lines are the cross sections of the point intensity estimates for empirical and simulated point patterns. The differences in densities (gray areas) are approximated numerically.

In this setup, the total number of events is still given by the empirical sample, i.e. the specific county the model was fitted on. Of course, predicting the overall number of conflict events (i.e. the severity of the civil war) is not in the focus of this study. Numerous socioeconomic, political, and military factors influence the severity of conflicts and it would be impossible to do justice to them in this article. I therefore decided to normalize the predicted intensities to one. As a result, the predictions reflect relative intensities scaled from zero to one.

5.3 Cross validation

In order to assess the predictive capabilities of the fitted models out-of-sample, a suitable cross-validation setup had to be defined as well. Basically, cross validation works by dividing the available empirical sample into a training and a test set. Models are fitted on the training set and then used to generate predictions for the test set. Those predictions are then validated against the empirically observed results from the test set. This setup serves as a more realistic test framework than simply assessing the in-sample predictions of statistical models – that is, their ability to replicate the test data they were trained on.[8]

In this case, I chose to apply a leave-one-out cross-validation scheme.[9] Models were fitted on all but one of the countries in the statistical sample. A prediction model was

[8] A typical problem that can arise in in-sample predictions is overfitting: instead of generalizing from the underlying data-generating process, an overfitted model tends to replicate the noise of the specific sample it was fitted on. Overfitting leads to low in-sample prediction errors combined with high out-of-sample prediction errors.

[9] One alternative setup frequently chosen in comparable projects is k-fold cross-validation. In this case, the small number of ten available cases allowed for using each single data point for a separate cross-validation run, thus providing the best possible approximation of the out-of-sample prediction error.

generated by averaging the $\beta-$coefficients of these models. The resulting model was subsequently used to predict the point pattern in the remaining country. This setup mimics the real-world challenge of predicting the spatial distribution of violence in future civil wars based on a set of historical conflicts. In the next section, I will present results for predictions of conflict intensity for all ten African cases.

6 Results

In order to assess the predictive capabilities of the models, I computed density differences both in-sample and out-of-sample. For the in-sample assessment, I fitted a series of Poisson models based on the introduced covariates and fitting techniques using the `spatstat` package for the `R` programming language. I simulated 100 distinct point patterns from the fitted models and calculated intensity surfaces. One expected density for each model was established by averaging over these simulations. Table 2 shows cumulative differences between the empirical density and the average simulated density. To generate a baseline against which the models could be compared, I generated 100 random point pattern consisting of an equal number of points as the empirical sample. The average difference in normalized intensities between random and empirical patterns serve as a baseline against which the predictions can be compared.

6.1 In-sample prediction error

In total, six model predictions plus the random baseline were simulated for each country. As a first test of the introduced setup and the predictive capabilities of the models, the introduced variables were added subsequently to the model specification. Acronyms in the top row of the prediction tables indicate the variables that were used: "p" stands for population, "c" for capital distance, "a" for accessibility, "w" for wealth, "b" for border distance, and "v" for vegetation. Results from a full-fledged model-averaging setup where each covariate's predictive performance is tested against a series of different model specifications are presented in the supplementary information. Table 2 shows cross-validation scores for the different model specifications and the random baseline. As discussed above, these scores are the average cumulative absolute difference between the empirical and the simulated point patterns. The last row in the table shows normalized cross-validation scores across countries with the random baseline having a value of 1. This row shows that the initial models that only use population as a predictor already yield half the cumulative error scores (.45) of the random baseline. As additional predictors are introduced, the scores drop to .26-.23. This setup also shows that not every predictor yields the same improvements. Model 2, using data on population centers and capital distances, clearly outperforms model 1, but subsequent additions of predictors only yield marginal returns.[10] Model 3 with an error score of .23 performs best in this setup. These results are encouraging as they demonstrate that the introduced data and modeling techniques can be used to replicate empirical patterns to some extent. However, the real test for the presented setup are predictions beyond the sample that the models were fitted on. Corresponding results can be found in table 3.

	Country	p(1)	pc(2)	pca(3)	pcaw(4)	pcawb(5)	pcawbv(6)	random
1	Cote d'Ivoire	0.30	0.26	0.22	0.24	0.26	0.25	0.39
2	Liberia	0.24	0.14	0.06	0.06	0.07	0.05	0.30
3	Guinea-Bissau	0.04	0.04	0.04	0.05	0.05	0.05	0.33
4	Sierra Leone	0.12	0.12	0.13	0.14	0.11	0.12	0.39
5	Algeria	0.12	0.01	0.01	0.01	0.01	0.01	0.50
6	Burundi	0.12	0.10	0.10	0.09	0.07	0.07	0.39
7	Rwanda	0.11	0.11	0.12	0.13	0.16	0.17	0.38
8	Congo	0.30	0.05	0.04	0.04	0.06	0.06	0.35
9	Congo, DRC	0.29	0.10	0.09	0.09	0.07	0.06	0.38
10	Chad	0.05	0.06	0.05	0.04	0.05	0.05	0.38
11	Sum	1.69	0.99	0.86	0.89	0.91	0.89	3.78
12	Normalized	0.45	0.26	0.23	0.24	0.24	0.24	1.00

Table 2: In-sample results based on differences between normalized empirical and simulated intensities.

6.2 Out-of-Sample Prediction

Table 3 shows cross-validation scores based on the leave-one-out cross-validation setup described in section 5.3. As one would expect, the cumulative error score across models is higher than in the in-sample setup (3.93 compared to 3.78). However, the out-of-sample predictions generally perform surprisingly well: For all but the simple population model, error scores below .3 of the normalized random baseline errors are attained. Interestingly, the lowest error scores are achieved for models 3 and 4 which only include 3-4 predictors each. The slightly lower performance of models 5 and 6 might be due to overfitting. Generally, the out-of-sample predictions work well and serve as a powerful reminder of the achievements of geographic and quantitative research on civil conflicts of the last decade. While measuring deviations between empirical and predicted densities is a good way to

[10]Please refer to the supplementary information for the β-estimates and results from a full model-averaging setup that shows each variable's contribution in a series of model specifications.

quantify the performance of prediction models, qualitative comparisons as introduced in the next section add another important angle to the empirical analysis.

	Country	p(1)	pc(2)	pca(3)	pcaw(4)	pcawb(5)	pcawbv(6)	random
1	Cote d'Ivoire	0.19	0.19	0.19	0.19	0.19	0.19	0.45
2	Liberia	0.10	0.08	0.06	0.06	0.06	0.06	0.27
3	Guinea-Bissau	0.04	0.05	0.04	0.04	0.04	0.04	0.35
4	Sierra Leone	0.21	0.21	0.20	0.20	0.19	0.18	0.38
5	Algeria	0.02	0.02	0.02	0.02	0.02	0.02	0.51
6	Burundi	0.07	0.07	0.07	0.09	0.09	0.11	0.39
7	Rwanda	0.10	0.10	0.10	0.11	0.15	0.17	0.39
8	Congo	0.09	0.09	0.09	0.09	0.10	0.10	0.35
9	Congo, DRC	0.11	0.11	0.11	0.11	0.11	0.11	0.40
10	Chad	0.41	0.16	0.07	0.07	0.08	0.08	0.44
11	Sum	1.35	1.08	0.96	0.99	1.03	1.07	3.93
12	Normalized	0.34	0.28	0.24	0.25	0.26	0.27	1.00

Table 3: Cross-validation results for predictions that were fitted on nine of the ten countries and then used to predict the remaining country.

6.3 Qualitative comparisons

Section 8 shows comparisons between empirical densities and predictions. For each of the ten countries in the sample, three plots were generated. The plot on the left show normalized intensity surfaces for the empirical patterns. The columns in the middle and on the right show model predictions based on model 3 which had the lowest error scores in the in-sample and out-of-sample predictions. In the middle column, model 3 was fitted on the country under investigation. In the right column, the cross validation model based on the estimates of the remaining cases was used to predict the country under investigation. The associated legend can be found below the country plots. As seen in section 8 on page 22, most of the out-of-sample predictions actually predict high-conflict areas. This is remarkable, as it both underscores the merit of the used datasets as well as the validity of the chosen modeling approach. These specific predictions also illustrate the merit of the technology for informing relief organizations and policy.

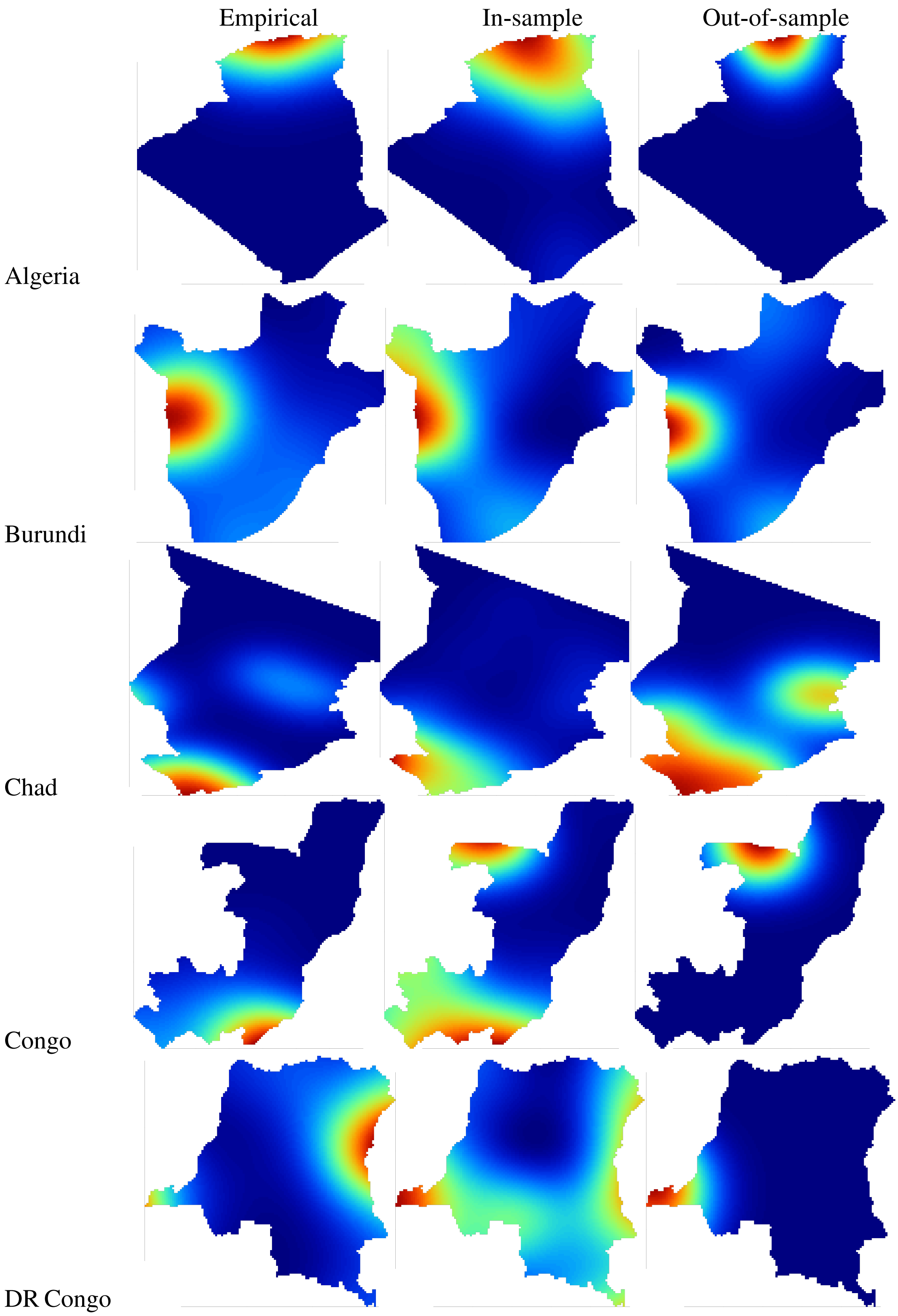
But how can we explain the fact that some conflicts are predicted quite well while others are not? A closer look at the (qualitatively) most obvious mispredictions –Ivory Coast and Republic of the Congo– can provide some answers. In the Ivorian case, the framework of a peripheral insurgency slowly advancing towards the center simply does not apply very well. Instead, ethnic and religious tensions between an predominantly Christian South and a predominantly Muslim North account for a large fraction of the conflict events (see McGovern, 2011). This entailed that the region of high-intensity fighting occurred along an East-West axis, while the model predictions place it in the northern periphery (see predictions in section 8).

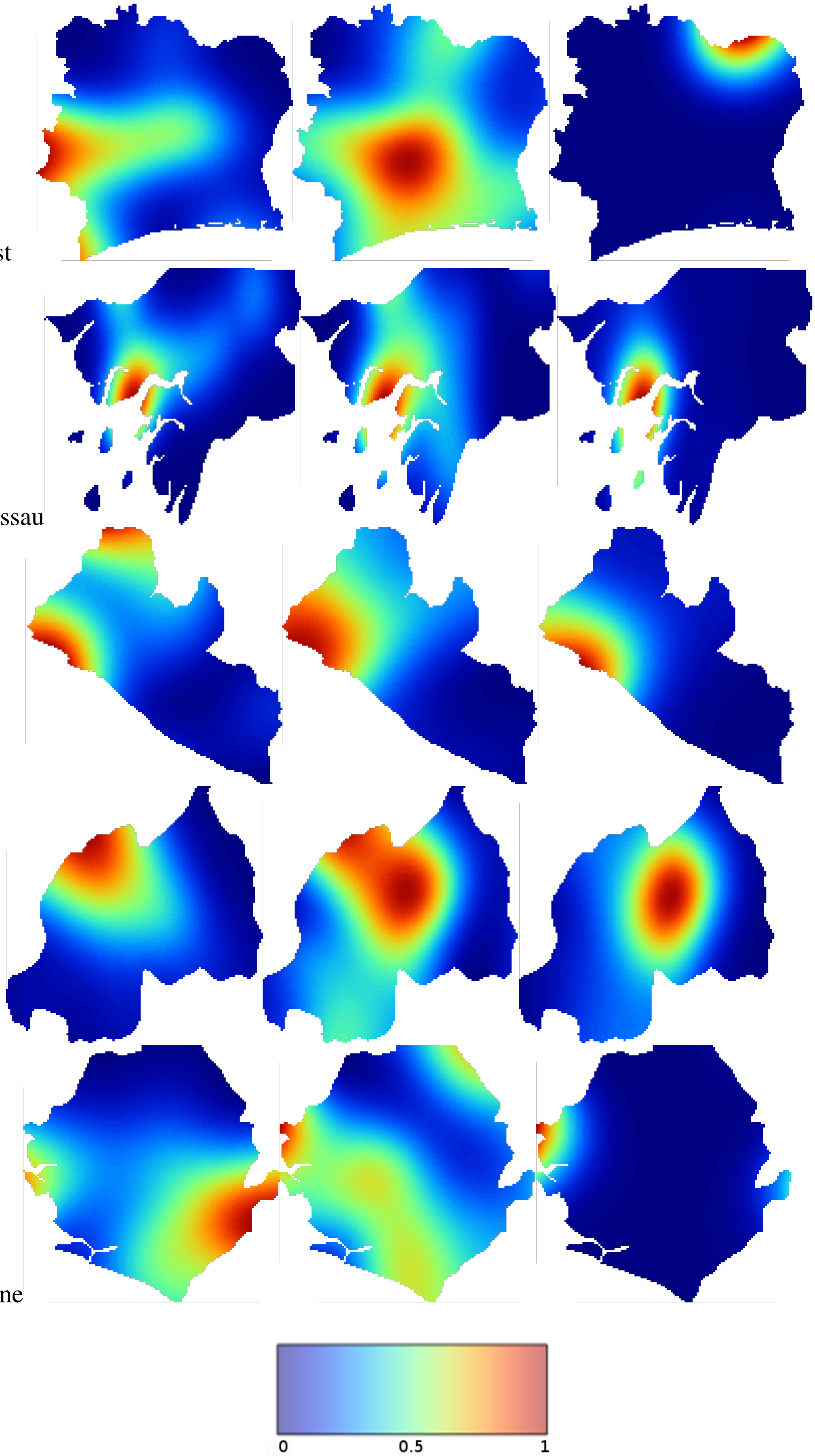
In the case of the Republic of Congo, much of the fighting happened during the height of the civil war in and around the capital city, Brazzaville. Fearing the outcome of a July 1997 Presidential election, followers of the top candidates Pascal Lissouba and Col. Denis Sassou Nguesso engaged in an armed struggle over control of the capital (see DeRouen and Heo, 2007, 129). While the uprising against then-ruling President Lissouba qualifies as a popular insurgency given the participation of numerous irregular fighters, Mao's ([1938] 1967) three-stage model for peripheral insurgencies fails to apply. Instead of building on a protracted and peripheral campaign, the warring parties opted for a conflict option that could be better described as a popular coup d'état. Instead of affecting remote areas, the fallout in political violence of this conflict clustered around the capital city in the far south of the country.

7 Discussion and Conclusion

Studies of armed conflict have identified a number of geographic conditions that correlate with guerrilla activity. Both protection from state power in terms of remoteness and the presence of strategic targets such as population centers affect the probability of armed clashes taking place in insurgencies. Causal effects of selected spatial covariates have been analyzed by a flurry of recent publications. However, established effects were only valid for specific spatial units, and only hold under cetris paribus conditions. An assessment of the external validity of this research program has been missing so far. Filling this gap, this paper has used geographic data on conflict as well as a series of theoretically prominent geographic covariates to predict the spatial distribution of conflict, both in-sample and out-of-sample. The results clearly communicate to what extent these variables actually improve predictions in direct comparison to an agnostic baseline: In-sample, cumulative error scores only amount to 25% of the cumulative error of the random baseline. In out-of-sample predictions, the error scores are slightly higher, but they still only amount to less than 30% of the random baseline error. In qualitative comparisons, the locations of high-intensity conflict zones are correctly predicted in 6 out of 10 countries. Two countries (Sierra Leone and the Democratic Republic of the Congo) have two distinct high intensity conflict areas, and only one of them is predicted correctly. In the two remaining countries (Ivory Coast and Republic of Congo) the predictions are incorrect. While more work needs to be done to identify and test predictors of violence and include more advanced modeling techniques, these results underscore the external validity of the insights generated by geo-quantitative research on civil conflicts and their potential merit for real-world applications.

8 Side-by-side comparisons of empirical densities and predictions





Ivory Coast

Guinea-Bissau

Liberia

Rwanda

Sierra Leone

Legend

References

A. Baddeley. Analysing spatial point patterns in r, workshop notes, version 3, 2008.

Adrian Baddeley and Rolf Turner. Practical maximum pseudoliklihood for spatial point patterns. *Australian and New Zealand Journal of Statistics*, 42(3):283–322, 2000.

Matthias Basedau and Jan Henryk Pierskalla. How ethnicity conditions the effect of oil and gas on civil conflict: A spatial analysis of africa from 1990 to 2010. *Political Geography*, 38(0):1–11, 2014.

M. Berman and T. R. Turner. Approximating point process likelihoods with glim. *Applied Statistics*, 41:31–38, 1992.

Ravi Bhavnani, Dan Miodownik, and Jonas Nart. Rescape: an agent-based framework for modeling resources, ethnicity, and conflict. *Journal of Artificial Societies and Social Simulation*, 11(2-7), 2008.

Ravi Bhavnani, Karsten Donnay, Dan Miodownik, Maayan Mor, and Dirk Helbing. Group segregation and urban violence. *Forthcoming in the American Journal of Political Science*, 2013.

Alex Braithwaite. *Conflict Hotspots: Emergence, Causes, and Consequences*. Ashgate, 2010.

Alex Braithwaite and Shane D. Johnson. Space-time modeling of insurgency and counterinsurgency in iraq. *Journal of Quantitative Criminology*, 28(1):31–48, 2012.

P.D. Broxton, X. Zeng, W. Scheftic, and P.A. Troch. A modis-based global 1-km maximum green vegetation fraction dataset, 2014.

Halvard Buhaug. Dude, where's my conflict? lsg, relative strength, and the location of civil war. *Conflict Management and Peace Science*, 27(2), 2010.

Halvard Buhaug and Scott Gates. The geography of civil war. *Journal of Peace Research*, 39:417–433, 2002.

Halvard Buhaug and Jan-Ketil Rød. Local determinants of african civil wars, 1970-2001. *Political Geography*, 25:315–335, 2006.

Halvard Buhaug, Scott Gates, and Päivi Lujala. Geography, rebel capacity, and the duration of civil conflict. *Journal of Conflict Resolution*, 53(4):544–569, 2009.

Halvard Buhaug, Kristian Skrede Gleditsch, Helge Holtermann, Gudrun Østby, and Andreas Forø Tollefsen. It's the local economy, stupid! geographic wealth dispersion and conflict outbreak location. *Journal of Conflict Resolution*, 55(5), 2011.

Charles Butcher. 'capital punishment': Bargaining and the geography of civil war. *Forthcoming in the Journal of Peace Research*, 2014.

CIA. Guide to the analysis of insurgency. Available online at http://www.fas.org/irp/cia/product/insurgency.pdf, 2009.

CIESIN. Gridded population of the world, version 3 (gpwv3). Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). 2005. Gridded Population of the World Version 3 (GPWv3): Population Grids. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw, 2005.

Oliver Crawford. *The Door marked Malaya*. Rupert Hart-Davis, 1958.

Karl DeRouen and Uk Heo. *Civil Wars of the World: Major Conflicts since World War II*. ABC-CLIO, 2007.

Peter J. Diggle. A kernel method for smoothing point process data. *Applied Statistics (Journal of the Royal Statistical Society, Series C)*, 34:138–147, 1985.

James D. Fearon and David D. Laitin. Ethnicity, insurgency and civil war. *American Political Science Review*, 97(1):75–90, 2003.

Hanne Fjelde and Lisa Hultman. Weakening the enemy: A disaggregated study of violence against civilians in africa. *Journal of Conflict Resolution*, pages 1–28, 2013.

David Galula. *Counterinsurgency Warfare: Theory and Practice*. Praeger Security International, 1964.

E. Gilmore, N. P. Gleditsch, P. Lujala, and J. K. Rød. Conflict diamonds: A new dataset. *Conflict Management and Peace Science*, 22(3):257–292, 2005.

Kristian Skrede Gleditsch and Michael Ward. Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, 50(1):17–31, 2012.

J. A. Goldstone, R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward. A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208, 2010.

Ernesto Guevara. *Guerrilla Warfare: Authorized Edition*. Ocean Press, 1961.

Johann Dittrich Hallberg. Prio conflict site 1989-2008: A geo-referenced dataset on armed conflict. *Conflict Management and Peace Science*, 29(2):219–232, 2012.

Håvard Hegre, Gudrun Østby, and Clionadh Raleigh. Poverty and civil war events: A disaggregated study of liberia. *Journal of Conflict Resolution*, 53(4):598–623, 2009.

Jeffrey Herbst. *States and Power in Africa - Comparative Lessons in Authority and Control*. Princeton University Press, 2000.

Jeffrey Herbst. African militaries and rebellion: The political economy of thread and combat effectiveness. *Journal of Peace Research*, 41(3):357–369, 2004.

Stathis Kalyvas. Warfare in civil wars. In Duyvesteyn I. and J. Angstrom, editors, *Rethinking the Nature of War*, pages 88–108. Abingdton, 2005.

Stathis Kalyvas. *The Logic of Violence in Civil Wars*. Cambridge University Press, 2006.

Stathis Kalyvas and Laia Balcells. International system and technology of rebellion: How the end of the cold war shaped internal conflict. *American Political Science Review*, 104(3):415–429, 2010.

Matthew A. Kocher, Thomas B. Pepinsky, and Stathis Kalyvas. Aerial bombing and counterinsurgency in the vietnam war. *American Journal of Political Science*, 55(2): 201–218, 2011.

Jason Lyall. Does indiscriminate violence incite insurgent attacks? evidence from chechnya. *Journal of Conflict Resolution*, 53(3):331–362, 2009.

Jason Lyall and Isaiah Wilson. Rage against the machines: Explaining outcomes in counterinsurgency wars. *International Organization*, 63(1):67–106, 2009.

Tse-tung Mao. *On Protracted War*. Foreign Language Press, 1967.

Robert W. McColl. The insurgent state: Territorial bases of revolution. *Annals of the Association of American Geographers*, 59(4):613–631, 1969.

Mike McGovern. *Making War in Côte d'Ivoire*. Chicago University Press, 2011.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill International Editions, 1997.

Andrew Nelson. Estimated travel time to the nearest city of 50,000 or more people in year 2000. Available online at http://bioval.jrc.ec.europa.eu/products/gam/, 2008.

William Nordhaus, Qazi Azam, David Corderi, Kyle Hood, Nadejda M. Victor, Mukhtar Mohammed, Alexandra Miltner, and Jyldyz Weiss. The g-econ database on gridded output: Methods and data, 2006.

John O'Loughlin and Frank Witmer. The localized geographies of violence in the north caucasus of russia, 1999-2007. *Annals. Association of American Geographers*, 100(3): 2379–2396, 2010.

Jan H. Pierskalla and Florian M. Hollenbach. Technology and collective action: The effect of cell phone coverage on political violence in africa. *American Political Science Review*, 107(02):207–224, 5 2013.

Gerard Prunier. *Africas World War*. Oxford University Press, 2009.

Clionadh Raleigh and Håvard Hegre. Introducing acled: An armed conflict location and event dataset, 2005.

Clionadh Raleigh and Håvard Hegre. Population size, concentration, and civil war. a geographically disaggregated analysis. *Political Geography*, 28(4):224–238, 2009.

Siri Camilla Aas Rustad, Halvard Buhaug, Ashild Falch, and Scott Gates. All conflict is local: Modeling sub-national variation in civil conflict risk. *Conflict Management and Peace Science*, 28(1):15–40, 2011.

I. Salehyan and K. S. Gleditsch. Refugees and the spread of civil war. *International Organization*, 60(2):335–366, 2006.

Idean Salehyan. *Rebels without Borders*. Cornell University Press, 2009.

Philip A Schrodt. Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 51(2):287–300, 2014.

James C. Scott. *The Art of Not being Governed*. Yale University Press, 2009.

Neil Sheehan. *A Bright Shining Lie: John Paul Vann and the USA in Vietnam*. Jonathan Cape, 1988.

Ralph Sundberg, Mathilda Lindgren, and Ausra Padskocimaite. Ucdp ged codebook version 1.5-2011. Available online at: http://www.ucdp.uu.se/ged/data.php, 2011.

Andreas Toellefsen, Håvard Strand, and Halvard Buhaug. Prio-grid: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374, 2012.

Michael D. Ward, Brian D. Greenhill, and Kristin Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):1–13, 2010.

Michael D. Ward, Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle. Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review*, 15(4):473–490, 2013.

Nils B. Weidmann and Idean Salehyan. Violence and ethnic segregation: A computational model applied to baghdad. *International Studies Quarterly*, forthcoming, 2012.

Nils B. Weidmann and Michael Ward. Predicting conflict in space and time. *Journal of Conflict Resolution*, 54(6), 2010.

Nils B. Weidmann, Doreen Kuse, and Kristian Skrede Gleditsch. The geography of the international system:the cshapes dataset. *International Interactions*, 36(1):86–106, 2010.

James E. Yonamine. Predicting future levels of violence in afghanistan districs using gdelt. Technical report, UT Dallas, 2013.

Andrew Zammit-Mangion, Michael Dewar, Visakan Kadirkamanathan, and Guido Sanguinetti. Point process modelling of the afghan war diary. *Proceedings of the National Academy of Sciences*, 2012.

Yuri M. Zhukov. Roads and the diffusion of insurgent violence: The logistics of conflict in russia's north caucasus. *Political Geography*, 31:144–156, 2012.