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

FIRST DRAFT

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.

*Paper prepared to be dragged through a cleansing fire of critical comments by Yuri Zhukov and other participants of the Peace Science Society Workshop "Disaggregation in Terrorism Studies", Philadelphia PA, October 9, 2014.

†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.

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 regional escalation coming. This episode demonstrates that areas 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 that draw 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). Moreover, the evaluation of predictions does not rely on a host of possibly violated assumptions, as is the case in more common goodness of fit measures.[2]

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

[2] Standard errors have become the central quantity of interest for the assessment of inferential results.

Moreover, instead of evaluating statistical predictions purely in-sample, this paper tests to what extent they improve out-of-sample predictions relative to a random baseline model.

The merit of this exercise is twofold: first, a direct comparison between the predictions of models and baseline serves as a reality check for the geographic research program, clearly communicating to what extent informed predictions exceed random guesses. Second, the applied methodology generates easily communicable predictions that could be utilized beyond basic research.

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. The superior ability of irregular forces to travel the desert already informed Lawrence's strategic decisions (Lawrence [1927]1998, 157). Mao ([1938] 1967, 7) assumed guerrilla warfare to be most feasible when employed in large countries (such as China) where the conventional forces of the incumbent or the invader tend to overstretch their lines of supply. In his three-stage model, Mao ([1938] 1967) stresses that rebels must initially establish bases in the periphery of their respective countries. In a second step, they gradually advance into more central areas and challenge the state in direct encounters before finally making a push for the capital city. Guevara likeweise emphasized the tactic of escaping the state's reach through the utilization of difficult terrain (Guevara, 1961, 10).

Consequently, counterinsurgency theory and conflict research have also identified geography as a decisive factor. John Paul Vann was troubled by the low degree of urbanization in South Vietnam in the early 1960s. With an estimated 85% of the population living

Standard errors do not necessarily reflect the predictive importance of explanatory variables (see Ward et al., 2010) and their correctness crucially relies on a series of assumptions. Observations must be independently and identically distributed, models must be correctly specified, must not be overfitted, and explanatory variables must not be endogenous to the dependent variable. While standard errors are convenient measures for the precision of a correlational relationship, violations of these underlying assumptions can be hard to assess and correct .

in the countryside, guerrilla recruitment could take place largely unnoticed by the central government (Sheehan, 1988, 50). McColl (1969) offered a geographic model that identified suitable territorial bases for guerrilla movements and used geographic information on terrain elevation and natural land cover to identify areas most prone to being used by guerrillas.

A classic model on the role of distance in interstate wars was presented by Boulding (1962). Boulding assumed that a state's ability to project power toward an enemy was dependent on both its military strength and the distance that separated the adversaries. His notion of a "Loss of Strength Gradient" (LSG) assumes that for every unit of distance, a certain number of personnel have to be subtracted from the fighting forces and added to the supply troops. In theory, putting this relationship into numbers allows for the exact calculation of the geographic limits in power projection. While heavily employed in the context of interstate conflict research (for example Lemke, 1995), the model has also been modified to serve in the context of civil war research (Herbst, 2004; Buhaug et al., 2008; Buhaug, 2010). Drawing on information on single conflict events, Buhaug (2010) showed that the basic premise of Boulding's (1962) "Loss of Strength Gradient" holds in domestic conflicts: a relatively more powerful incumbent can push an ongoing insurgency further into the periphery. Clashes between incumbent and insurgent occur most often where belligerents' relative strengths are equal.

Deviating from models of interstate war, several studies have also stressed the primarily local determinants of fighting in civil conflicts (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 determine much of the violence in civil conflicts (Kalyvas, 2005).[3]

Local socio-economic conditions have consequently moved into the focus of empirical studies. Drawing on spatially disaggregated data on wealth, 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,

[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).

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. Schutte and Weidmann (2011) 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; Schutte and Donnay, 2014). Again, differences in the average causal effects of indiscriminate or selective attacks on the levels of subsequent counterattacks underline that a simple local association of socioeconomic and geographic conditions and instances of conflict falls short of a full explanation.

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 capable of testing whether or not specific variables have an estimated marginal 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 regression designs, we must acknowledge that they do not produce the type of information that decision makers and relief organizations care about most. Instead, tangible predictions about the location and timing of violence are more of a concern. Consequently, political scientists have embarked on generating predictions for which countries are likely to experience civil wars (Weidmann and Ward, 2010; Gold-

stone 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 attempts 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.

Of course, assessing this predictive gain requires a suitable empirical setup. The paper proceeds as follows: in the next section, I will identify central variables for the prediction of variation in conflict intensity from the above-referenced literature. After that, a generic setup for predicting violence based on these variables is presented, 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 a baseline model.

3 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.

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 a straightforward quantity of interest. The resulting predictions therefore communicate where violence is used by military or paramilitary forces within civil wars.

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 conquering 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 borderlands to Pakistan. Distances to capital cities and international boundaries were calculated based on Weidmann et al. (2010).

Remoteness

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 offers a suitable operationalization for remoteness.

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. But to what extent are these factors capable of predicting the spatial variation in 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.

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, Ugan-

dan 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 1943 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 1km² (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.

4 Modeling approach

As described above, data on conflict events, geo-coded information on terrain, and distances to borders and cities were obtained from various sources. 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 post-hoc choice of a specific cell size. Ideally, a non-parametric technique would

be used for mapping conflict events to covariate information.

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 formal properties, their approach to multivariate inference, as well as the chosen setup for prediction, cross-validation, and extrapolation.

4.1 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. Generally speaking, the 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 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 select the area around the points for aggregating covariate information automatically and non-parametrically. Before focusing on the implementational details of PPMs, I will briefly review the assumptions underlying the models that will be used for predicting conflict events. After that, I will provide a closer look at fitting PPMs to data.

Underlying assumptions for spatial Poisson processes

Any quantitative model of data-generating processes must strike a balance between mathematical tractability and theoretical adequacy. Very much in favor of the first requirement is the spatial Poisson process, which can serve as a suitable starting point for predictive modeling. For the spatial Poisson process, a discrete number of points are created in

n-dimensional space based on a Poisson distribution with expected value and variance λ.[4] 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 λ is uniform for the entire observational window. For my analysis each observational window represents one country experiencing civil conflict. Of course, this renders it ineligible for modeling high- or low-risk areas within countries. But it can serve as an agnostic baseline model against which more eligible predictions can be compared.

The introduction of subregions within the spatial window is a way to model variations in point intensity across space. 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 still Poisson-distributed. 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.

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 or Voronoi diagrams. For

[4]The corresponding probability mass function is $Pr(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, for natural positive numbers X and k.

[5]Cox processes that generate offspring events from initial events will be included in future iterations of this paper.

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.

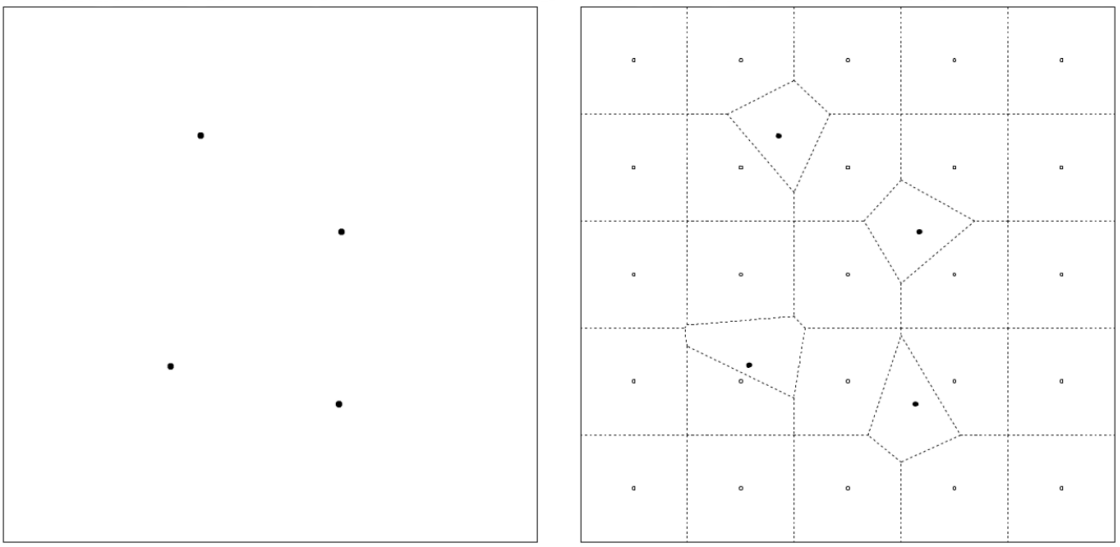


Figure 1: Illustration of a quadrature scheme based on Dirichlet tessellation (figure taken from Berman and Turner, 1992)

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 generally 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

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, such as Drichilet tessalation or a Voronoj diagram, the observational window is split up into areas associated with one quadrature point each.

2. For each quadrature area j, weights are computed according to $w_j = \frac{area(u_j)}{area(W)}$ for each $u_j \varepsilon W$.

3. For each quadrature area j, 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, a response variable is computed according to $y_j = z_j / w_j$.

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 w in a log-linear Poisson regression.

For the problem at hand, a suitable type of model must be selected. A wide variety of models can be fitted with this generic setup. As a staring point, I decided to model the spatial distribution of conflict events as a spatial inhomogeneous Poisson process, while more complex models will be added in future iterations of this paper.

4.2 Cross validation

In order to assess the predictive capabilities of the fitted models, a suitable cross-validation setup must be defined. 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 with the test sample. 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.[6]

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

[6] 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.

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 generally mimics the real-world challenge of predicting where major conflict zones will emerge in future civil wars based on a set of historical conflicts.

A remaining challenge is comparing the empirical pattern to the model's predictions. A simple and very general validation approach is to simulate point patterns from the spatial probability distribution of the PPMs.[7]

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.[8] Of course, simulated point patterns vary from simulation run to simulation run. To establish average densities, I simulated point patterns from the prediction model 100 times for each country in the cross-validation sample. The density estimates were obtained non-parametrically from Gaussian kernels with empirically estimated bandwidths. This approach easily generalizes for more advanced PPMs, but could be used with any framework that predicts events at distinct geo-locations, such as agent-based models, or grid-cell based econometric models. Figure 2 depicts empirical and simulated conflict events, as well as corresponding density surfaces.

The prediction error is computed based on the absolute differences in the densities for empirical and simulated events.[9] Figure 3 illustrates the comparison and the calculation of average prediction errors visually. In the next section, I will briefly discuss the case selection for the predictive study. After that, I will present prediction results for major conflict zones in ten African insurgencies.

5 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

[7] An alternative approach would have been likelihood-based statistics. The problem here is that likelihood-based inference is only valid if the underlying modeling assumptions are met. For Poisson processes, this cannot be assumed as point-to-point interaction is generally omitted. Simulating from the probability distribution and comparing simulated patterns to empirical ones is therefore a more general validation strategy.

[8] The Metropolis-Hastings Algorithm otherwise familiar from Bayesian statistics is generally used to simulate point patterns from spatial probability distributions.

[9] Density surfaces are represented as fine-grained arrays. The mean average error for an array with J cells is $MAE = \frac{\sum_1^J abs(emp_j - sim_j)}{J}$.

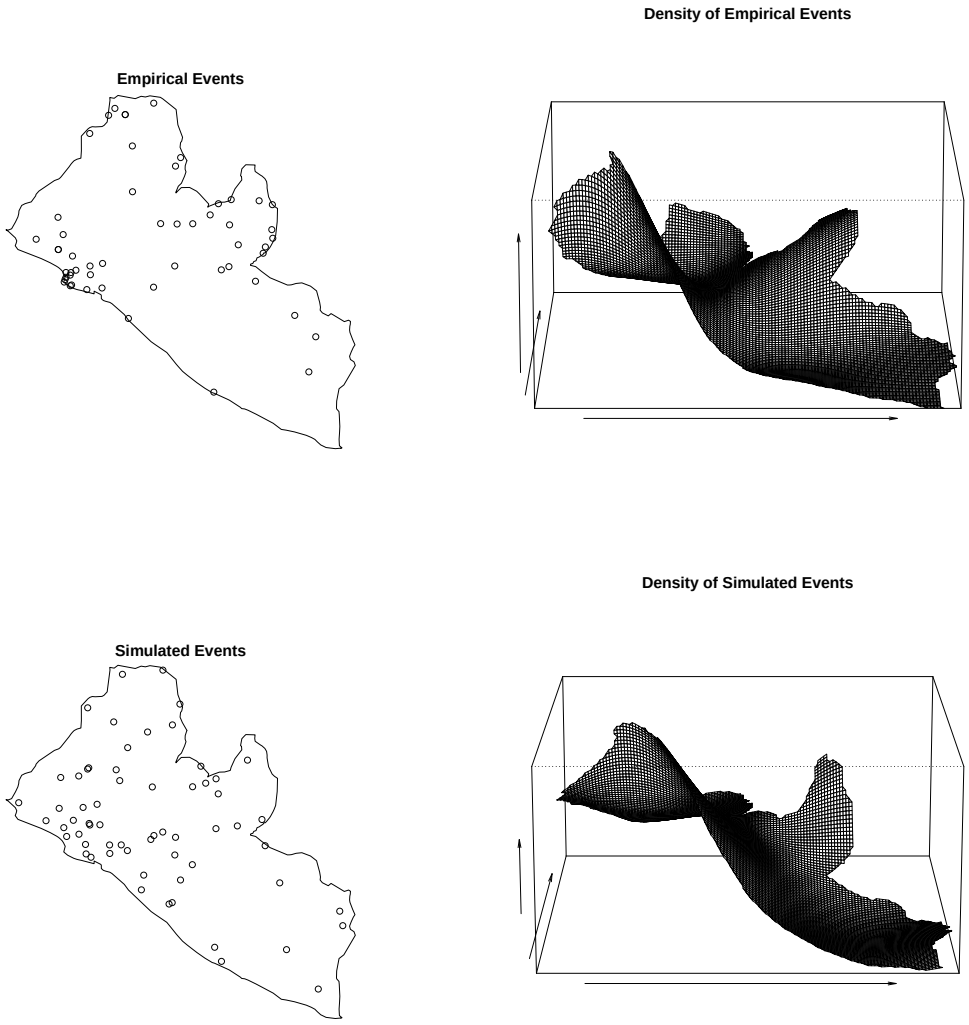


Figure 2: Example of empirical and simulated events from the Second Liberian Civil War (1999-2003). On the top left, the empirical events from the GED dataset can be seen for this conflict. On the top right, a corresponding Gaussian density surface is visible. In the bottom row, simulated events and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

rebels are not recognized as belligerents and heavily rely on civilian assistance to wage a guerrilla war against the state. However, not all civil or irregular conflicts take on the form of 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 conflict in the post-World War II period. In the GED sample, lethal clashes from 42 African countries are coded for the period 1990 to 2010. Drawing on a separate dataset by Lyall and Wilson (2009), I identified 11 cases of insurgency that are

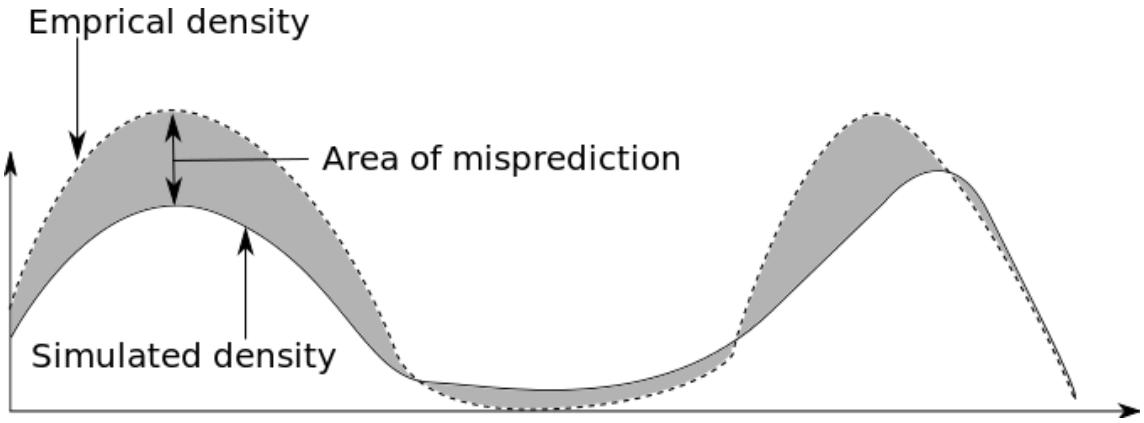


Figure 3: Depiction of the error metric used for the prediction models. The two lines are the cross sections of the density surfaces for empirical and simulated point patterns. The differences in densities are approximated numerically.

covered in the GED. [10] I decided to exclude Djibouti because the country is too small for meaningful statistical analysis given the resolution of the covariates. Table 1 shows the remaining cases that were used for the analysis.

No.	GW number	Country	War start	War end
1	615	Algeria	1992-01-01	2002-12-31
2	490	Democratic Republic of Congo (Zaire)	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.

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 density surfaces. I established one expected density for each model by averaging over these simulations. Table 2 shows cumulative

[10] 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).

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 differences in densities between random and empirical points serve as a baseline against which the predictions can be compared.

6.1 In-sample prediction error

In total, seven PPMs plus the random baseline were estimated 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. Instead of a full-fledged model-averaging setup where each covariate's predictive performance is tested against a series of different model specifications, I decided to simply add covariates sequentially to the model specification for the first version of the paper. 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 (.47) of the random baseline. As additional predictors are introduced, the scores drop to .28-.24. 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.[11] 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.

[11] Future iterations of this paper will yield a more comprehensive model-averaging setup to cleanly compare predictive improvements across predictors.

	Country	p(1)	pc(2)	pca(3)	pcaw(4)	pcawb(5)	pcawbv(6)	pwab(7)	random
1	Cote d'Ivoire	0.29	0.28	0.23	0.21	0.23	0.18	0.25	0.43
2	Liberia	0.22	0.15	0.06	0.07	0.05	0.04	0.05	0.27
3	Guinea-Bissau	0.04	0.04	0.04	0.05	0.05	0.06	0.04	0.16
4	Sierra Leone	0.12	0.11	0.13	0.13	0.12	0.12	0.13	0.38
5	Algeria	0.12	0.01	0.01	0.01	0.03	0.02	0.06	0.51
6	Burundi	0.12	0.10	0.10	0.09	0.08	0.10	0.08	0.39
7	Rwanda	0.11	0.11	0.12	0.13	0.16	0.16	0.15	0.38
8	Congo	0.33	0.05	0.04	0.04	0.05	0.05	0.07	0.34
9	DR Congo	0.31	0.10	0.09	0.09	0.08	0.08	0.11	0.37
10	Chad	0.05	0.06	0.05	0.05	0.06	0.05	0.05	0.40
	Sum	1.71	1.01	0.87	0.87	0.91	0.86	0.99	3.62
	Normalized	0.47	0.28	0.24	0.24	0.25	0.24	0.27	1.00

Table 2: In-sample results

6.2 Out-of-Sample Prediction

Table 3 shows cross-validation scores based on the leave-one-out cross-validation setup described in section 4.2. As one would expect, the cumulative error score across models is higher than in the in-sample setup (3.95 compared to 3.62). However, the out-of-sample predictions generally perform surprisingly well: For all but the simple population model, CV scores below .3 of the normalized random baselines models are attained. Interestingly, the lowest error scores are attained for models 3, 4, and 7, 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 for yet-unseen conflicts work surprisingly well and serve as a powerful reminder of the achievements of geographic and quantitative research on civil conflicts of the last decade. Measuring deviations between empirical and predicted densities is a good way to quantify the performance of prediction models. Another way to assess the quality of the predictions are qualitative comparisons as introduced in the next section.

	Country	p(1)	pc(2)	pca(3)	pcaw(4)	pcawb(5)	pcawbv(6)	pwab(7)	random
1	Cote d'Ivoire	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.42
2	Liberia	0.11	0.07	0.06	0.06	0.05	0.05	0.06	0.31
3	Guinea-Bissau	0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.35
4	Sierra Leone	0.21	0.21	0.20	0.20	0.20	0.19	0.20	0.39
5	Algeria	0.02	0.02	0.02	0.02	0.09	0.09	0.01	0.51
6	Burundi	0.07	0.07	0.07	0.09	0.11	0.11	0.07	0.39
7	Rwanda	0.10	0.10	0.10	0.11	0.15	0.16	0.11	0.40
8	Congo	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.35
9	DR Congo	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.38
10	Chad	0.38	0.15	0.10	0.08	0.08	0.08	0.07	0.44
	Sum	1.35	1.07	0.99	1.00	1.12	1.12	0.95	3.95
	Normalized	0.34	0.27	0.25	0.25	0.28	0.28	0.24	1.00

Table 3: Cross-validation results

6.3 Qualitative comparisons

Section 7 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 density surfaces for the empirical patterns. The columns in the middle and on the right show model predictions based on model 7. In the middle column, model 7 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 densities have no associated legends, as they are all normalized to 1. The qualitative comparisons only serve to test whether the high intensity conflict areas (shown in yellow and red) can be approximately correctly predicted. As seen in section on page 21, 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.[12]

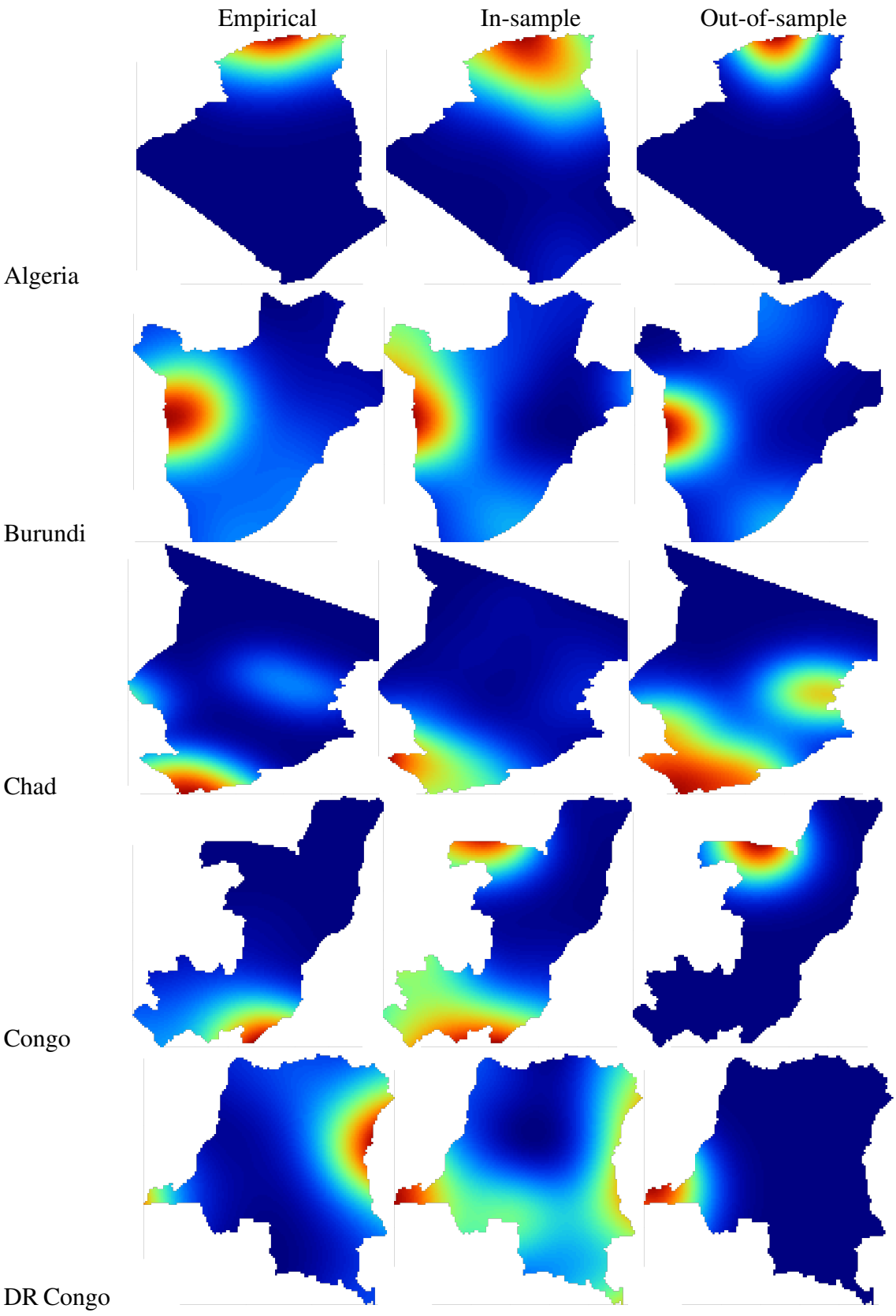
7 Discussion and Conclusion

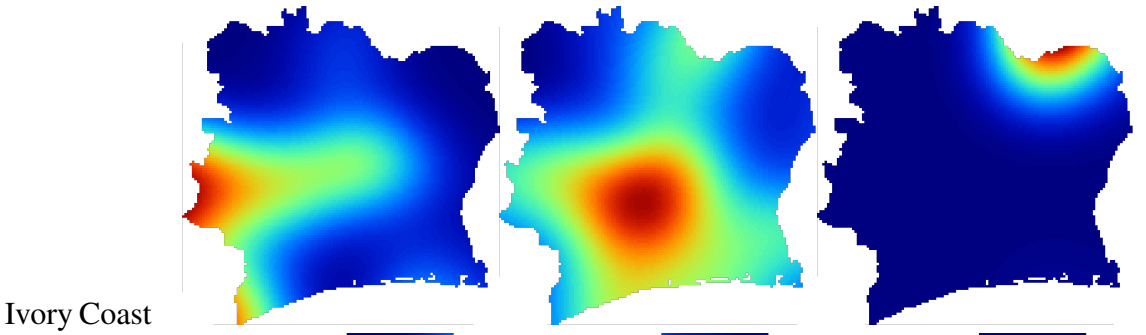
Counterinsurgency studies and contemporary studies of armed conflict have identified a number of geographic conditions that correlate with guerrilla activity. Both protection from state power in terms of remoteness and strategic targets such as population centers affect the intensity of clashes in insurgencies. 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 spatial covariates to predict the spatial distribution of conflict, both in-sample and out-of-sample. The results clearly communicate to what extent geographic 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 coun-

[12] Future iterations of this paper will present more detailed case studies for the mispredicted cases and discuss possible explanations for variation in prediction performance across countries.

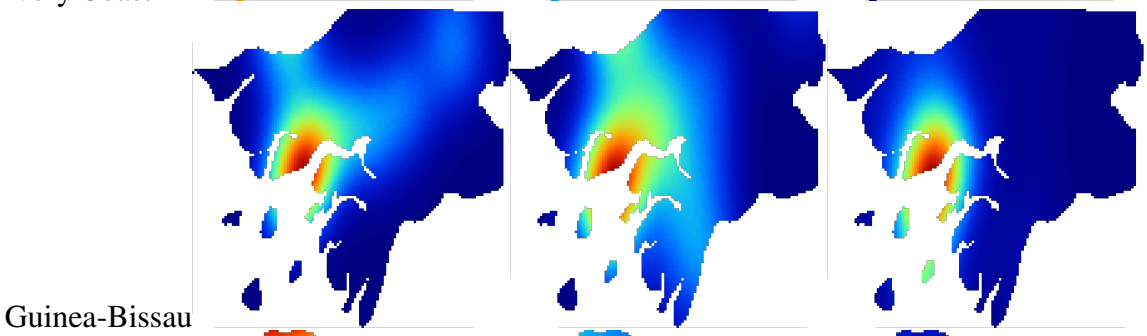
tries (Sierra Leone and the Democratic Republic of the Congo) have two distinct high intensity conflict areas, and only one of them is predicted correctly. In the two remaining countries (Ivory Coast and Republic of Congo) the predictions are incorrect. While more work needs to be done to identify and test predictors of violence and include more advanced modeling techniques, these preliminary results underscore the external validity of the insights generated by geo-quantitative research on civil conflicts and their merit to real-world applications.

Side-by-side comparisons of empirical densities and predictions

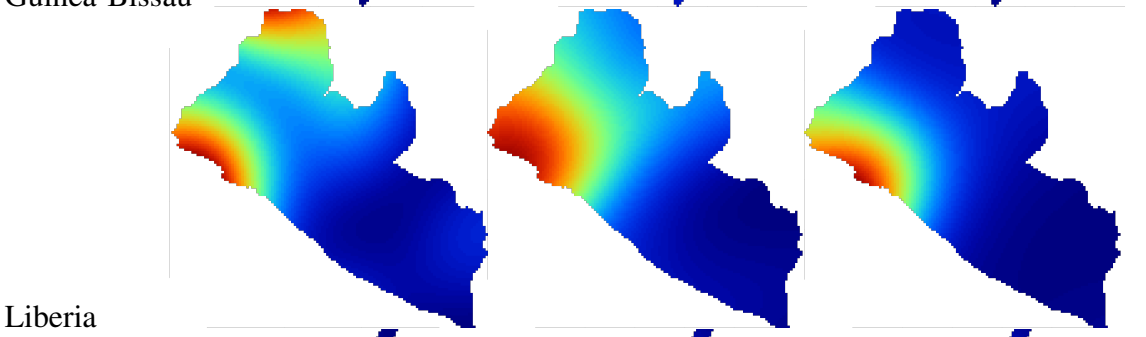




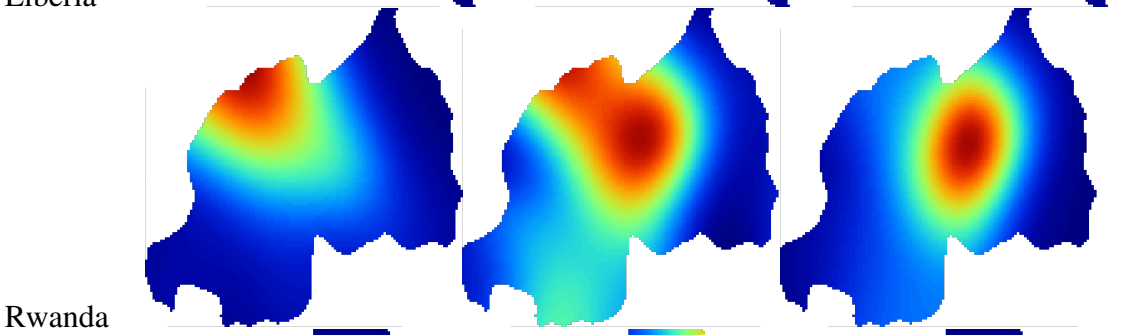
Ivory Coast



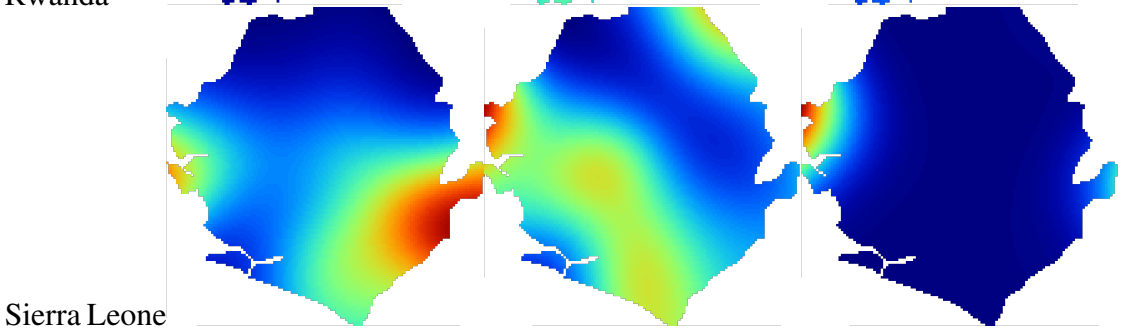
Guinea-Bissau



Liberia



Rwanda



Sierra Leone

References

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

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.

Kenneth Boulding. *Conflict and Defense: A General Theory*. Harper, 1962.

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, Lars-Erik Cederman, and Jan Ketil Rød. Disaggregating ethnonationalist civil wars: A dyadic test of exclusion theory. *International Organization*, 62(3):531–551, 2008.

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.

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.

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.

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.

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.

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.

Thomas E. Lawrence. *Revolt in the Desert*. Combined Publishing, 1998.

Douglas Lemke. The tyranny of distance: Redefining relevant dyads. *International Interactions*, 21(1):23–38, 1995.

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.

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.

Sebastian Schutte and Karsten Donnay. Matched wake analysis: Finding causal relationships in spatiotemporal event data. *Political Geography*, 41:1–10, 2014.

Sebastian Schutte and Nils B. Weidmann. Diffusion patterns of violence in civil war. *Political Geography*, 30(3):143–152, 2011.

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.