Supporting Information for

"A New Geography of Civil War:
A Machine Learning Approach to Measuring the Zones of Armed Conflicts"

Kyosuke Kikuta

Supporting Information 1. OCSVM Hyper-parameter Tuning

The hyper-parameter tuning of the OCSVM, and one-class classification more broadly, is not as straightforward as that for conventional classification problems. In this supporting information, I detail the main problem with this method, existing approaches for addressing it, and my solution.

Problem

In the literature of machine learning, a common approach to hyper-parameter tuning is cross-validation. Researchers split the sample into k groups, remove one of the groups, fit a model with the rest of the data, and calculate the predictive performance with the omitted data. This procedure is repeated for all groups and all possible combinations of hyper-parameter values. Finally, the hyper-parameter values that have the highest predictive performance are chosen. Standard performance metrics in binary classification are accuracy $\frac{TP+TN}{n}$, where TP and TN are the numbers of true positives and true negatives, and F1 score, $2\frac{p \cdot r}{p+r}$, where p is precision $\frac{TP}{PP}$ (PP is the number of positive predictions) and r is recall $\frac{TP}{CP}$ (CP is the number of condition positives). In one-class classification, however, these performance metrics cannot be directly calculated; the true negative rate and precision require $y_i = 0$ observations. As a result, cross-validation cannot be used without any modification.

Existing Approaches

Broadly speaking, the literature of one-class classification takes two approaches to the problem. The first approach extends the cross-validation scheme by introducing a new performance metric that can be used even in one-class classification. Lee and Liu (2003), for instance, propose a matric $\frac{r^2}{Pr(f=1)}$, where $Pr(f = 1)$ is the frequency of positive predictions. Similarly, Banerjee, Burlina,

and Diehl (2006) propose $\frac{r}{(\#\ of\ SV)/n}$, where $\#\ of\ SV$ is the number of observations used as parts of a support vector.

In my application, however, implementing these cross-validation techniques is problematic. Because the event locations in the UCDP GED tend to cluster at a few points (Mogadishu in case of Somalia), creating a tiny circle around those points means that the recall $r = \frac{TP}{CP}$ will be non-zero, while the prediction frequency $Pr(f = 1)$ and the number of support vector observations will approach zero. These produce near-infinities of $\frac{r^2}{Pr(f=1)}$ and $\frac{r}{(\#\ of\ SV)/n}$. However, those tiny conflict zones are artifacts of those performance metrics; by making the denominators near-zero, these performance metrics will tend to become nearly infinite. This problem of a near-zero denominator is unique to these metrics and does not exist in conventional metrics, such as accuracy and F1 score.[1]

An alternative to cross-validation is a heuristic approach. In a recent article, Ghafoori et.al (2018) propose the following procedure for hyper-parameter selection and demonstrate its usefulness via a simulation study.[2]

1. Calculate the average distance of each observation to the nearest l neighbors,

   $s_i = \frac{1}{l}\sum_{i \in H_l} ||x_i - x_j||$, where H_l is a set of the nearest l neighbors;

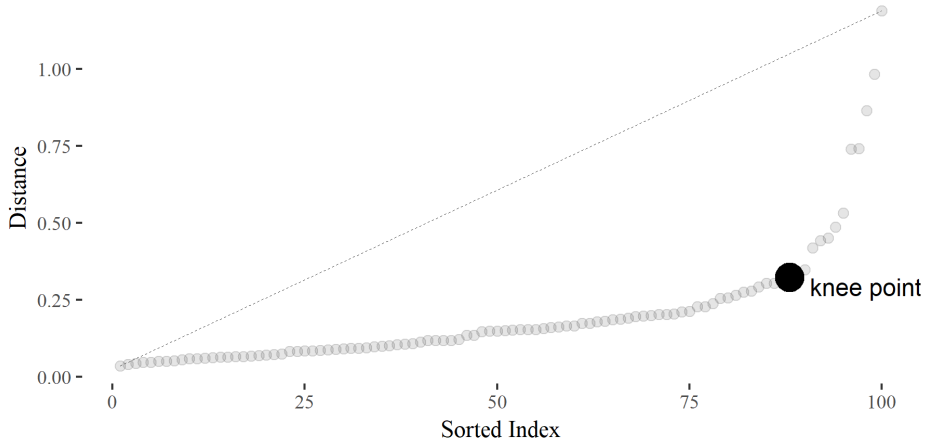
2. Sort the data from the smallest to the largest values of s_i;

[1] I also tried to add some value α to the denominators to address the near-zero denominator problem, but it turns out that the results are sensitive to the choice of α.

[2] For the other heuristic methods and their problems, refer to Ghafoori et.al (2018).

3. Find a knee point s_m in $\{s_i : i = 1, \dots, n\}$, that is, an observation that has the longest distance from the line connecting s_1 and s_n;

Figure A1.1. An Example of a Knee Point



NOTE: The figure shows an example of a knee point. The data contain two variables of 100 observations that are generated from an independent exponential distribution. The vertical and horizontal axes are the mean distances and the sorted indexes of the observations respectively.

4. The optimal ν is $\frac{n-m}{n}$;

5. The optimal γ is $\frac{1}{s_m}$.

Ghafoori et.al (2018) provide mathematical justifications for this procedure. Unlike the cross-validation techniques, this method does not have the problem of near-zero denominators and hence can be used even when conflict events tend to concentrate in a few locations. Furthermore, since the heuristic method does not require a grid search of the hyper-parameters, the method is much faster than the cross-validation methods. In the implementation to the UCDP GED, I set l as the 1% of conflict events. I emphasize that this choice is rather arbitrary; even though the OCSVM does not depend on the areal-unit assumptions, the OCSVM, or any one-class classification, is not totally free from arbitrary assumptions. Unlike the areal-unit assumptions, however, the parameter

l is readily interpretable; the parameter represents the prior expectation of a possible number of outliers (Ghafoori et al. 2018). This allows researchers to choose the hyper-parameter based on their substantive knowledge.

Implementation

Due to the problems discussed above, the cross-validation techniques result in unrealistically tiny conflict zones, which are of little use in application. The results in the main paper and the new conflict zone dataset are therefore based on the heuristic method proposed by Ghafoori et al. (2018). Because Ghafoori et al.'s method relies on the estimation of a knee point, and because a knee point cannot be estimated precisely with small sample sizes, I limit the cases to conflict episodes that have more than three conflict events. I also suspect that conflict zones are not well defined for such infrequent conflicts.

Supporting Information 2. Preprocessing of the UCDP GED

Because the conflict events in the UCDP GED are different in their spatial and temporal precision (for some events, we know the exact locations and even dates, while we only know broader administrative units or months for the other events), I resample the locations and dates of conflict events so as to account for the reporting precision. I first resample the locations of conflict events, based on their spatial precision, as seen in Table A2.1.

Table A2.1. Spatial Precision and Resampling of Conflict Events

Precision		Resampling
1	The exact location is known.	No resampling.
2	An event occurred within 25km radius of a known point.	Resampling within a 25km radius of the reported location.
3	Only the second-order administrative unit is known.	Resampling within the second-order administrative unit.
4	Only the first-order administrative unit is known.	Resampling within the first-order administrative unit.
5	An event occurred along line features, such as rivers and roads.	Resampling within the country.
6	Only known at a country level.	Resampling within the country.
7	An event occurred in international waters or airspace.	Dropped from analysis.

NOTE: The left column shows the spatial precision of conflict events in the UCDP GED. The right column shows the resampling strategy for each level of spatial precision.

Note that the precision 6 and 7 are very rare in the UCDP GED and hence unlikely to change the results. The dates of conflict events are also resampled within a range of event dates recorded in the UCDP GED. The spatial and temporal resampling is repeated 100 times with bootstrapped samples. After the data are standardized (min-max scaling), the OCSVM is fitted to each of the resampled data.

Supporting Information 3. Bootstrapping

I calculate the confidence intervals via parametric bootstrapping. For each of the 100 spatio-temporal resamples (which is detailed in Supporting Information 2), the OCSVM is fitted, and the decision values are calculated. I then calculate the means and standard deviations of the decision values in each location. With the assumption of normality, the 95% upper and lower bounds of the decision values are calculated. If a decision value is more than zero, the location is predicted to be a part of a conflict zone.

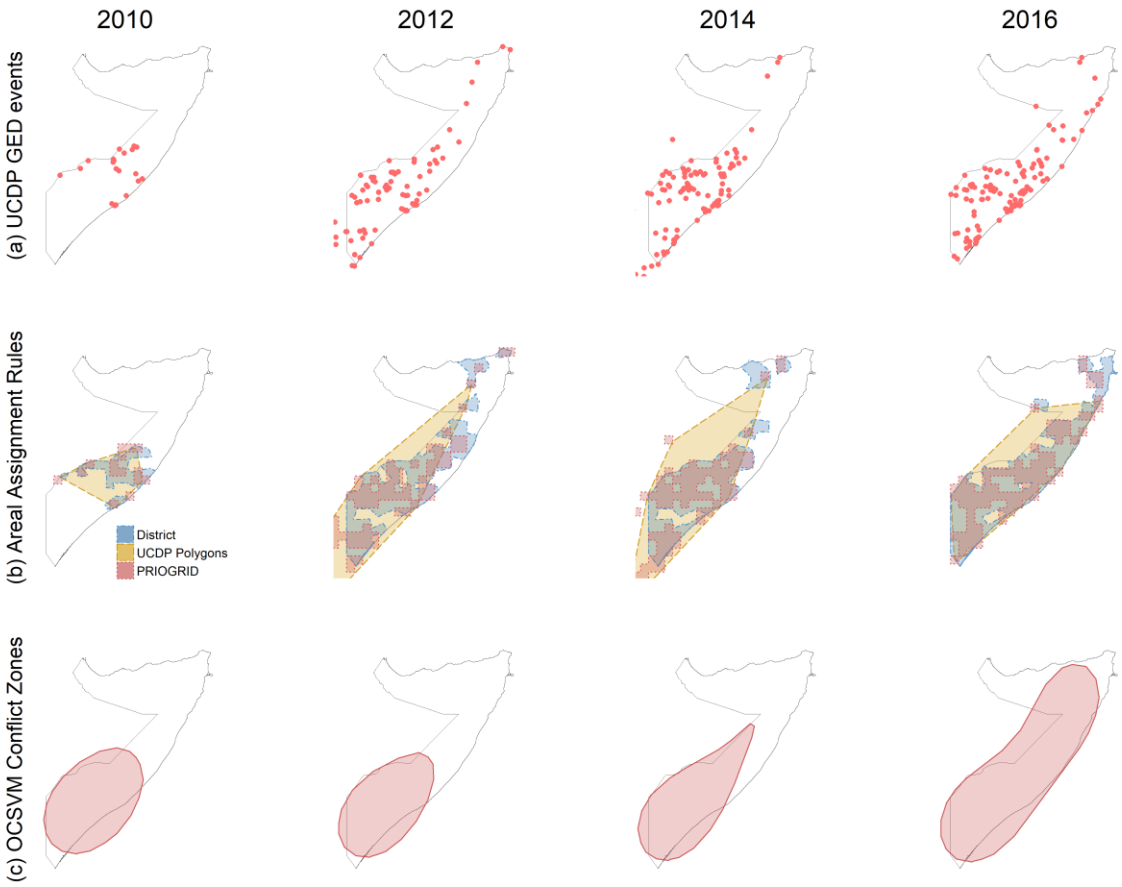
The parametric bootstrapping is particularly useful in the implementation of the UCDP GED. Because the spatio-temporal resampling is computationally expensive, it is not feasible to generate a large number of bootstrapped samples. As a result, a "hard" ensemble[3] does not generate smooth conflict zones. With the aid of the normality assumption, the parametric bootstrapping provides smooth conflict zones. Admittedly, this procedure rests on the normality assumption, but I believe this is the best feasible method of predictive inference in this application.

[3] In a "hard" ensemble, for each of the 100 resamples, the OCSVM is fitted and a binary prediction is made. If over a majority of the 100 predictions are $y = 1$ in a given location, the location is predicted to be a part of a conflict zone. The lower and upper bounds of the estimates are obtained by collecting a set of locations which more than 2.5% or 97.5% of the estimates classify as a part of a conflict zone. This procedure usually requires a larger number of resamples.

Supporting Information 4. Conflict Zones over Time

The following figure (Figure A4.1) shows how the distributions of conflict events and conflict zones changed over time in the case of the Somali Civil War for the period of 201-2016 (in the later half of the 2010s, there are not so many conflict events, and hence the conflict zones are empty). As seen in the first row of the figure, the fighting locations are mostly centered around the central region, in which the state capital Mogadishu is located. The conflict then expanded to the south and north in the later 2010s. The existing methods, as shown in the second row of Figure A4.1, more or less reflects those trends, though the convex hull method yields conflict zones that include large parts of Ethiopia, in which there is no conflict events. The OCSVM estimates (the third row of Figure A4.1) also reflect those overall trends, but the changes are smoother; unlike other zones, the OCSVM estimates are not heavily influenced by a single event, and they capture the overall trends in the distribution of conflict events. This is not surprising as the OCSVM explicitly account for temporal dependency.

Table A4.1. Spatial Precision and Resampling of Conflict Events



NOTE: The first row shows the distribution of conflict events relating to the conflict episode of the Somali Civil War. The second row shows the conflict zones estimated by the district assignment, the assignment of the PRIOGRID cells, and OCSVM.

Supporting Information 5. Replication of Daskin and Pringle (2018)

The following table (Table A5.1) shows (a) the original estimate of Daskin and Pringle (2018), (b) the estimate replicated with the old version of the UCDP Polygons, (c) the estimate replicated with the updated version of the UCDP Polygons, and (d) the estimate replicated with the OCSVM-based conflict zones. The difference between (a) and (b) reflects differences in data handling and computation of conflict frequency measures. The difference between (b) and (c) reflects differences between the old and updated versions of the UCDP Polygons. The difference between (c) and (d) reflects differences between the convex hull method and OCSVM. As seen in the table, the estimate becomes different and statistically insignificant only when I use the OCSVM-based conflict zones. In other cases, the estimates indicate even larger effect sizes. [4]

Table A5.1. Replication of Daskin and Pringle (2018): Full Results

(a) Original Estimate	(b) Replication with the old UCDP Polygons	(c) Replication with the updated UCDP Polygons	(d) Replication with the New Conflict Zones
-0.57	-0.87	-0.85	-0.10
[-0.86, -0.29]	[-1.23, -0.51]	[-1.20, -0.49]	[-0.40, 0.21]

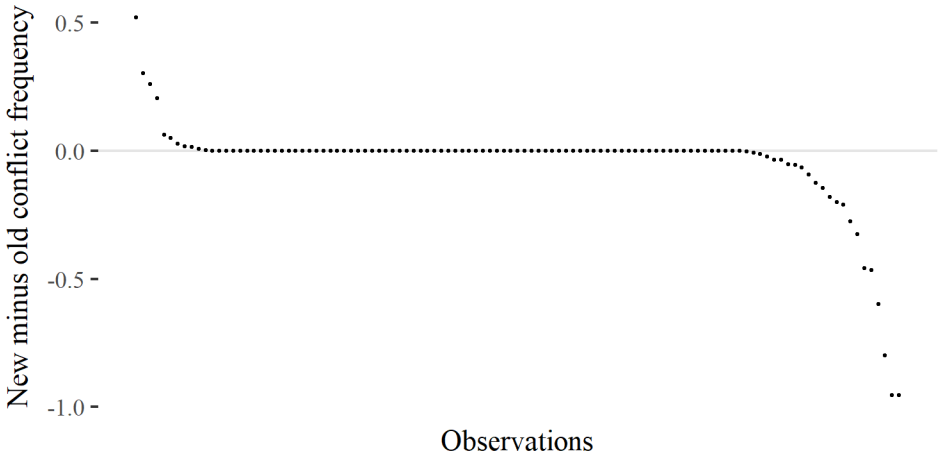
NOTE: The table shows the regressions of mammal population trajectories on the average proportions of conflict areas in protected areas in Africa. The column (a) to (d) show the original estimate, the estimate with the old version of the UCDP Polygons, the estimate with the updated version of the UCDP Polygons, and the estimate with the OCSVM-based conflict zones. In each column, the regression coefficient and corresponding 95% 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.

[4] Unfortunately, even with my best efforts, I cannot make an exact replication of Daskin and Pringle (2018)'s conflict frequency measure. Because the codes for the data compilation and GIS operations are not available, I cannot analyze what causes the differences.

Why Do the Estimates Become Different?

A question that is not directly addressed in the main paper is why the new conflict zones change Daskin and Pringle's findings. In this section, I answer this question by looking at individual cases. The following figure (Figure A5.1) maps the differences between the original and new conflict frequency measures. As seen in the figure, there are several observations for which the conflict frequencies take different values even though the differences are negligible for the majority of cases. This means that neither UCDP Polygons nor OCSVM systematically over- or under-states the conflict frequencies.

Figure A5.1. Differences in the Conflict Frequency Measures

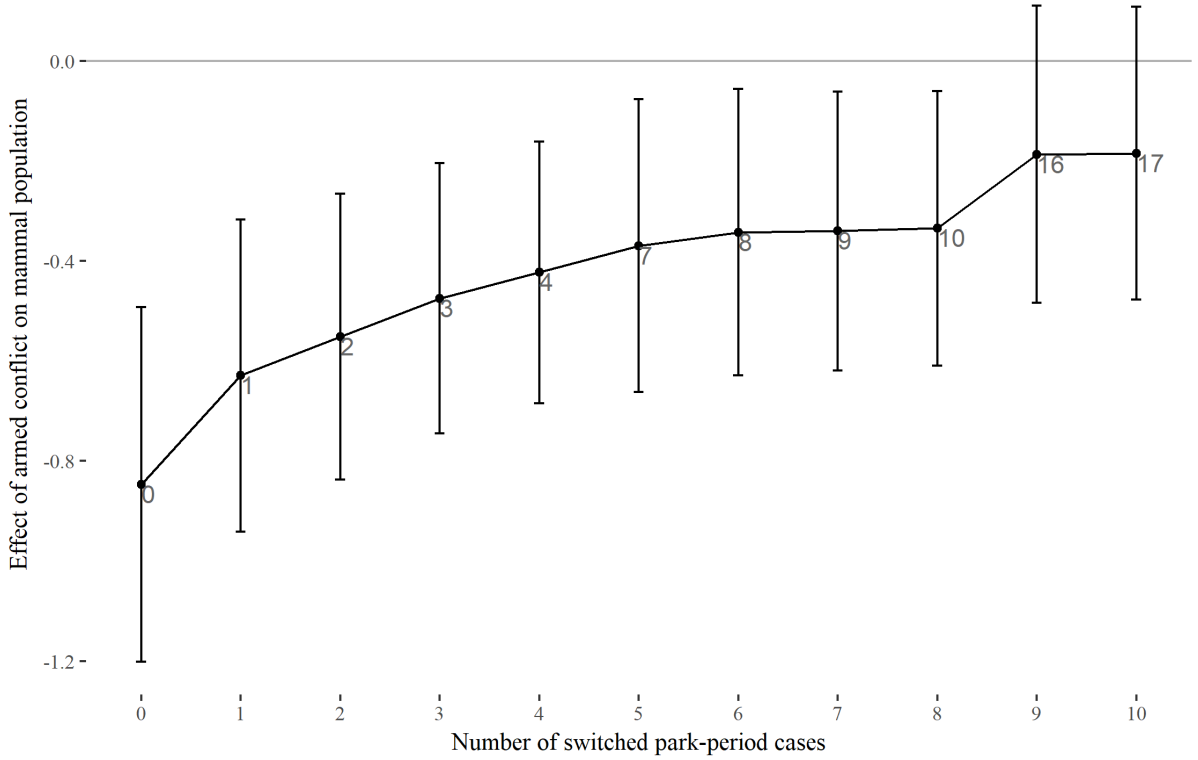


NOTE: The vertical axis shows the conflict frequencies based on the OCSVM estimates minus those based on the updated UCDP Polygons. The park-period cases are sorted from the largest to the smallest values of the differences.

In the following figure (Figure A5.2), I repeat Daskin and Pringle (2018)'s regression analysis while replacing the conflict frequencies of the top m different park-period cases with the new OCSVM-based conflict frequencies. For example, the second interval to the right is the estimate and confidence interval when I use the conflict frequency measure based on the updated UCDP Polygons except for the most different park-period case, for which the OCSVM-based

conflict frequency measure is used. Because there can be multiple park-species observations for a given park-period case (that is, multiple species are observed for a given park-period case). I also add the corresponding number of park-species observations for which the new OCSVM-based conflict frequency measure is used. As seen in Figure A5.2, replacing the conflict frequencies of nine different park-period cases—which corresponds to sixteen park-species observations—is sufficient to overturn the main estimate of Daskin and Pringle (2018).

Figure A5.2. Estimates with Switched Observations



NOTE: The vertical axis shows the estimated effect of conflict frequencies on mammal population. The horizontal axis shows the number of park-period cases for which the OCSVM-based conflict frequency measure is used (for the rest, the measure is based on the updated version of the UCDP Polygons). The numbers annotated in the figure is the number of park-species observations for which the OCSVM-based conflict frequency measure is used (for a given park-period case, there can be multiple park-species observations).

The following table (Table A5.2) presents the details of the nine most different cases. The first four columns show the country and park names as well as the start and end years of species observations. The last five columns indicate the population trajectory (λ), the conflict frequency measure with the updated UCDP Polygons, the OCSVM-based conflict frequency measure, and whether there is any qualitative evidence for the presence of armed conflict. The qualitative evidence is based on academic literature and non-academic reports. The absence of qualitative evidence means either the absence of conflict in a national park, the absence of information about actually-existing conflict, or my failure to find the information. As I emphasize in the paper, it is much more difficult to prove the absence than showing its presence. Moreover, because there was less attention devoted to the security conditions in protected areas of Africa in the 1990s and early 2000s, the qualitative inquiry must be compromised. My strategy is therefore to focus on the evidence about the presence of conflict and balance it to the overall tendency of the zoning methods.

Table A5.2. Nine Most Different Cases

Country	Park	Start	End	λ	Conflict (Convex Hull)	Conflict (OCSVM)	Qualitative
Mozambique	Marromeu	1990	1994	0.60	0.60	0.08	No
Zimbabwe	Gonarezhou	1989	1993	0.50	0.30	0.00	No
Uganda	Queen Elizabeth	1992	2002	1.13	0.36	0.68	No
South Africa	Addo-Elephant	1990	1993	1.06	0.42	0.88	No
Zimbabwe	Dande	1989	2003	1.02	0.00	0.47	Yes
Uganda	Kidepo Valley	2005	2008	1.03	0.00	0.60	Yes
Zimbabwe	Chewore	2001	2010	1.06	0.00	0.80	No
Kenya	Kerio Valley	1997	2002	0.95	0.00	0.95	Yes
Uganda	Murchison Falls	1995	2005	1.04	0.04	1.00	Yes

NOTE: The table shows park-period cases that have the first to ninth largest absolute differences in the original and new conflict frequency measures. The first four columns are the names of countries, protected areas, and start and end years of species records. The last four columns are the population trajectories, the conflict frequencies based on the UCDP Polygons dataset, those based on the OCSVM estimates, and existence/absence of qualitative evidence for conflict presence.

Regarding the Dande Safari Area, Kidepo Valley National Partk, Kerio Valley National Reserve, and Murchison Falls National Parks, the OCSVM predicts the presence of conflict, and I am able to find qualitative evidence for the predictions. The qualitative sources refer to the presence of conflict in the protected areas;

> "[In 2000,] there have also been occasional conflicts, as well as rumors that the party leadership [ZANU-PF] feared the Association [the Zimbabwe National Liberation War Veterans Association]'s leadership to become too powerful. The farm invasions were accompanied by plenty of violence against members of the opposition" (Spierenburg 2004 p.39).[5]
>
> "For decades [decades before 2014] there was a sustained conflict between neighbouring tribes and the people and wildlife in Kidepo. Heavily armed raiders attacked villages to steal cattle, and also poached in the park" (Almond 2014 p.71);
>
> "In March 2001, a particularly fierce raid took place, with a Pokot attack on Murkutwo village in Marakwet District [a district within the Kerio Valley] leaving at least 40 people killed" (Elfversson 2016 p.2073);
>
> "Gunman from the Lord's Resistance Army shot dead a British tourist yesterday in an ambush in the north of the country. According to the Foreign Office, the man had gone to help a group of tourists

[5] In a chapter about the history of Dande.

> whose vessel capsized in the Murchison Falls national park in north-east Uganda" (Howden 2005).

Although the UCDP Polygons tend to be more inclusive than the OCSVM-based conflict zones, the UCDP Polygons actually fail to include those four protected areas due to its outlier detection; under the 20%-5% rule of the UCDP Polygons, conflict events near the protected areas are considered as outliers, and hence the conflict zones do not contain those protected areas. The OCSVM, by contrast, takes a more statistically systematic approach to the outlier detection and hence accurately include the three protected areas inside the conflict zones.

For the remaining protected areas, Gonarezhou National Park, Queen Elizabeth National Park, Addo-Elephant National Park, Marromeu National Reserve and Chewore Wildlife Safaris, I do not find qualitative evidence for the presence of conflict. I cannot definitely state whether this is due to the absence of conflict, the absence of information, or my failure to find information. Furthermore, the Queen Elizabeth National Park and Addo-Elephant National Park are especially difficult cases as both UCDP Polygons and new conflict zones predict some degrees of conflict presence, but they disagree with its extent. Given these facts, I am hesitant to draw a definite conclusion from these cases.

Overall, the OCSVM-based conflict zones are usually smaller and tighter (so the OCSVM does not predict too many $Y = 1$) but they still accurately assign the four protected areas to conflict zones. By contrast, the UCDP Polygons are generally more inclusive, but they fail to correctly classify those protected areas. Thus, it appears that the OCSVM provides a more accurate picture of conflict zones, and that the differences in the regression estimates are driven by the misclassification by the UCDP Polygons.

Although we must be careful about making a generalization from a single replication, a practical implication is that the conflict zones should be used for the purpose of macro-level comparison. Even though Daskin and Pringle (2018)'s analysis is mostly based on macro-level comparison, quite a few protected areas are small and hence subject to measurement errors. The problem is further compounded by the small sample size; even though average measurement errors would approach zero as the number of observations increases (though it will still cause attenuation biases in regression analyses), such convergence does not occur with a small sample. In fact, Daskin and Pringle (2018)'s sample contains 172 park-species observations, in which there are only 96 unique protected areas. As a result, even though the measurement errors seem to have no systematic pattern, just switching the conflict frequencies in the nine protected areas can overturn the results. Thus, it is advised for future studies to check the sizes and number of geographical units of analysis.

Supporting Information 6. Replication of Beardsley et al. (2015)

I also replicate Beardsley et al. (2015), which uses the UCDP Polygons to measure the movements of rebel groups and analyzes the effect of rebels' ethnic claim and military strength on the degrees of their movements. The authors measure the rebels' movement by calculating the intersection of conflict zones in t and $t-1$ (where t is a given year) and taking the proportion of the intersecting areas in the $t-1$ conflict zones.[6] Using 257 rebel groups as units, the authors find that rebels move across locations when they do not make ethnic claims and when rebels are weaker than a government in its military capabilities.

I re-estimate their main model (Model 1 in Table 1 in Beardsley et al. 2015) with their original variable (exact replication), the replication with the old UCDP Polygons, the replication with the updated UCDP Polygons, and the replication with the OCSVM-based rebels' movement measure. Other specifications are the same as those in Beardsley et al. (2015).[7] The following table (Table A6.1) summarizes the results of the replications.

[6] The code for making the rebels' movement measure is not available. I am able to make a near-exact replication by taking the proportion of the intersecting areas in the conflict zones at a time t (instead of $t-1$).

[7] Refer to Beardsley et a. (2015) for details.

Table A6.1. Replication of Beardsley et al. (2015)

The effect of ethnic claim			
(a) Original Estimate	(b) Replication with the old UCDP Polygons	(c) Replication with the updated UCDP Polygons	(d) Replication with the New Conflict Zones
0.87	0.84	0.59	0.43
[0.11, 1.62]	[0.10, 1.59]	[-0.04, 1.22]	[-0.31, 1.17]

The effect of rebels' weakness			
(a) Original Estimate	(b) Replication with the old UCDP Polygons	(c) Replication with the updated UCDP Polygons	(d) Replication with the New Conflict Zones
-0.70	-0.73	-0.78	-0.19
[-1.29, -0.12]	[-1.32, -0.14]	[-1.30, -0.27]	[-1.01, 0.64]

NOTE: The table shows the regressions of conflict zone overlaps on rebels' ethnic claim and their military weakness relative a government. The models are quasi-binomial regressions estimated with MLE. The upper and lower panes show the estimated effects of rebels' ethnic claim and their military weakness. The column (a) to (d) show the original estimate, the estimate with the old version of the UCDP Polygons, the estimate with the updated version of the UCDP Polygons, and the estimate with the OCSVM-based conflict zones. In each column, the regression coefficient and corresponding 95% confidence intervals are reported. The control variables and other specifications are the same as those in Model 1 of Table 1 in Beardsley et al. (2015). n=257.

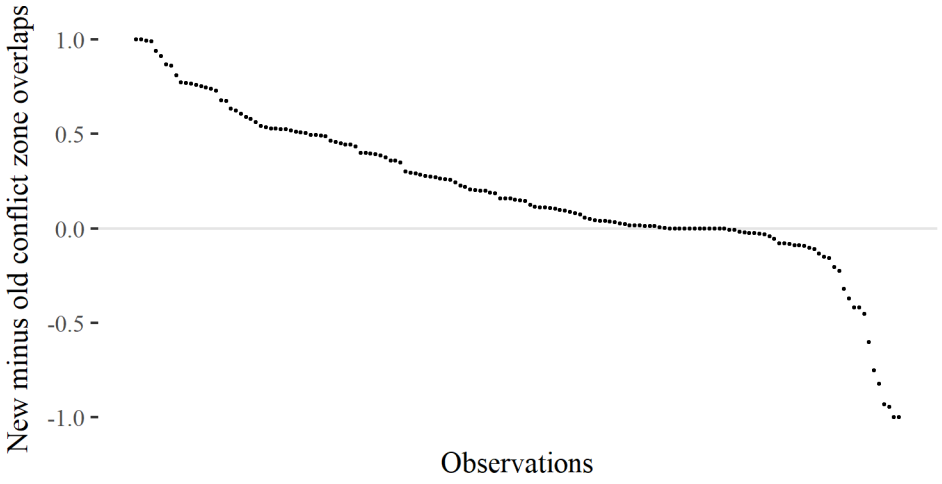
The estimates replicated with the old or updated version of the UCDP Polygons only make small changes. Only when I use the updated version of the UCDP Polygons, the effect of ethnic claim becomes smaller, though the effect is still statistically significant at a 0.1 level. These results are, however, overturned with the OCSVM-based measure; neither the effect of ethnic claim nor the effect of rebel's military weakness is statistically significant, and the effect sizes become smaller.

These differences are mostly explained by the fact that the OCSVM accounts for temporal dependency and hence allows smoother changes in conflict zones.[8] By contrast, the UCDP

[8] Note that the OCSVM *allows*, not *assumes*, the temporal dependency. If locations of conflict events suddenly change, the OCSVM also allows sudden changes in conflict zones.

Polygons create zones for each year, assuming that there would be no temporal dependency across years (thus, the conflict zones can be different whether an event occurred on 31st of December or 1st of January). As a result, as seen in Figure A6.1, the UCDP Polygons tend to overstate the temporal variation in rebels' mobility. In fact, in nearly 70% of the observations, the UCDP Polygons have smaller overlaps in conflict zones (thus higher degrees of rebels' movement). The large temporal variation results in a larger variance in the outcome variable and hence larger effect sizes in the regression analysis.

Figure A6.1. Differences in the Conflict Zone Overlaps



NOTE: The vertical axis shows the conflict zone overlaps based on the OCSVM estimates minus those based on the updated UCDP Polygons. The observations are sorted from the largest to the smallest values of the differences.

Although I must be careful about drawing implications from a single replication, it appears that conflict zones can sway empirical findings when we use the zones for creating variables. Both Beardsley at al. (2015) and Daskin and Pringle (2018) use the conflict zones for the purpose of measurement. While Daskin and Pringle (2018)'s findings are sensitive to measurement errors in a few observations, Beardsley et al. (2015)'s findings are subject to systematic measurement errors. Although Daskin and Pringle (2018)'s problem might be addressed by increasing the sample size,

the same solution may not be applicable to Beardsley et al. (2015); there is a systematic tendency in the measurement errors (larger variance in the outcome variable), and such a tendency will not disappear even with a large sample size. Thus, it is advised for future studies to consider whether zoning methods and their underlying assumptions can cause systematic measurement errors and how the measurement errors can bias empirical estimates.

Supporting Information 7. Replication of Fjelde and Hultman (2014)

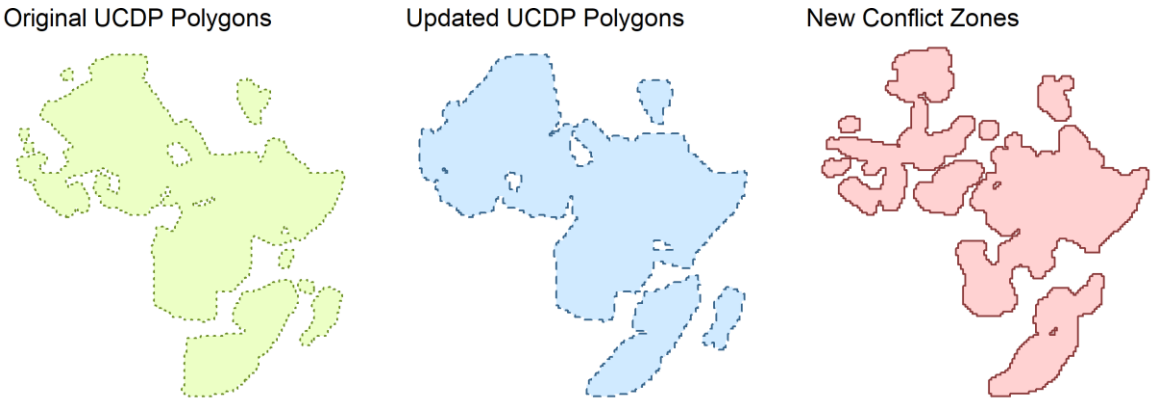
While both Daskin and Pringle (2018) and Beardsley et al. (2015) use the conflict zones for creating variables, the conflict zones can be used for other purposes. Fjelde and Hultman (2014) is such an example; the authors use the conflict zones for selecting relevant observations. Fjelde and Hultman (2014) analyze the effect of ethnic constituencies on violence during civil war. Because the authors are interested in dynamics in conflict-affected locations, they limit the observations to those within conflict zones. The units of analysis are the pairs of the PRIOGRID cells and years. The PRIOGRID cells are limited to those within the zones of state-based conflict in the static version of the UCDP Polygons. Using the sample, the authors find that there are a larger number of violence against civilians in enemies' ethnic constituencies.

I replicate the main models of Fjelde and Hultman (2014; Model 1 and 3 of Table 1 in their article) by using the PRIOGRID cells within the old version of the UCDP Polygons, the updated version of the UCDP Polygons, and the OCSVM-based conflict zones. Other specifications, including control variables and regression models, are the same those in the original analysis.[9] The following figure (Figure A7.1) compares those three conflict zones in Africa.[10] As seen in the figure (although the old and updated versions of the UCDP Polygons are similar) the OCSVM-based conflict zones are more contained.

[9] Refer to Fjelde and Hultman (2014) for details.

[10] Because the old version of the UCDP Polygons includes only Africa, I also limit the observations to those in Africa.

Figure A7.1. Comparison of Conflict Zones



NOTE: The figure shows the static version of conflict zones in Africa. The first to third panes correspond to the old version of the UCDP Polygons, its updated version, and the OCSVM-based conflict zones.

The following table (Table A7.1) presents the results of the replications. As seen in the table, the results are similar regardless of the underlying conflict zone data. Although the analysis with the OCSVM-based conflict zones indicates smaller effect sizes, it does not alter the statistical inferences. These results are not surprising given the large number of observations. With the samples selected by the UCDP Polygons, there are more than 70 thousand cell-year observations, and even with the OCSVM-based sample, there are over 60 thousand observations. The large sample size ensures that the findings would not be heavily influenced by a small portion of observations.

Table A7.1. Replication of Fjelde and Hultman (2014)

The effect of rebels' ethnic constituency on government's violence			
(a) Original Estimate	(b) Replication with the old UCDP Polygons	(c) Replication with the updated UCDP Polygons	(d) Replication with the New Conflict Zones
1.78	1.85	1.95	1,52
[1.15, 2.42]	[1.23, 2.48]	[1.32, 2.58]	[0.89, 2.15]

The effect of government's ethnic constituency on rebels' violence			
(a) Original Estimate	(b) Replication with the old UCDP Polygons	(c) Replication with the updated UCDP Polygons	(d) Replication with the New Conflict Zones
0.87	0.88	0.96	0.75
[0.18, 1.56]	[0.19, 1.57]	[0.26, 1.66]	[0.03, 1.47]
n=70,185	n=72,418	n=76,871	n=61,653

NOTE: The table shows the regressions of violence against civilians on rebels' and government's ethnic constituencies. The models are negative binomial regressions estimated with MLE. The upper and lower panes show the estimated effects of ethnic constituency on enemy's violence. At the bottom, the number of observations are reported. The column (a) to (d) show the original estimate, the estimate with the old version of the UCDP Polygons, the estimate with the updated version of the UCDP Polygons, and the estimate with the OCSVM-based conflict zones. In each column, the regression coefficient and corresponding 95% confidence intervals are reported. The control variables and other specifications are the same as those in Model 1 and 3 of Table 1 in Fjelde and Hultman (2014).

Although one must be careful about making hasty generalizations, it seems that when conflict zones are used for sample selection and/or when the sample size is large, the estimates are less sensitive to the measurement errors in conflict zones. As illustrated by the replication of Daskin and Pringle (2018), when one uses conflict zones for the purpose of measurement and the measurement error is less systematic, the bias arises if the sample size is small. Similarly, when one uses conflict zones for measurement but the measurement error is systematic as seen in the replication of Beardsley et al. (2015), the bias can potentially persist regardless of sample sizes. By contrast, when use conflict zones for sample selection, the bias may not be large if the sample size is large. Although I am not certain about the biases when conflict zones are used for sample selection and the sample size is small, my conjecture is that the analysis with a small sample would

be sensitive to the inclusion or exclusion of a few observations, and hence the findings could be sensitive to the selection of conflict zones. It is a task of future research to conduct a large-scale replication analysis and hence to systematically analyze the biases due to the measurement errors in conflict zones.

Reference List

Almond, Gary. 2014. *Splendid Isolation*. https://www.wildplacesafrica.com/files/kidepo_valley_travel_report.pdf (accessed on 20 November 2019).

Beardsley, Kyle, Kristian Skrede Gleditsch, and Nigel Lo. 2015. "Roving Bandits? The Geographical Evolution of African Armed Conflicts." *International Studies Quarterly* 59 (3): 503–16.

Banerjee, A., P. Burlina, and C. Diehl. 2006. "A Support Vector Method for Anomaly Detection in Hyperspectral Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 44 (8): 2282–91.

Elfversson, Emma. 2019. "The Political Conditions for Local Peacemaking: A Comparative Study of Communal Conflict Resolution in Kenya." *Comparative Political Studies* 52 (13–14): 2061–96.

Fjelde, Hanne, and Lisa Hultman. 2014. "Weakening the Enemy: A Disaggregated Study of Violence against Civilians in Africa." *Journal of Conflict Resolution* 58 (7): 1230–57.

Ghafoori, Z., S. M. Erfani, S. Rajasegarar, J. C. Bezdek, S. Karunasekera, and C. Leckie. 2018. "Efficient Unsupervised Parameter Estimation for One-Class Support Vector Machines." *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Howden, Daniel. 2005. "Rebels Murder British Tourist in Uganda Park." *Independent* (9 November 2005).

Koumanelis, Mimi. 2006. "Widespread Elephant Slaughter Discovered in Chad – National Geographic Partners Press Room." *National Geographic,* August 30, 2006.

http://press.nationalgeographic.com/2006/08/30/widespreadelephantslaughterdiscoveredinchad/ (accessed on1 January 2018).

Lee, Wee Sun, and Bing Liu. 2003. “Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression.” In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 448–455. ICML’03. Washington, DC, USA: AAAI Press.

Poilecot, Pierre, Etienne Bemadjim N’Gakoutou, and Nicolas Taloua. 2010. “Evolution of Large Mammal Populations and Distribution in Zakouma National Park (Chad) between 2002 and 2008.” *Mammalia* 74 (3): 235–246.

Spierenburg, Marja. 2004. *Strangers, Spirits, and Land Reforms: Conflicts about Land in Dande*, Northern Zimbabwe. Leiden, Netherland: Brill Publishers.