Political Science Research and Methods (2022), **10**, 97–115
doi:10.1017/psrm.2020.16

ORIGINAL ARTICLE

A new geography of civil war: a machine learning approach to measuring the zones of armed conflicts

Kyosuke Kikuta*

Osaka School of International Public Policy, Osaka University, 1-31 Machikaneyamacho, Toyonaka, Osaka, 560-0043, Japan
*Corresponding author. Email: kikuta@osipp.osaka-u.ac.jp

(Received 22 April 2019; revised 24 November 2019; accepted 8 January 2020; first published online 13 May 2020)

Abstract

Where do armed conflicts occur? In applied studies, we may take ad hoc approaches to answer this question. In some regression studies, for instance, a single conflict event can cause an entire province to be classified as a conflict zone. In this paper, I fill this void of knowledge by developing a machine learning method that is less dependent on the areal-unit assumptions and can flexibly estimate conflict zones. I apply the method to a conflict event dataset and create a new dataset of conflict zones. A replication of Daskin and Pringle (2018, *Nature* 553, 328–332) with the new dataset indicates that the effect of civil war on mammal populations is much smaller than the original estimate.

Key words: Armed conflict; conflict zones; machine learning

Where does armed conflict occur? Despite the plethora of subnational studies on civil war, we still lack clear answers to this question, which we may think of as a mere nuisance. In a number of regression studies, for instance, scholars use specific areal units, such as administrative boundaries or grid cells, and assume that the presence of a combatant event means that the entire unit is a conflict zone. These areal assignments are so common that we may not recognize that they are in fact assumptions. For example, a number of studies using the PRIOGRID (Tollefsen *et al.*, 2012) assume that if one or more events occur in a grid cell, the entire 55-km-by-55-km cell would be affected by the conflict (Buhaug *et al.*, 2011; Pierskalla and Hollenbach, 2013; Fjelde and Hultman, 2014). Other scholars use large administrative units, such as provinces (Cunningham and Weidmann, 2010; Fjelde and von Uexkull, 2012; Ritter and Conrad, 2016), and rely on a similar set of assumptions. Although these studies carefully defend their choices of areal units and measurements, none check the robustness of their findings with alternative units.[1]

The areal-assignment assumptions are, however, consequential for our understanding of civil war. As an example, the following figure (Figure 1), maps the zones of the Somali Civil War (1989–2017) made by the different areal-unit assignment rules but with the same dataset of conflict events (UCDP GED; Sundberg *et al.*, 2010). If one assigns the conflict events to the grid cells (PRIOGRID; Tollefsen *et al.*, 2012; red dotted polygons in Figure 1), the conflict zones tightly fit the conflict event locations (dot points in Figure 1). In contrast, if one uses the second-order administrative units (districts; blue dot-dashed polygons in Figure 1), the conflict zones grow to include lands within Somalia. The UCDP Polygons Dataset (Croicu and Sundberg, 2012; yellow dashed polygons in Figure 1)—a commonly used conflict zones dataset—indicates an even

[1]This is despite the long-standing attention to the so-called modifiable areal unit problem (Buhaug and Lujala, 2005).

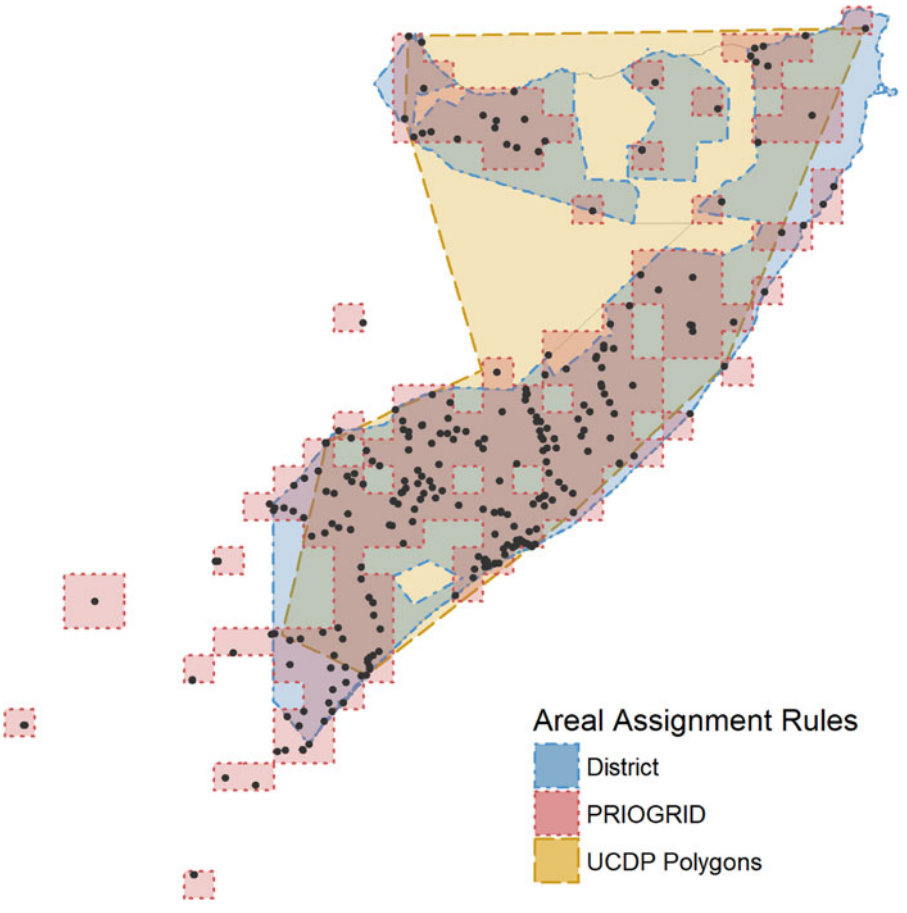

Figure 1. Zones of the Somali Civil War.
Note: The figure maps conflict zones of the Somali Civil War (1989–2017) created by existing zoning methods. All of the results are created from the same dataset of conflict events (UCDP GED; black dots).

larger area that includes Ethiopia's Ogaden region, which has no record of conflict events. The profound differences in how conflict zones can be defined from a given set of underlying data suggest that empirical findings may be sensitive to the choice of areal unit. How can we define conflict zones in a way that is less dependent on areal-unit assumptions?

I argue that the extant zoning methods rest on strong assumptions about the areal assignment of war zones, which can potentially result in misleading pictures. I demonstrate this by formalizing a zone as a summary function that maps locations and (if necessary) other substantive information onto the presence/absence of conflict events. From this perspective, these approaches not only impose strong constraints on the zoning function, but also assume that the mapping has no stochastic error. However, since a conflict zone is a function, we can readily apply statistical methods to estimate the zones.

Statistically estimating conflict zones presents a special challenge, however; while we can observe the presence of conflict events and their locations, we do not have direct observations about their absence. Although one might consider that the lack of recorded conflict events within particular geographical boundaries—such as grid cells or administrative units—would constitute absence data, the construction of the absence data is not as straightforward as one might think. Importantly, it requires pre-defined areal units, and the results may differ depending on which areal units one uses. Furthermore, since locations near conflict events are less likely to be "real" absence observations than locations farther from the events, one might also need to build a sampling scheme that accounts

for the spatial heterogeneity. However, all of these procedures require additional assumptions that make the zoning exercise sensitive to researchers' arbitrary choices. Ideally, we would like to estimate conflict zones without relying on the pseudo-absence data or pre-defined areal units.

In this paper, I address these problems by using the one-class support vector machine (OCSVM), which is an unsupervised machine learning method commonly used for outlier detection (Schölkopf *et al.*, 2000). Unlike other methods, the OCSVM requires only presence data, allowing us to estimate the conflict zones even without any pre-defined areal units. Even though the OCSVM does not use absence data and is less powerful than other statistical methods, it allows us to construct conflict zones with fewer assumptions and is therefore suitable for creating data infrastructure for broader application. In order to provide such infrastructure, I apply the OCSVM to the UCDP GED and create a new dataset of conflict zones. With this new dataset, I replicate Daskin and Pringle's (2018) study on civil wars' effect on wildlife. The results suggest that the actual ecological costs of civil war are much smaller than the original estimate.

1. A conflict zone as a representation

I consider a conflict zone to be a concise *representation* of the geographic distribution of conflict. A "population" conflict zone is an area within which conflict takes place and thus generates conflict events. An "estimated" conflict zone, by contrast, is an area in which conflict is likely to take place given our observations of conflict events. In reality, the population conflict zone may not exist; we cannot draw a line such that conflict takes place one millimeter inside of it, while conflict does not exist one millimeter outside of it. Thus, as is common in structural parameter estimation (such as the utility maximization theory of a logistic regression), the data generation process should be considered a theoretical construct. The key question is not whether the population conflict zones are "true", but whether they are *useful* for specific purposes.

Conflict zones are useful for certain purposes. For instance, by having conflict and non-conflict zones, we can directly compare human, economic, and environmental costs of civil war inside and outside of the conflict zones (Ghobarah *et al.*, 2003; Daskin and Pringle, 2018). Moreover, the conflict zones can be used for the purpose of issuing travel advisories. In fact, as can be seen in the travel advisory maps, it is more helpful to display zones of high risk instead of the precise locations of violent events. Finally, the method that this paper proposes can be potentially used for other mapping exercises, such as poverty maps, crime zones, state controls over territories, hazard maps, and zones of racial segregation, all of which have substantive applications.

This paper is agnostic with respect to the definitions of "conflict" and "conflict events." I assume that conflict events are presented as point locations,[2] and that the term "conflict" in "conflict events" and "conflict zones" have the same meaning, but this study does not depend on a particular definition of conflict. Since there are a number of studies about the concepts of armed conflict (Sundberg *et al.*, 2010), violence (Kalyvas, 2006), civil war (Sambanis, 2004), peace (Campbell *et al.*, 2017), and territorial controls (Tao *et al.*, 2016; Anders *et al.*, 2017), I focus on the concept of a zone and ask readers to refer to those studies.

Finally, this paper is primarily interested in a binary measure of conflict zones.[3] Although continuous indicators of conflict risks might be more nuanced and useful for some purposes (Anders *et al.*, 2017; Campbell *et al.*, 2017), a dichotomous zone has at least one clear advantage: providing a new geographical unit of analysis that allows us to compare conflict and non-conflict areas. What motivates this study is not to estimate a "true" distribution of conflict events or to create as precise as possible description of events. Rather, the goal is to provide a concise representation of conflict as a part of data infrastructure.

[2]A conflict event can also be expressed by a polygon, such as an administrative unit. Extending the following theory and methods to polygonal conflict event data is relatively straightforward.

[3]The proposed OCSVM method can easily be extended to multinomial or ordered outcomes.

1.1 Formalizing a conflict zone

This paper makes a conceptual shift in the geography of civil war; I conceptualize locations as *predictors* of conflict instead of *units* of analysis. This conceptualization allows me to create a new areal unit – a conflict zone – without assuming any prior areal units. Consider a set of conflict events, $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n : y_i = 1 \text{ for } i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is a vector of longitude and latitude (and if necessary other predictors) of an event i,[4] which I call a *location*, and y_i is an indicator of the presence and absence of conflict. A *zoning function* f_Y is a function that maps every location on the earth to the sample space of Y,

$$f_Y : \boldsymbol{X} \rightarrow S_Y \text{ for } \boldsymbol{x} \in G,$$

where G is the entire surface of the globe. Intuitively, as seen in Figure 2, a zoning function tells us whether each location belongs to a zone of a certain conflict. A *conflict zone* is an uncountable set of locations, $A_c = \{\boldsymbol{x} \in G : f_Y(\boldsymbol{x}) = 1\}$, and a *non-conflict zone* is its complement, $A_{\neg c} = \{\boldsymbol{x} \in G : f_Y(\boldsymbol{x}) = 0\}$. Our goal is therefore to estimate a zoning function that approximates the population zoning function and hence best summarizes the conflict events.

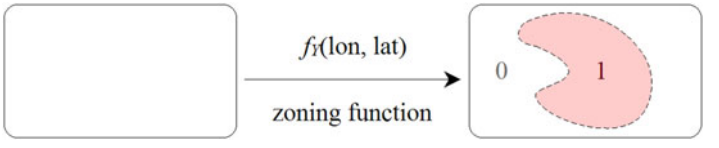

Figure 2. Zoning function.
Note: The figure shows a stylized example of a zoning function that maps every location to a conflict zone ($Y = 1$; red area) and non-conflict zone ($Y = 0$; remaining white area).

One advantage of this formalization is that we can now define the *fitness* of zoning. Let $\tilde{f}_Y$ be the population zoning function and $\hat{f}_{Y|\boldsymbol{X}}$ be a zoning function estimated from data. The population zoning function represents the underlying data generation process of conflict events, while the estimated zoning function is our estimate of the data generation process. The difference between the population and estimated zoning functions is then defined by a loss function $L(\tilde{f}_Y, \hat{f}_{Y|\boldsymbol{X}})$. Our objective is therefore to find $\hat{f}_{Y|\boldsymbol{X}}$ that minimizes the expected value of the loss function, $E_{\boldsymbol{X}}[L(\tilde{f}_Y, \hat{f}_{Y|\boldsymbol{X}})]$. Under certain conditions (Friedman, 1997; Valentini and Dietterich, 2004), the expected loss function is decomposed into bias and variance terms;

$$E_{\boldsymbol{X}}[L(f_Y, \hat{f}_{Y|\boldsymbol{X}})] = E_{\boldsymbol{X}}[g(\underbrace{L_1(f_Y, E_{\boldsymbol{X}}[\hat{f}_{Y|\boldsymbol{X}}])}_{\text{bias}}, \underbrace{L_2(\hat{f}_{Y|\boldsymbol{X}}, E_{\boldsymbol{X}}[\hat{f}_{Y|\boldsymbol{X}}])}_{\text{variance}})],$$

where g is a generic function that is increasing with L_1 and L_2. The L_1 term represents a systematic difference between the population and estimated zoning functions (*bias*), while the L_2 term indicates how random noise can alter our estimate (*variance*). When a zoning function is too inflexible and thus *underfitted* to data, the zoning function is heavily influenced by our assumptions, resulting in a large bias. By contrast, when a zoning function is *overfitted* to data, the function is extremely sensitive to random noise, indicating a large variance. Thus, estimating the population zoning function requires striking a delicate balance between bias and variance.

1.2 Fitting problems in deterministic methods

From the bias-variance perspective, deterministic methods of zoning like those commonly used in conflict studies are suboptimal. In fact, they tend to risk *both* underfitting and overfitting. Because those methods impose relatively strong constraints on the zoning function, the estimated

[4]The examples include the dates of conflict events and geographical characteristics.

zoning functions are dependent on those assumptions and potentially biased (unless those constraints were in fact correct). For instance, although we might use simple polygon assignment rules, such as assigning an administrative unit polygon as part of a conflict zone if it contains one or more conflict event, this method presumes the following functional form;

$$\hat{f}_{\text{polygon}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{x} \in P_{\text{conflict}} \\ 0 & \text{otherwise} \end{cases},$$

where P_{conflict} is a set of polygons that have at least one conflict event; if, for example, there is one or more conflict events at the eastern border area in Ogaden, the entire Ogaden region is assumed to be affected by the conflict. Although the UCDP Polygons (Croicu and Sundberg, 2012) take a more sophisticated approach (called a convex hull method), it also assumes that the shapes of conflict zones are convex, which may not always be realistic. In the case of the Somali Civil War, for instance, the convex hull method cannot account for the concave shape of Somalia (yellow dashed zone in Figure 1), resulting in a conflict zone that mistakenly includes the Ethiopian Ogaden region (despite the fact that no conflict event is reported in Ethiopian Ogaden).

Even worse, because the deterministic rules do not account for stochastic errors in our observations,[5] they also tend to overfit the data. For instance, if there is a single combat event in a far distant location (say, bombing in Paris by the combatants of the Sri Lankan Civil War), the polygon assignment method treats the surrounding areas as a part of the conflict zone. Thus, even if the deterministic approaches were to minimize the differences between the zoning function and *observed* data, the zoning function may not be optimal.

Fortunately, we can avoid these shortcomings by using statistical methods. With statistical learning methods, we can assume a fairly flexible zoning function and systematically account for random errors. An easy way to understand the statistical approach is a logistic regression (even though it is not flexible); one could estimate a logistic regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ and then use the estimated model as a zoning function. However, as I discuss in the next section, extending statistical methods to the zoning problem is not as straightforward as one might expect.

2. Statistical approaches to zoning: problems of presence-only data

A methodological challenge is that even though we have data on the presence of conflict events, we do not have direct observations about the absence of conflict events. As a result, y_i *always takes a value of 1 in our sample*, and thus conventional methods, such as logistic regression, cannot be used without further innovations. Although the presence-only data do not draw much attention and are rarely recognized as a problem in political science, this problem arises in other fields, including the conservation sciences (Mack and Waske, 2017), genetics (Mei and Zhu, 2015), and text analyses (Lee and Liu, 2003).

2.1 Positive-absence (PA) data approach

The most straightforward approach is the positive-absence (PA) data methods. The idea is that we "make up" absence data and then apply conventional classification methods. To create the pseudo-absence data, one might assign areal units, such as grid cells or administrative boundaries, to the conflict events and then treat the remaining areal units as absence data. Alternatively, one could build more sophisticated sampling schemes that account for spatial relationships (Mei and Zhu, 2015). Once s/he creates absence data, a variety of classification methods are readily available.

A drawback to the PA approach is its sensitivity to the absence-data generation. Researchers must specify the areal units or sampling schemes, and it is well known that those choices can

[5]The UCDP Polygons dataset uses so-called the 20–5 percent rule to drop outliers (see Croicu and Sundberg (2012) for details). However, even without such a rather arbitrary rule, one can readily use statistical tools to remove outliers.

greatly influence the estimates (Phillips *et al.*, 2009). Even worse, because both estimation and cross-validation rest on the pseudo-absence data, there is no established way to evaluate different absence-data sampling schemes. Thus, without strong substantive reasons to justify particular methods of absence-data generation, it is difficult to use the PA methods.

2.2 Positive-unlabeled (PU) data approach

Unlike the PA methods, the positive-unlabeled (PU) data methods do not treat the pseudo-absence data as genuine $Y = 0$ observations. Instead, the PU methods treat the outcome of the pseudo-absence data as *indeterminate*. For instance, the maximum entropy method (Phillips and Dudík, 2008), which is one of the most widely used methods in the species distribution modeling, estimates the probability distribution of Y over a specific extent using observed events. The estimated probability distribution is then used for predicting zones as well as assigning specific probabilities to the unlabeled data.

Although the PU method is the current standard in the literature on species distribution modeling, recent studies have shown that the PU methods are actually dependent on how one defines the scopes of unlabeled data (VanDerWal *et al.*, 2009). In conflict studies, Schutte (2017) applies a point process model (PPM) to predict zones of ten insurgent wars in Africa. Although the author correctly refers to the problems of areal-unit assumptions, the PPM actually depends on particular areal assumptions, including the geographical scope of the analysis and the specification of the grid cells.[6] Thus, although the PPM and more generally the PU methods are great departures from the deterministic methods, they are still confined by the areal-unit assumptions.[7] At the crux, the deterministic, PA, and PU methods suffer the same problem; they are sensitive to the choices of pre-defined areal units.

2.3 Positive-only (PO) approach

The positive-only (PO) methods can provide a possible solution to the areal-unit problems (Mack and Waske, 2017). Unlike the PA or PU methods, the PO methods solely rely on presence data without requiring absence data or pre-defined areal units. The PO approach therefore can be considered as a *minimalist* approach to conflict zoning; even though the PO methods can be less informative as they do not utilize unlabeled data, they do not require strong assumptions and hence allow broader applications. In general, while the PA and PU approaches are useful when one's objective is to make the best possible zones for a few conflicts with field-level knowledge, the PO approach is more suitable when one would like to create database infrastructure for the purpose of broader application. This paper aims at the latter objective and hence develops a PO method.

3. Statistical method of zoning: one-class support vector machine (OCSVM)

The OCSVM is an unsupervised machine learning method and one of the most popular among the PO approaches. There are several applications in the fields of text analysis (Lee and Liu, 2003), species distribution models (Mack and Waske, 2017), and gene science (Mei and Zhu, 2015). The advantages of OCSVM over other PO methods are that it is particularly useful for handling continuous predictors and that the hyper-parameter tuning is relatively well understood.[8]

[6]The tessellation algorithm in the PPM requires grid cells or sampling schemes.

[7]Although Schutte (2017) proposes a cross-validation scheme, it relies on a performance metric that favors predictions similar to the Gaussian kernel densities. It is, however, unclear why predictions need to be similar to the Gaussian kernel (and if so, we should use the Gaussian kernel in the first place).

[8]Possible alternatives are isolation forest (Liu *et al.*, 2008) and autoencoder (Hinton and Salakhutdinov, 2006). In general, the isolation forest is suitable to categorical predictors, while the autoencoder is particularly useful when there exist a large

Conceptually (but not algorithmically), the OCSVM can be considered as a two-step procedure; transforming data with a fairly flexible function φ and then fitting the tightest enclosing circle to the transformed data.[9] As seen in Figure 3, the function φ maps the observed m predictors to m-dimensional Cartesian space so that the data are centered at $\boldsymbol{b}$ (in Figure 3, $m = 2$). Although such an m-to-m function is hard to even express, it is mathematically sufficient to define its kernel, $K(\boldsymbol{x}_j, \boldsymbol{x}_k) = \varphi(\boldsymbol{x}_j)^T\varphi(\boldsymbol{x}_k)$, which maps two m-length vectors to *a scalar* and hence is mathematically tractable.[10] The Euclidian distance, for instance, would be such a kernel, but we can use more flexible kernels as well. A standard choice is the radial basis function;

$$K_{rbf}(\boldsymbol{x}_j, \boldsymbol{x}_k) = \exp(-\gamma||\boldsymbol{x}_j - \boldsymbol{x}_k||),$$

where γ is a kernel parameter, which represents the influence of a single observation on the overall estimate. Larger γ indicates a tighter fit to every observation. The support vector machine with the radial basis function is so flexible that it can approximate to any finite function (so-called universal approximator; Hammer and Gersmann, 2003).

Given a specified kernel, the OCSVM searches for the tightest circle that encloses the transformed data points (the red dashed circle in the right pane of Figure 3). However, because it is not desirable to fit the circle too tightly to the data and risk overfitting, we also allow several observations to be outside of the circle (four data points in the right pane of Figure 3). This provides a guard against overfitting. Formally, the loss function and corresponding optimization

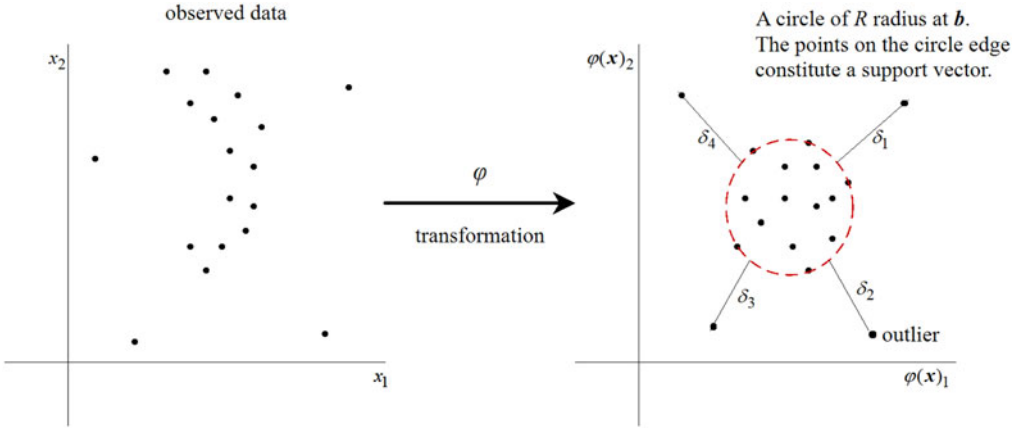

Figure 3. Stylized example of the OCSVM.

Note: The figure shows an example of the OCSVM with hypothetical data. The left pane plots the observed events with respect to two predictors, x_1 and x_2 (say, longitude and latitude). The right pane plots the observations transformed by a flexible function φ. In the right pane, the red circle is the fitted OCSVM, the points on the edge of the circle constitute a support vector, and points outside of the circle are outliers (events that reflect stochastic errors). By transforming the circle back to the original space, one can obtain the estimated zone as well as the outliers.

number of predictors. Given the continuous predictors and low dimensionality, I use the OCSVM in the present analysis. Although spatial interpolation methods, such as kriging and Gaussian kernel, can be used as well, these methods require an arbitrary threshold for a binary classification.

[9]For more detailed mathematical treatment, refer to Schölkopf *et al.* (2000). The Support Vector Data Description (SVDD) is mathematically equivalent to the OCSVM in a standard setup.

[10]The input data must be standardized.

problem is expressed as;

$$\min\left(E_X[\mathrm{L}(f_Y, \hat{f}_{Y|X})]\right) \approx \min_{R,b,\delta}\left(R^2 + \frac{1}{\nu}\frac{\sum_{i=1}^{n}\delta_i}{n}\right);$$

with constraints of;

$$||\varphi(\boldsymbol{x}_i) - \boldsymbol{b}|| \leq R^2 + \delta_i \text{ and } \delta_i \geq 0 \quad \forall i \in \{1,\ 2,\ \ldots,\ n\},$$

where R is a radius of the circle. We would like to have a circle that encloses the points as tightly as possible (minimizing R^2), but we also want the circle to be sufficiently inclusive and thus not so far from the outliers (minimizing $\sum_{i=1}^{n}\delta_i$). The parameter ν controls the weights of those two opposing forces; large ν allows many outliers, while small ν means an inclusive circle. By solving the optimization problem for $\hat{R}$, $\hat{\boldsymbol{b}}$, and $\hat{\boldsymbol{\delta}}$, we get the OCSVM approximation to the population zoning function;[11]

$$\hat{f}_{\mathrm{OCSVM}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad ||\varphi(\boldsymbol{x}) - \hat{\boldsymbol{b}}|| \leq \hat{R}^2 \\ 0 & \text{otherwise} \end{cases}.$$

Since the two hyper-parameters γ and ν (both of which control the balance between underfitting and overfitting)[12] are not directly estimated, I follow Ghafoori *et al.* (2018) to choose the optimal values.[13] The predictive intervals are obtained via bootstrapping.

4. Performance comparison I: simulation analysis

One advantage of using statistical methods is their ability to separate a systematic pattern of conflict events from non-systematic errors. In this section, I compare the performance of both deterministic and statistical methods by conducting a couple of simulation analyses. I first define a population conflict zone as the entire territory of Nigeria or Somalia. I choose these countries because they have perhaps the most convex and concave shapes among African countries. I then randomly draw 1000 locations within the territory and add random noise;

$$\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{v}_i;$$

$$\tilde{\boldsymbol{y}}_i \underset{\text{iid}}{\sim} U_{\text{poly}};$$

$$\boldsymbol{v}_i \underset{\text{iid}}{\sim} N(0,\ \sigma^2) \text{ for } i \in \{1,\ 2,\ \ldots,\ n\},$$

where U_{poly} is a uniform distribution over the territory of Nigeria or Somalia, $\tilde{\boldsymbol{y}}_i$ is a location within the territory, and $\boldsymbol{v}_i$ is noise drawn from a normal distribution of mean zero and variance σ^2. I vary the size of the noise σ from 0 to 1 degree (~0 to 111 km).[14] We are supposed to have no information about $\tilde{\boldsymbol{y}}_i\boldsymbol{v}_i$, or their distributions with only having the data $\boldsymbol{y}_i$. Our task is to infer the population conflict zone from the observed data $\boldsymbol{y}_i$.

[11]For a proof of Lagrangian optimization, refer to Schölkopf *et al.* (2000).

[12]While large γ makes a zone tight to every observation and hence "snaky," small ν makes a zone inclusive and hence "stretched." Both make the zone sensitive to outliers.

[13]For the detail of the hyper-parameter selection, see Supporting Information 1.

[14]Any larger value makes the performances of all of the methods equally worse.

In the following analysis, I compare the performances of the PRIOGRID and district assignments, the convex hull (deterministic approaches), support vector machines (SVM; PA data approach), maximum entropy method (MAXENT; PU data approach), and one-class support vector machine (OCSVM; PO approach).[15] The convex hull method is supplemented with a deterministic rule for outlier removal, which is used in the UCDP Polygons dataset (so-called 20–5 percent rule).[16] Since the SVM and MAXENT require pseudo-absence or unlabeled data, I randomly sample locations and use them as pseudo-absence or unlabeled data.[17] Finally, I evaluate the performance by calculating the accuracy of the predictions (the proportion of correctly predicted conflict and non-conflict area across the entire area). I repeat the simulation for 1000 times for each value of σ and calculate the average accuracies.[18]

4.1 Results

The following figure (Figure 4) shows the results of the simulation analyses. On average, the OCSVM has a higher performance than the other methods in both simulations. Although the PRIOGRID assignment performs relatively well, the performance is sensitive to the addition of small amounts of noise especially in the case of Somalia, which is not surprising given its deterministic nature. In both simulations, the district assignment exhibits relatively low performance; in the case of Somalia, its accuracy quickly deteriorates and then becomes comparatively stable. The convex hull method works well only when a population conflict zone is convex. When the assumption of a convex conflict zone is violated, the accuracy becomes much lower.

Among the statistical methods, only OCSVM has high performance in both simulations. While the performance of the MAXENT is as high as OCSVM's in the case of Nigeria when there is a large amount of noise, the MAXENT has the second lowest accuracy in the Somalia simulation. Similarly, the performance of the SVM is somehow equivalent to that of the PRIOGRID assignment for Somalia, but it does poorly for the Nigeria simulation. These results reinforce the fact that SVM and MAXENT are sensitive to pseudo-absence data generation. Overall, the OCSVM exhibits the highest and most stable performance.

5. Performance comparison II: validation with the Rohingya crisis

Although it is usually difficult to validate conflict zones with real-world data as we rarely have absence observations (and without absence observations, we cannot calculate accuracy), the case of the Rohingya Crisis provides a unique analytical opportunity. Specifically, the United Nations Institute for Training and Research (UNITAR) analyzes high-resolution satellite images to measure the levels of housing destruction at 900 Rohingya villages in Myanmar for the period of 31 August 2017 to 31 March 2018 (UNITAR, 2018). Importantly, the dataset contains information about *both* presence *and* absence of housing destruction in each village. Although housing destruction might not be a valid indicator of conflict, I can at least analyze to what extent the conflict zones (or zones of housing destruction) validly reflect the reality. If the UNITAR data

[15]I choose these methods because (i) PRIOGRID and districts are the spatial units that are the most commonly used in conflict studies, (ii) the convex hull method is used in the UCDP Polygons dataset, (iii) the OCSVM is a natural extension of the SVM and hence they are more comparable, and (iv) the MAXENT is the current standard in the literature of species distribution models.

[16]See Croicu and Sundberg (2012) for details of the 20–5 percent rule.

[17]In particular, the pseudo-absence or unlabeled data are uniformly drawn from a rectangle that has maximum and minimum extents equal to those of the observed data. When I draw them from the global extent, the performances of the SVM and MAXENT become much lower. Following the convention, the number of the pseudo-absence data are set as equal to the sample size of presence data. The hyper-parameters of the SVM are tuned with cross-validation.

[18]The confidence intervals are very small and hence not reported.

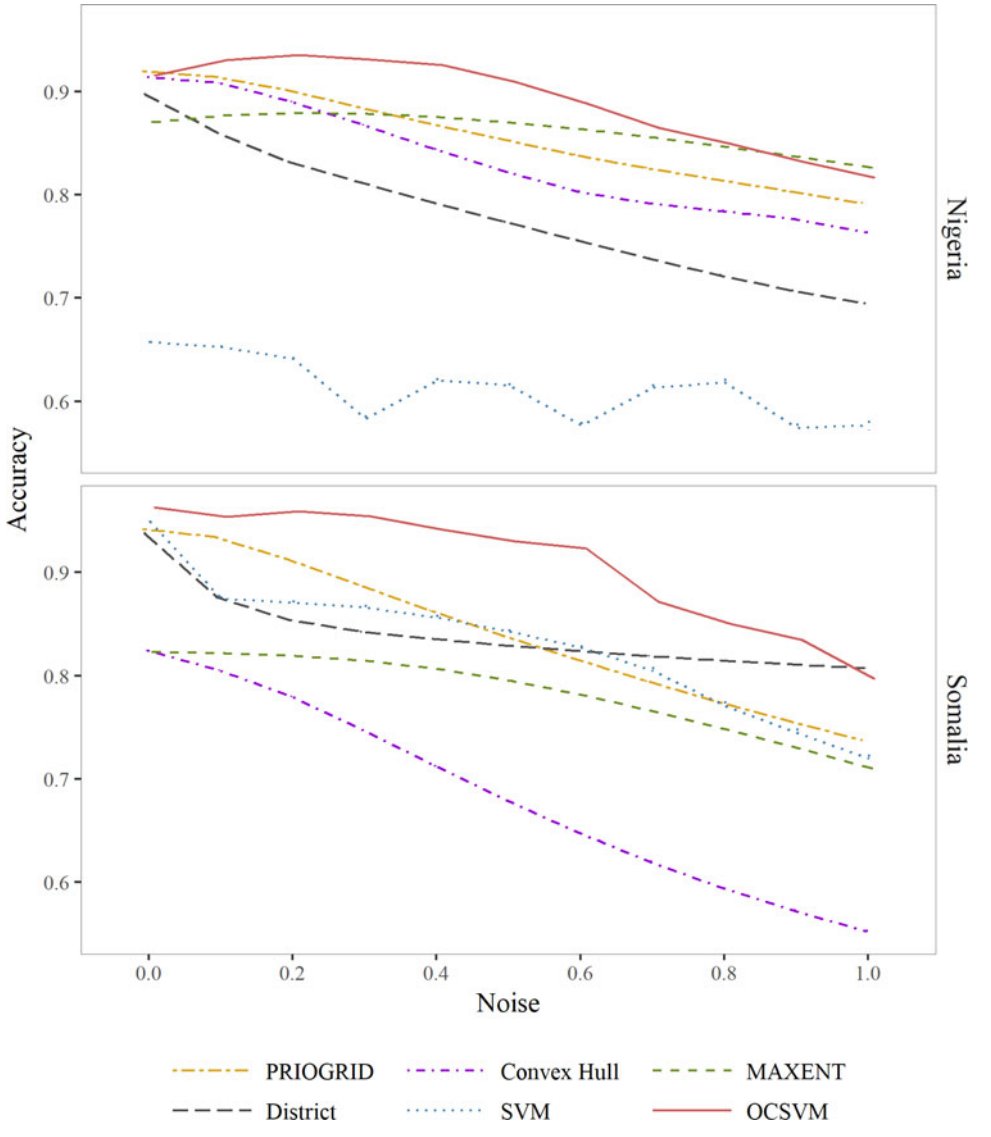

Figure 4. Simulation: performance comparison.
Note: The figures shows the performances of PRIOGRID assignment (yellow dot-long-dashed line), district assignment (gray long-dashed line), a convex hull (purple dot-dashed line), support vector machine (blue dotted line), maximum entropy method (green dashed line), and OCSVM (red solid line). The upper and lower panes show the results when the territories of Nigeria and Somalia are used in the simulation respectively. The horizontal axis shows the level of noise in the observed data (σ). The vertical axis shows the accuracy. The confidence intervals are very small and hence not reported.

indicate "few" or more destruction, it is considered as evidence for the presence of conflict, and hence the outcome variable takes a value of 1.

I conduct a two-fold cross-validation test with the housing destruction data. I first randomly split each of the destroyed and unaffected villages to two groups. The assignments of the PRIOGRID cells and township polygons,[19] convex hull, SVM, MAXENT, and OCSVM are then applied to

[19]The townships are the third-order administrative units next to districts. There are only five PRIOGRID cells and three townships in the Rohingya region (there is only one district). Despite this fact, I use the PRIOGRID and townships because

one half of the affected villages, and the corresponding conflict zones are estimated. I calculate the accuracies of the conflict zones by comparing them to the other half of the affected villages and one half of the villages that were unaffected.[20] The same exercise is done by replacing the groups. The two-fold cross-validation is repeated 500 times (thus, $2 \times 500 = 1,000$ simulation outputs). Finally, the average accuracy is calculated.[21] The other specifications are the same as those in the simulation analyses.

As seen in Figure 5, the OCSVM exhibits the highest performance, indicating that the OCSVM better reflects the reality of the Rohingya Crisis. Nonetheless, it should be noted that the performance is not very high in the *absolute* term; only about seven out of ten times, the OCSVM correctly distinguish affected and unaffected villages. This reflects the generic difficulties of one-class classification. Thus, as mentioned above, the OCSVM should not be considered as substitutes for detailed field-level knowledge. Having said that, however, the OCSVM marks improvement compared to the extant methods; the OCSVM increases the probability of correct predictions by 0.25, 0.2, 0.05, and 0.03 compared to the PRIOGRID and polygon assignments, the MAXENT, the convex hull, and the SVM respectively.

Although the SVM exhibits a performance similar to the OCSVM, the SVM's accuracy varies substantially across simulations. Indeed, the standard deviation of SVM's accuracy is 0.065, which is far larger than any of the other methods (the standard deviations of the other methods are below 0.03). This is not surprising because the SVM relies on the random sampling of absence data and hence is subject to additional noise. Next, even though the convex hull also exhibits a

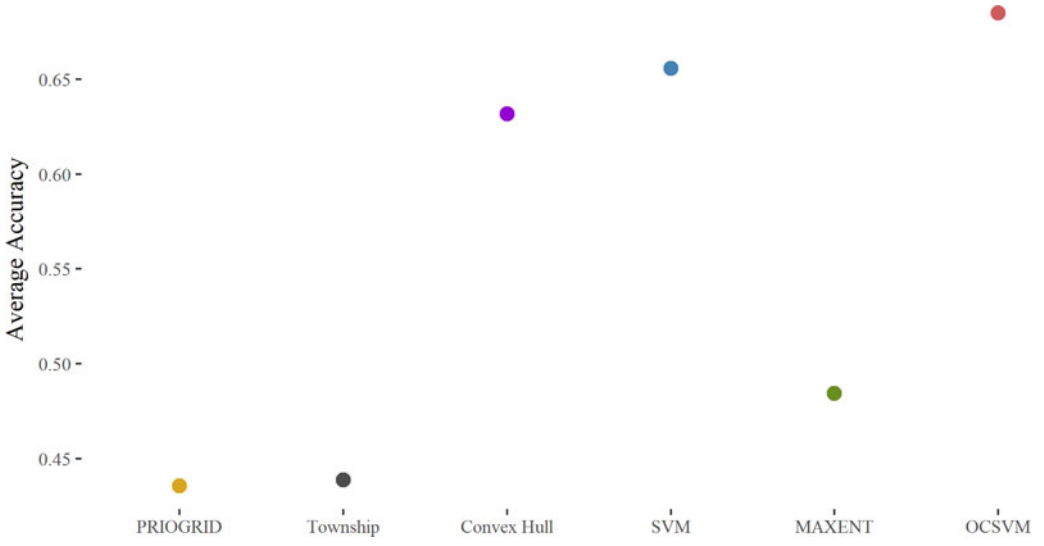

Figure 5. Validation: performance comparison.
Note: The figure shows the results of the two-fold cross-validation tests. The vertical axis is average accuracy over 500 cross-validation tests (thus $2 \times 500 = 1000$ simulations). The confidence intervals are very small and hence not reported.

these are often used in macro-level analysis. The goal of this paper is to develop a method for macro-level comparison, and the analysis in this section attempts to validate the macro-level methods with micro-level data. Although it might be interesting to use smaller grid cells or polygons, I do not evaluate these areal assignments. The reasons are (i) due to the sheer sample size, they cannot be used in analysis with a large number of countries, (ii) there is no global dataset of village or equivalent administrative polygons, and as a result (iii) to my best knowledge, there is no study that uses those fine-grained units in a large number of countries.

[20]The other half of the unaffected villages are not used in order to maintain the ratio of the affected and unaffected villages.

[21]The confidence intervals are very small and hence not reported.

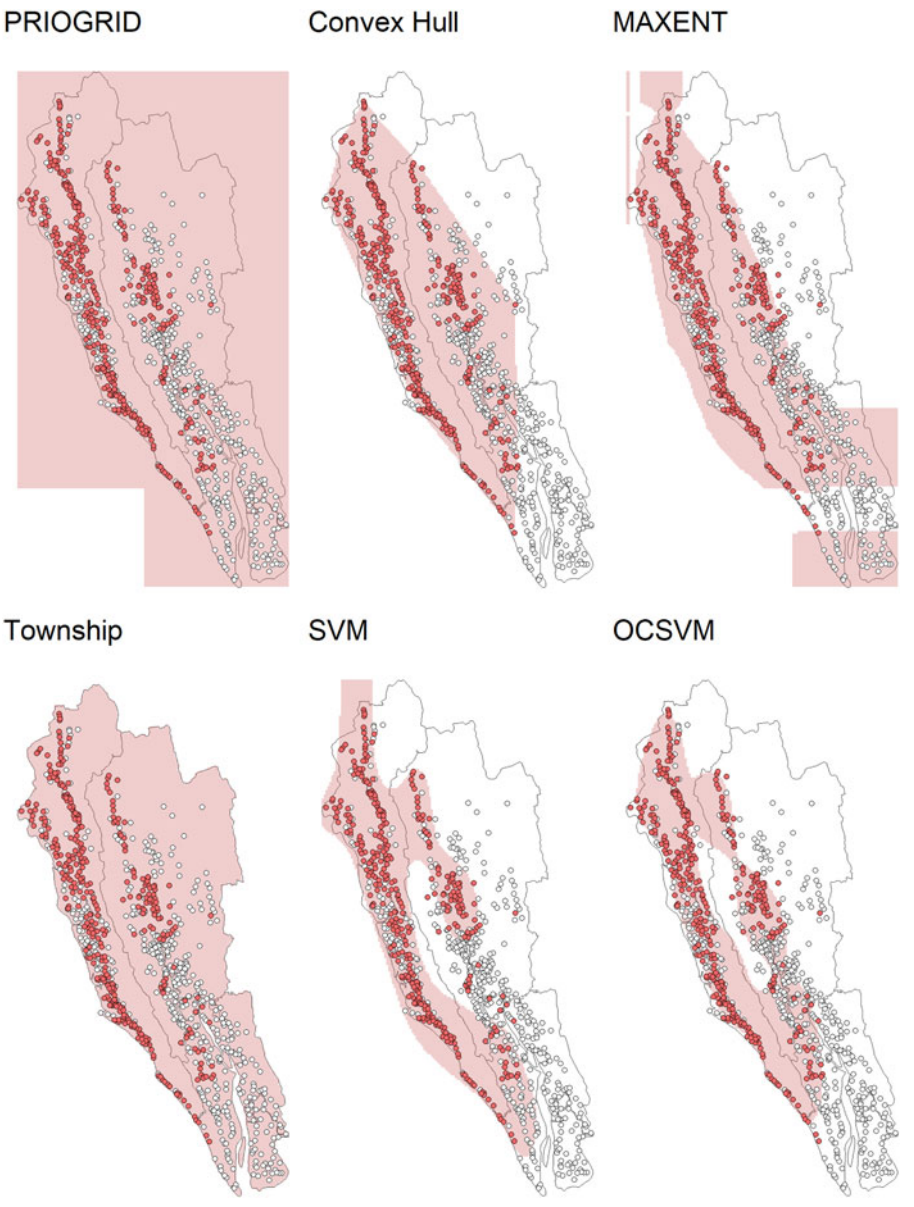

Figure 6. Estimated zones of the Rohingya housing destruction.
Note: The figure shows the zones of the Rohingya housing destruction estimated by the six methods. The red points show the villages that suffered "few housing destruction" or more between 31 August 2017 and 31 March 2018. The white points villages that did not suffer any housing destruction. Those data are derived from UNITAR (2018). The red areas are the estimated zones.

relatively high performance, it includes the central mountain areas in which there is no housing destruction or conflict (the upper middle pane of Figure 6). Because there is no observation in the mountain areas, these mis-predictions are not reflected in the accuracy metric, which creates the impression that the convex hull would be as accurate as the OCSVM. The OCSVM, on the other hand, does not include those central mountain areas.

Compared to those statistical methods, the MAXENT exhibits very low accuracy. As seen in the upper right pane of Figure 6, the MAXENT is unstable outside the extent of the presence

observations. Moreover, even within this extent, the predictions are too inclusive and therefore inaccurate. Finally, the PRIOGRID and township polygon assignments have the lowest accuracies, which is not surprising given the large sizes of the grid cells and townships. Because a majority of the UCDP GED events are also reported at the levels of villages, towns, or cities, those findings cast a doubt on the validity of those polygon assignments. Overall—even though none of these methods can substitute field-level knowledge—the OCSVM exhibits the highest performance, indicating its potential use for macro-level analysis.

6. New conflict zones: application to the UCDP GED

I apply the OCSVM to the UCDP GED (version 19.1), a conflict event dataset commonly used not only in political science but also in other fields (Daskin and Pringle, 2018).[22] An armed conflict is defined as "[a]n incident where armed force was by an organized actor against another organized actor, or against civilians, resulting in at least 1 direct death at a specific location and a specific date" (Sundberg, Lindgren, and Padskocimaite, 2010: 2). Although recent studies point out reporting biases in the dataset (Weidmann, 2015, 2016), the reporting biases require solutions at the level of event data collection. Thus, they are beyond the scope of this paper. The following analysis is readily replicable with more accurate event data. The new conflict zone data can also potentially be used in conjunction with the calibration method proposed by Donnay *et al.* (2018).

I estimate conflict zones with and without using the conflict event dates as an additional predictor so that I can create both time-variant and time-invariant conflict zones. Each conflict event is weighted by the casualties so that events of higher casualties have larger weights in the estimation. With the event data, I separately estimate the conflict zones for each dyad of actors. The UCDP GED specifies a conflict name (which I call "conflict episode") and names of two involved actors (which I call "conflict dyad") for every conflict event.[23] Therefore, in the example of the Iraqi Insurgency, I create conflict zones for battles between the government and Islamic State, battles between the government and Ansar al-Islam, and so forth.[24] Because each dyad is always assigned to a single episode—which in turn belongs to either state-based, one-sided or non-state conflict type—the dyadic conflict zones can be easily aggregated to zones at the levels of conflict episodes or types.

I do not include any geographic or climatic predictors so that the conflict zones are solely based on the UCDP GED and hence those predictors can be used in later analyses.[25] These features are intended to match those of the UCDP Polygons dataset.[26] The goal here is to provide a reliable alternative to the UCDP Polygons dataset, which has not been updated for the past 8 years and only includes Africa.[27] Moreover, the roles of additional predictors are rather limited in the OCSVM. If predictors do not affect the conflict locations, there is no reason to include them. By contrast, if predictors can affect the locations of conflict events, such effects are already reflected in the conflict events themselves. Although the predictors may still provide efficiency gains, they

[22]Although the OCSVM is readily applicable to other event datasets, I choose the UCDP GED as the dataset has accompanying zonal data, UCDP Polygons.

[23]In implementation, I use the corresponding IDs in the UCDP GED (*conflict new id* and *dyad new id* variables).

[24]Due to small sample sizes, zones cannot be estimated for dyads of three or fewer events. Although this is clearly a disadvantage of the statistical method, it does not mean that the deterministic rules could create valid conflict zones for those cases. Given the limited number of events, I even doubt that defining zones is substantively meaningful for those cases.

[25]I also account for geographic and temporal precision of the conflict events by resampling their locations and dates. For the full details of the preprocessing, see Supporting Information 2.

[26]Depending on research interests, it is recommended to drop the cases relating to Al-Qaeda's transnational terrorism (UCDP GED conflict ID 418 and 608), which have exceptionally large conflict zones. The following analyses exclude those cases.

[27]Confirmed on 20 November 2019.

do not reduce biases in the estimate. Finally, the predictors also limit the possible usage of the conflict zones. If one would include geographic predictors, for instance, it prevents us from analyzing the relationship between those predictors and conflict zones in the causal analysis. This is less than attractive not only because we cannot answer those substantive questions, but also because we can no longer use those exogenous variables for the purpose of causal identification.

As a final note, recall that the conflict zones are not real geographical objects but concise summaries of conflict events, and hence the conflict zones are primarily used for *macro-level* analysis. For instance, it makes less sense to compare areas one or few kilometers inside and outside of the conflict zones, as the approximation errors are usually larger than such a small scale. Thus, the dataset also comes with estimates of the approximation errors. Specifically, I use parametric bootstrapping to provide the standard errors and corresponding 95 percent lower and upper bounds of the conflict zones.[28] The interval estimates can be used for the purpose of sensitivity analysis.

The new conflict zone dataset—Wzone—is publicly available in time-varying (daily; 1989–2018) and static versions at the levels of conflict dyads and episodes. Any geo-spatial covariate can be incorporated to Wzone by calculating the mean or other metrics within each zone. Conversely, the Wzone dataset can be integrated to PRIOGRID (Tollefsen *et al.*, 2012) and other spatial datasets by calculating the proportion of conflict zones within a spatial unit. The integration with PRIOGRID will allow researchers to access a wide array of covariates for further analysis.[29]

6.1 Results

The following figure (Figure 7) is the time-invariant estimates of conflict zones. The left and right panes are the UCDP Polygons dataset and the OCSVM estimates respectively.[30] For graphical purposes, the figure shows only the zones of state-based conflicts in Africa. Consistent with my argument, the UCDP Polygons tend to be less flexible but more sensitive to outliers. For example, while the UCDP Polygons contains substantial amounts of ocean areas for the case of Mozambique (blue; bottom right of Figure 7), the OCSVM estimates are mostly along the coastal lines. As I argued, because a majority of conflict events occurred inside the coastal lines, even without any covariates about terrain, the OCSVM properly accounts for the spatial distribution.

A more noticeable and perhaps important difference is the sensitivity to the outliers. For quite a few conflicts, the UCDP Polygons indicate larger conflict zones than those in the OCSVM estimates, including those in Algeria (brown; top left) and Angola (orange; bottom left). In the case of the Algerian Civil War, for instance, the conflict was mostly fought within the northern region of Algeria. A few terrorist attacks, however, squeeze the UCDP conflict zone to the outside of the country, including Mauritania, Mali, Niger, and a large area of the Sahara Desert. By contrast, the OCSVM estimate is contained within the northern coastal regions of Algeria, more accurately representing the nature of the civil war.[31]

7. The ecological costs of armed conflict: replication of Daskin and Pringle (2018)

Finally, I replicate Daskin and Pringle's (2018) study on the ecological consequences of armed conflict to demonstrate how the zones could alter the inferences they made. I choose the *Nature* letter to examine the potentially broad implications of the zoning problem and to

[28]For the details of the bootstrapping, see Supporting Information 3.

[29]I emphasize that the new conflict zone dataset is *not* an alternative to PRIOGRID itself, but it is an alternative to *the conflict zone variables* in the PRIOGRID dataset, which currently use the minimum circle that encloses all events of a given conflict.

[30]I update the UCDP Polygons dataset with the latest version of the UCDP GED (version 19.1).

[31]In Supporting Information 4, I also provide a focused comparison of conflict zones in the case of the Somali Civil War.

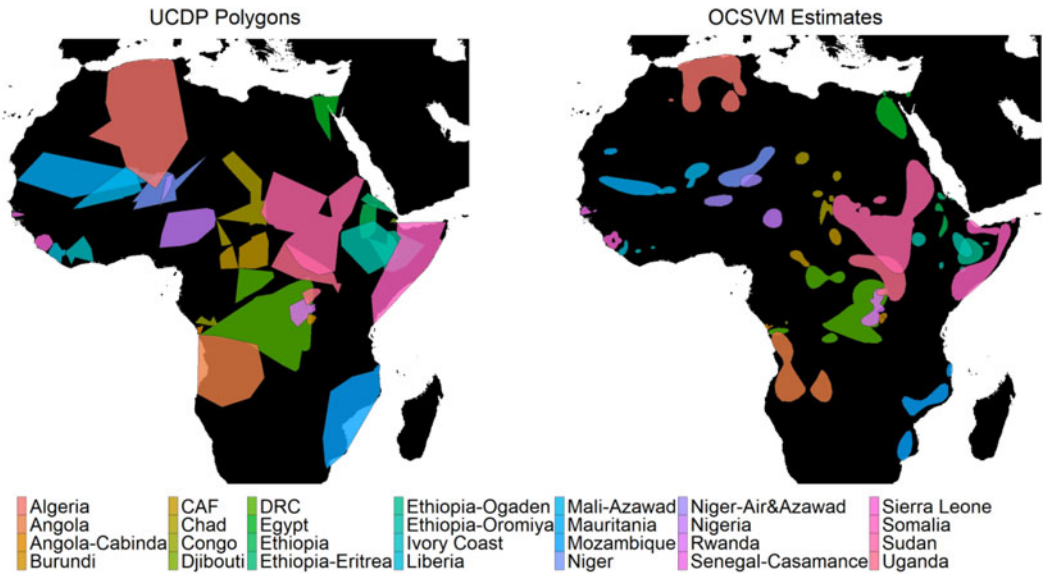

Figure 7. Time-invariant estimates of conflict zones in Africa.
Note: The figure maps the time-invariant estimates of conflict zones. The left and right panes are the updated UCDP Polygons dataset and the OCSVM estimates respectively. The conflict names are shown at the bottom with corresponding colors. For graphical purposes, the conflict zones are limited to those of state-based conflicts in Africa. For readers of monotone prints, please refer to the online article.

highlight an issue that is understudied in political science.[32] The article, which was published on 18 January 2018, has already been cited by 52 newspapers, including the New York Times and the Economist (8 August 2018).[33] Their sample is cross-sectional and comprised of 172 park-species combinations in Africa.[34]

The outcome variable is the annualized finite rate of population change,

$$\lambda = \left(\frac{d_{t=1}}{d_{t=0}}\right)^{1/(y_{t=1}-y_{t=0})},$$

where $d_{t=0}$ and $d_{t=1}$ are the densities of a wild large herbivores in the beginning and end years of mammal population records ($y_{t=0}$ and $y_{t=1}$ respectively). The lambda measures the ratio of the population size at the end of a year and the population size at the beginning of the year. The value $\lambda = 0.9$, for instance, indicates that if there are 100 animals at a beginning of a year, their population decreases to 90 at the end of that year. The densities of wild large herbivores are compiled by "systematically reviewing academic and grey literature" (Daskin and Pringle, 2018: 329).[35] Their key predictor is the proportion of conflict zones averaged over the years of mammal population records. While the authors use the UCDP Polygons dataset, I use the new conflict zone dataset and calculate a proportion of zones within each protected area. In the following section, I compare the results with the updated version of the UCDP Polygons and the results with the

[32]*Nature* letters are "short reports of original research focused on an outstanding finding whose importance means that *it will be of interest to scientists in other fields*" (https://www.nature.com/nature/for-authors/formatting-guide; 19 October 2018). To my best knowledge, this paper is the first response to Daskin and Pringle (2018) in political science. I also conduct two additional replications. See the last paragraph in the next subsection.

[33]See the author's website (https://joshdaskinecology.com/publications; accessed on 8 August 2018) for the latest information.

[34]The sample includes 96 protected areas and 30 species.

[35]See Daskin and Pringle (2018) for the details.

Table 1. The effects of armed conflicts on the mammal population

Replication with the updated UCDP polygons	Replication with the new conflict zones
−0.85 [−1.20, −0.49]	−0.10 [−0.40, 0.21]

Note: The table shows the regressions of mammal population trajectories on the average proportions of conflict areas in protected areas in Africa. The left and right columns show the results based on the updated UCDP Polygons and the OCSVM estimates respectively. In each column, the regression coefficient and corresponding 95 percent confidence intervals are reported. The control variables are human population density, proportion of urban areas, and drought frequency, which are included in the "best" model of Daskin and Pringle (2018). $n = 172$.

OCSVM estimates, while keeping the other specifications intact so that the only difference lies in the zoning methods.[36]

7.1 Results

The following table (Table 1) compares the results based on the updated UCDP Polygons and the OCSVM estimates. While the original finding (left columns in Table 1) indicates a statistically and substantively significant association between conflict zones and the decline of the mammal population, these results are not consistent with my conflict zones dataset. In fact, with the new zones, we cannot draw meaningful inferences from the data.

The differences become even clearer once we consider the effect sizes. The following figure (Figure 8) compares the trajectories of the hypothetical mammal population, which has an initial size of 100,000. For each of the estimated effects in Table 1, I calculate the population trajectory in a protected area that does not at all belong to conflict zones (blue dotted line) and that in an area totally belonging to conflict zones (red solid line). As seen in pane (a) of Figure 8, according to Daskin and Pringle (2018), the mammal population is stable or only slightly decreases without armed conflict, but it drastically decreases in conflict zones; in each year of the armed conflict, the population is estimated to decline to about 85 percent of the initial size. This means that within 5 years of armed conflict, the population would decrease to less than 1 percent of the initial size.

The estimates with the new conflict zones, however, indicate more modest and perhaps realistic trajectories (pane (b) of Figure 8); in each year of armed conflict, the population is predicted to decline to 90 percent of the initial size. Although this estimate is still large given the prolonged nature of armed conflict (the population decreases to about 59 percent of the initial size within five years of armed conflict), it at least does not mean that fighting would nearly eradicate the animals within a few years.[37] The results are also indeterminate. There is no definite evidence that mammal population decreases in conflict zones or that the rate of population loss is higher than that in non-conflict zones. Given the relatively large difference in the mean estimates, the null result is probably due to the small sample size. Future studies need to collect more observations to increase the power of the analysis.

I also conduct two additional replications, which are detailed in Supporting Information 6 and 7. Although I refrain from drawing a definite conclusion given the small number of replications, it appears that the measurement errors tend to have large impacts when we use the conflict zones for creating variables and/or when the sample size is small. The analysis with a small sample can be heavily influenced by systematic or non-systematic measurement errors in a few observations.

[36]I use the "best" specification in Daskin and Pringle (2018; the authors use model-selection and model-averaging techniques). In Supporting Information 5, I also present the original estimates and the results replicated with the old version of the UCDP Polygons.

[37]In Supporting Information 5, I also examine the reasons why the estimates are so different depending on the conflict zones.

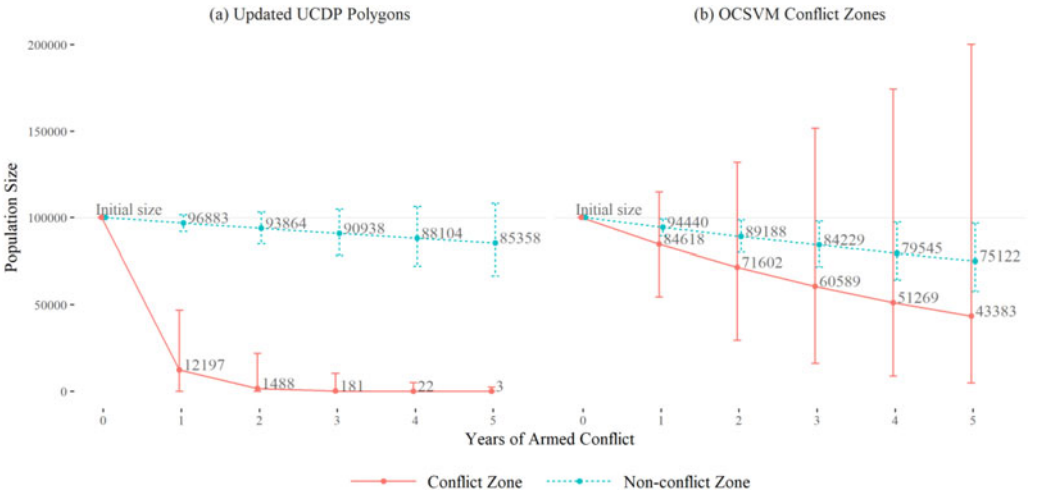

Figure 8. The ecological costs of armed conflict.
Note: The figure shows the estimated trajectories of a hypothetical mammal population. The initial size of the population is 100 thousands. The left and right panes show the population trajectories estimated with the UCDP Polygons (exact replication of Daskin and Pringle (2018)) and with the OCSVM-based conflict zones. The dotted blue lines are the population trajectories in protected areas any part of which does not experience armed conflict. The solid red lines are the population trajectories in protected areas which totally belong to conflict zones.

In fact, the new conflict zones also substantially alter the results of Beardsley *et al.* (2015), who use the UCDP Polygons for measuring rebels' movement in an analysis with a relatively small sample (n = 257). By contrast, the new measure does not alter the main findings of Fjelde and Hultman (2014), who use the conflict zones for selecting a sample in an analysis with large panel data.[38]

These results, however, do not mean that a larger number of observations can *always* mitigate the biases from measurement errors. In fact, the measurement errors in Beardsley *et al.* (2015) have systematic patterns that will not disappear even with a large sample. This demonstrates that the biases in empirical estimates can persist. Thus, it is advised for future studies to carefully assess the underlying assumptions of conflict zones and the patterns of the measurement errors. If the measurement errors are not systematic, a large sample can help (even though it can cause attenuation biases). If the measurement errors are systematic, however, the empirical findings must be taken with great caution.

8. Conclusion

In conflict studies, the selection of areal units is so common that people may not recognize that the areal assignment is indeed an assumption. Without properly understanding *where* armed conflict takes place, however, we cannot know *why* armed conflict occurs or *what* its consequences are. In this paper, I have addressed the areal-unit problems by developing a theory, method, and dataset of conflict zones. I define a zone as a summary function that maps locations and other relevant information onto the presence and absence of armed conflict. This formalization clarifies that the zoning exercise is essentially a statistical problem—it is a matter of how we can infer a zoning function from observed data of conflict events. I answer this question by applying the OCSVM, which unlike other deterministic or statistical methods does not depend on a predefined areal unit. I apply the OCSVM to the UCDP GED conflict event dataset and create a new dataset of conflict zones. The replication of Daskin and Pringle (2018) indeed indicates

[38]I choose these two studies as they are the most cited articles that use the UCDP Polygons (confirmed on 20 November 2019 at Google Scholar).

that zones can potentially alter our inferences about the ecological costs of armed conflict in statistically and substantively significant ways.

Although this paper is primarily interested in armed conflict and applies the method to the UCDP GED, the theory and method can be applied to the other conflict data, such as ACLED (Raleigh *et al.*, 2010), SCAD (Hendrix and Salehyan, 2013), and ICEWS (Boschee *et al.*, 2015), and potentially to other topics in the social sciences, including poverty mapping, crime zones, state controls over territories, hazard maps, and zones of racial segregation. Although application to those topics will certainly require extensions and modifications, the framework of this paper provides a way to think about the problems and thus to develop suitable methods. I hope this paper facilitates our understandings on the geography of armed conflict and, more broadly, the areal-unit assumptions in political science.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2020.16.

Acknowledgements. I am grateful to Michael G. Findley, Stephen Jessee, Yuta Kamahara, the two anonymous reviewers, and the editors of PSRM for thoughtful comments. I also express appreciation to Ross Buchanan for writing support. This paper was presented at the IR Workshop at the University of Texas at Austin, the Workshop on Armed Conflict and Political Economy of Development at Kyoto, and American Political Science Association Annual Meeting in 2018.

References

Anders T, Xu H, Cheng C and Satish Kumar TK (2017) Measuring territorial control in civil wars using hidden Markov models: a data informatics-based approach. Available at http://arxiv.org/abs/1711.06786 (Accessed 1 January 2018).

Beardsley K, Gleditsch KS and Lo N (2015) Roving bandits? The geographical evolution of African armed conflicts. *International Studies Quarterly* **59**, 503–516.

Boschee E, Lautenschlager J, O'Brien S, Shellman S, Starz J and Ward M (2015) ICEWS coded event data. Available at http://dx.doi.org/10.7910/DVN/28075 (Accessed 1 January 2018).

Buhaug H and Lujala P (2005) Accounting for scale: measuring geography in quantitative studies of civil war. *Political Geography* **24**, 399–418.

Buhaug H, Gleditsch KS, Holtermann H, Østby G and Tollefsen AF (2011) It's the local economy, stupid! Geographic wealth dispersion and conflict outbreak location. *Journal of Conflict Resolution* **55**, 814–840.

Campbell SP, Findley MG and Kikuta K (2017) An ontology of peace: landscapes of conflict and cooperation with application to Colombia. *International Studies Review* **19**, 92–113.

Croicu MC and Sundberg R (2012) UCDP GED conflict polygons dataset codebook version 1.1-2011. Available at https://ucdp.uu.se/downloads/ged/ucdp-ged-polygons-v-1-1-codebook.pdf (Accessed 1 January 2018).

Cunningham KG and Weidmann NB (2010) Shared space: ethnic groups, state accommodation, and localized conflict. *International Studies Quarterly* **54**, 1035–1054.

Daskin JH and Pringle RM (2018) Warfare and wildlife declines in Africa's protected areas. *Nature* **553**, 328–332.

Donnay K, Dunford ET, McGrath EC, Backer D and Cunningham DE (2018) Integrating conflict event data. *Journal of Conflict Resolution* **63**, 1337–1364.

Fjelde H and Hultman L (2014) Weakening the enemy: a disaggregated study of violence against civilians in Africa. *Journal of Conflict Resolution* **58**, 1230–1257.

Fjelde H and von Uexkull N (2012) Climate triggers: rainfall anomalies, vulnerability and communal conflict in Sub-Saharan Africa. *Political Geography* **31**, 444–453.

Friedman JH (1997) On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**, 55–77.

Ghafoori Z, Erfani SM, Rajasegarar S, Bezdek JC, Karunasekera S and Leckie C (2018) Efficient unsupervised parameter estimation for one-class support vector machines. *IEEE Transactions on Neural Networks and Learning Systems* **29**, 1–14.

Ghobarah HA, Huth P and Russett B (2003) Civil wars kill and maim people: long after the shooting stops. *American Political Science Review* **97**, 189–202.

Hammer B and Gersmann K (2003) A note on the universal approximation capability of support vector machines. *Neural Processing Letters* **17**, 43–53.

Hendrix CS and Salehyan I (2013) Social Conflict in Africa Database (SCAD). Available at https://www.strausscenter.org/scad.html (Accessed 1 January 2018).

Hinton GE and Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)* **313**, 504–507.

Kalyvas SN (2006) *The Logic of Violence in Civil War*, 1st Edn, New York, NY: Cambridge University Press.

Lee WS and Liu B. (2003) Learning with positive and unlabeled examples using weighted logistic regression. In Fawcett T and Mishra N (eds). *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. Washington, DC, USA: AAAI Press, pp. 448–455.

Liu FT, Ting KM and Zhou Z (2008) Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422.

Mack B and Waske B (2017) In-depth comparisons of MaxEnt, biased SVM and one-class SVM for one-class classification of remote sensing data. *Remote Sensing Letters* **8**, 290–299.

Mei S and Zhu H (2015) A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Scientific Reports* **5**, 8034.

Phillips SJ and Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175.

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J and Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**, 181–197.

Pierskalla JH and Hollenbach FM (2013) Technology and collective action: the effect of cell phone coverage on political violence in Africa. *American Political Science Review* **107**, 207–224.

Raleigh C, Linke A, Hegre H and Karlsen J (2010) Introducing ACLED: an armed conflict location and event dataset special data feature. *Journal of Peace Research* **47**, 651–660.

Ritter EH and Conrad CR (2016) Preventing and responding to dissent: the observational challenges of explaining strategic repression. *American Political Science Review* **110**, 85–99.

Sambanis N (2004) What is civil war? Conceptual and empirical complexities of an operational definition. *Journal of Conflict Resolution* **48**, 814–858.

Schölkopf B, Williamson RC, Smola AJ, Shawe-Taylor J and Platt JC. (2000) Support vector method for novelty detection. In Solla SA, Leen TK and Müller K (eds). *Proceedings of the 12th International Conference on Neural Information Processing System*. Cambridge, MA: MIT Press, pp. 582–588.

Schutte S (2017) Regions at risk: predicting conflict zones in African insurgencies. *Political Science Research and Methods* **5**, 447–465.

Sundberg R, Lindgren M and Padskocimaite A (2010) UCDP GED codebook version 1.0-2011. Available at http://ucdp.uu.se/downloads/ged/ucdp-ged-polygons-v-1-1-codebook.pdf (Accessed 1 January 2018).

Tao R, Strandow D, Findley M, Thill J-C and Walsh J (2016) A hybrid approach to modeling territorial control in violent armed conflicts. *Transactions in GIS* **20**, 413–425.

Tollefsen AF, Strand H and Buhaug H (2012) PRIO-GRID: a unified spatial data structure. *Journal of Peace Research* **49**, 363–374.

UNITAR (2018) Myanmar: buthidaung, Maungdaw, and Rathedaung townships/Rakhine State. Available at https://www.unitar.org/maps/map/2727 (Accessed 20 November 2019).

Valentini G and Dietterich TG (2004) Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. *Journal of Machine Learning Research* **5**, 725–775.

VanDerWal J, Shoo LP, Graham C and Williams SE (2009) Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling* **220**, 589–594.

Weidmann NB (2015) On the accuracy of media-based conflict event data. *Journal of Conflict Resolution* **59**, 1129–1149.

Weidmann NB (2016) A closer look at reporting bias in conflict event data. *American Journal of Political Science* **60**, 206–218.

Cite this article: Kikuta K (2022). A new geography of civil war: a machine learning approach to measuring the zones of armed conflicts. *Political Science Research and Methods* **10**, 97–115. https://doi.org/10.1017/psrm.2020.16