A Morden Approach to Conflict Prediction
or: Dude, Here's Your Conflict
or: Where Men Rebel

Simon Polichinel von der Maase

September 2018

Contents

Abstract

To come..

1 Introduction

This paper presents a modern computational approach to conflict prediction and forecasting. It is constructed as a unified framework to asses the probability that a given sub-national geographical location will experience battle-related deaths as a consequence of intra state conflict. The challenge is handled as a forecasting problem, thus the prediction target is whether or not the given location experiences any battle-related deaths *next year*. Or in mathematical terms at $t + 1$. The given geographical unit is a PRIO-grid cell, which is a cell of 0.5×0.5 decimal degrees. I am not trying to estimate the number of deaths; only the presence of deaths. Thus the target is binary, taking either the value of 1 or 0 with 1 denoting the presence of fatal conflict. In mathematical terms; $target \in \{0, 1\}$. Importantly, however, the estimate will not be binary but probabilistic; between 0 and 1. In mathematical terms; $target \in [0, 1]$.

This is a preliminary project, thus the aim is as much to explore fruitful approaches and methods for future research as it is making good predictions. As a consequence a lot of this paper is dedicated to choosing what goes into the model and evaluating what comes out.

The model itself consists of an ensemble of Extreme Gradient Boosting classifiers (xgboost), which is an special boosting algorithm based on ensembles of regression tress well suited for both large date-sets and rare events detection.

What goes in to the model is a roster of features derived from both the theoretical and empirical literature regarding civil wars and intra-state conflicts. The relationship does not need to be causal, but I do demand a salient theoretical link between the target and the features - which would make the feature useful in a practical long-term forecasting scenario.

What comes out of the model is first of all probabilities of conflict deaths in a given geographical cell at a given year which are evaluated against the actual observations. But just as importantly the xgboost algorithm allows for evaluation of the individual features importance in the prediction process. This insight is paramount when deciding how to improve the framework and where to invest resources in future endeavours.

The model constructed is remarkable reliable in out-of-sample prediction with an AUC (ROC) of XX. Even more impressive is the fact that i captures all most all events, with a recall of XXXX when applying a threshold of 0.5 - that is when I classify all events with a predicted probability of conflict above 0.5 as conflicts. Unfortunately, when faced with the unbalanced nature of the real world the model do generate relatively many false positives as underscored with a precision at XXXX and an average precision (precision recall curve) of XXXX. Never-the-less the model does reach a accuracy of XXXX; and that is while still capturing most of the rare events. Presented in a less technical jargon: If I categories all predictions with a probability of conflict of 0.5 or above

as predicted conflicts then given all the conflicts - pertaining to some given year - my model will on average correctly classify XXXXX%(recall) of the conflicts which will occur the following year. However these events only make op XXXX% of events classified by the models as conflicts. Thus, I do predict a more dangerous world then we actual do live in. Still while the number of false positives is rather high compared to true positives it is rather low compared to the total number of observations and thus model is over all correct in its assessment XXXXX% (accuracy) of the time.

The Features credited with most of the prediction power are generally features pertaining directly to the temporal and spatial dimensions of conflict; that is how far is a cell from the nearest conflict; How many conflicts have the cell previous seen; How many fatalities where 'this' year reported in the country which the cell belongs to ect.. This is followed by features pertaining to the population size of the cell and the size of the country which the cell belongs to. Then grievance based features enters; such as whether the cell is economically deprived relative to the country as a whole and whether the cell is inhabited by politically excluded ethnic groups. Lastly come features pertaining to greed and resources such as absolute economic capacity and the presence of oil.

Going forward with future endeavours I recommend allocating resources to create a more unified and systematic framework modeling the temporal-spatial evolution of conflicts as a function of it self. For one thing it seems the most productive way to improve prediction power, but just as important the conflict data are more often updated and the last entry closer to the presen year then the contextual features used in this project. Thus the conflict data itself presents itself as better suited for real life forecasting applications.

The to following sections will first present the motivation and then the research question and design. This is followed by a introduction to the data sources and a presentation of the included features. Next I present the predictive framework and a thorough analysis of the derived results. This is proceeded by a discussion regarding future challenges and suggestions for improvements. Lastly a conclusion sums up the main findings.

1.1 The 'Why'

> *[...] the estate of Man can never be without some incommodity or other; and [...] the greatest, that in any form of Government can possibly happen to the people in generall, is scarce sensible, in respect of the miseries, and horrible calamities, that accompany a Civill Warre;* (Hobbes, 1651, 128)

The perils and miseries of civil war and internal conflict have plague mankind all throughout

history. Presumably intra-state conflicts have been around as long as there have been states to exercise conflict within - and the last century has been no exception. Since the conclusion of the Second World War intra-state conflicts have been far more common then inter-state wars (Collier and Hoeffler, 2004, 563); Over five times as many people have died in intra-state conflicts compared to inter-state wars (Collier and Hoeffler, 2004, 563); and since 1960 over one half of all nations have experienced some sort of violent internal conflict leading to fatalities (Blattman and Miguel, 2010, 3-4).

Importantly, internal conflicts should not be viewed as internal affairs of little concern to other then the inflicted host and its allied. Examples of spillover effects facilitating the spread of conflicts across boarders a ample. At country level, having a country located in a conflict ridden neighbourhood have been shown to be a robust predictor of internal conflict (Hegre and Sambanis, 2006; Goldstone et al., 2010). Internal conflict is thus a highly destructive and potentially contentious malaise only to be ignored by the most imprudent or hostile observers. Understanding how internal conflicts originates and spreads in order to prevent or mitigate the destruction is indeed as crucial as ever.

Encouragingly, developments in statistical techniques, data availability and computational power makes the endeavour slightly more feasible with each passing year [1]. One recent development is the shift in focus from cross country comparison towards disaggregated analyzes on sub-country unites. Given the nature - and indeed definition - of intra-state conflicts, this development is a promising step towards a better understanding the phenomenon (Cederman and Gleditsch, 2009). The disaggregated approach is further enhanced by evermore accessible geospatial software and powerful new machine learning algorithms. As I will show these developments can aid us in predict future conflict zone as well as generate novel insight into the processes which accompanies and facilitates internal conflicts.

1.2 The 'How'

The project at hand can be summed up by two research questions:

First research question(Q_1)**:** To what extent is it possible to predict the geographic location of future intra-state conflicts using a modern computational approach.

Second research question(Q_2)**:** What phenomenons and feature presents themselves as the most important in this predictions effort, and what can be done to create an even more informative feature space in the future.

[1] Unless, of course, conflict is inherently shrouded in ontological uncertainty rather the epidemiological uncertainty as implied by Gartzke (1999)

To answer these questions I use data obtained from the Uppsala Conflict Data Program (UCDP) (Sundberg and Melander, 2013; Croicu and Sundberg, 2017). This data includes counts of conflict deaths along with both coordinates of the scene and estimated time of the event. The coordinates a linked up to specific geographical cells of 0.5.5 decimal degrees derived from the PRIO grid database (Tollefsen et al., 2012). I aggregate the conflict data to sum up the yearly number of conflict deaths in a given cell. This measure i further dichotomized such that it only indicates whether or not a given cell experienced conflict deaths or not. I then 'lead' the measure, effectively lagging all explanatory features to come. Or put another way; I shifts the target feature one year behind such that the explanatory features of e.g. 2006 will try to predict conflicts in 2007. To further mimic the forcasting nature of the problem the model is created only on the basis of data from 1990 through 2005. Data from 2006 and onward are reserved for model evaluation through out-of-sample prediction.

To create the explanatory features I borrow from both the Uppsala conflict data itself and from the large number of features available from the PRIO grid database. To mitigate overfitting, secure robustness and aid in the quest of new insights I constrain myself to create features which corresponds to theoretical salient phenomenons described in the larger literature on civil wars and internal conflicts. I do not damand the teoritical connection to be causal; it can be spurious as long as it comes before the conflict in time and that spurious link is theoretically meaningful. The aim is to predict conflict - not explain it. Due to limitation in the data available from the PRIO grid database the date only covers 1990 through 2010. This is naturally a big obstacle if the framework was ever to be applied in practice, and as such I shall return to this challenge in the discussion.

With target and predictors in place, I construct a predictive framework using an ensemble of xgboost algorithms. I use an ensembles of xgboost algorithms to generate a more robust results and to facilitated insights into the uncertainty inherent in the prediction effort. For robustness I construct is both a model including all observations and one including only new "cell-onsets".

Before moving on it should be noted that the phrase "To what extent [...]" present in Q_1 implies the caveat: " - Given the scope of the project and the limited computational resources at my disposal. As this is preliminary research I will spend more time evaluating the results and discussing how to improve future frameworks then I will training and fine tuning the model and optimizing hyper parameters. The next section will serve as a more complete introduction to the data source briefly presented above.

2 The Data Sources

The project at hand utilize two different data source. The he Uppsala Conflict Data Program (UCDP) (Sundberg and Melander, 2013; Croicu and Sundberg, 2017) and the PRIO grid(Tollefsen et al., 2012). The first subsection below presents the UCDP while the follow presents the PRIO grid database.

2.1 UCDP

Most central to the endeavour at hand lies - of course - the data regarding intra-state conflict itself. This data is obtained through the Uppsala Conflict Data Program (UCDP) (Sundberg and Melander, 2013; Croicu and Sundberg, 2017). Specifically I utilize the UCDP Georeferenced Event Dataset (GED) Global version 18.1 (UCDP, 2017). The dataset contains records of conflict fatalities and the corresponding coordinates. As mentions I utilized data from 1990 through 2010 but data from XXXX through XXXX is available in the database. Conflict fatalities are here defined as:

> *"An incident where armed force was used 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."*(Croicu and Sundberg, 2017, 38).

Further definitions regarding armed force, organized actor ect. can be found in (Croicu and Sundberg, 2017, 10-11).

The project at hand limits itself to intra-state conflict, thus I only included incidents which do *not* include two different nations as the organized actors[2]. As presented already this data source provides the prediction target; the presence of conflict deaths in a given cell at a given year(+ 1). The data from UCDP is very detailed in regards to both temporal and spatial location, however for the endeavour at hand I aggregate the time unites to *years* and the spatial resolution to the 0.5×0.5 decimal degrees grid cells provided by the PRIO grid.

As I will show later on, many of the most important predictive features are also derive from this data source; e.g. the distance from a specific cell to the nearest conflict and the number of previous conflicts in a given cell. The fact that conflict patterns do them selves hold a lot of prediction power regarding future conflict zones might not surprise the reader, but I find the implications of this insight rather consequential in regards to future improvement of the framework, which I shall

[2]Naturally the distinction between a intra-state conflict and a proxy war can be very hard to uphold in practise. Never-the-less, though proxy wars are effectively between two states, the phenomenon is arguable more similar to intra-state conflicts and civil wars then to all out open warfare between to nations

return to in the discussion.

2.2 PRIO

Given the geo-referenced nature of the UCDP data, the number of interesting data sources one could enhance it with are virtually endless. However, the aggregation of various geo referenced and geo-spatial data accompanied with the appropriate grid construction and feature engineering can be a time consuming endeavour. While it is certainly a interesting and most likely fruitfully undertaking, it is one which will be saved for less preliminary work then the project at hand. Conveniently the Peace Research Institute Oslo (PRIO) has created what they call a "unified spatial data structure" (Tollefsen et al., 2012, 1). More specifically they have divided the world - excluding Greenland and Antarctica - into grid cells of 0.5 x 0.5 decimal degrees. For each cell PRIO has gather a large selection of features, including economic, geographic and demographic features (Tollefsen et al., 2012). Naturally the PRIO GRID also extent itself across time, and the most current data includes cell-years from 1946 to 2015 - with some divergence in the data coverage across the years (Tollefsen et al., 2016). On one side more complete data is available for more recent years; especially after 1989. On the other side no or little data is avialbel for the most recent years. Thus I utilize data from 1990 through 2010. This data includes relatively few random missing observations which a handle as described in the appendix, subsection 8.1. Some variables a only account for in 5-years intervals. This is handled through cell-specific linear interpolation as illustrated and described in the appendix, subsection 8.2.

The PRID grid is constructed as geo-spatial data and primed for collaboration with the UCDP data. As such merging and handling these two data source is a trivial task. For now I refrain from the temptation to included all variables available through the PRIO grid data base, and instead chose handpick features which are most often and constantly associated with internal conflict in the corresponding litterateur. This is done for four main reasons. Firstly, it is an effort to reduce potential noise and overfitting. Notably, the the xgboost algorithm utilized is essentially a self regularizing ensembles of decisions tress and thus it's performance should not be overly burden by being presented with a large feature spaces including irrelevant features; never-the-less it is a prudent precaution. A second reason for the initial handpicking is also to serve convertibility; a smaller feature space i easier communicated to the reader. Thirdly, having less raw features to focus on allows for more time spend on feature engineering, manipulating the raw features from the database to more theoretical sound features corresponding to the actual mechanisms purposed in the literature. It should be noted that since the xgboost algorithm utilizes decisions tress it is capable of constructing relevant interactions it self. Thus I shall not manually create any interaction links. Fouth; initial test showed that most of the prediction power comes from a very limited number of features and no amount of extra features succeeded in improving the model further.

The following section will present the included features and also discuss some notable absentees.

3 The Included Features

In this section I will introduce the roster of features used in the endeavour at hand. Some features are readily available from on of the two data sources, other requires some feature engineering before they correspond the the theoretical phenomenon believed to be connected the internal conflicts. I draw on insights from both the country aggregated civil war literature and the more resent disaggregated conflict litterateur. As such some features are cell specific, while others are country specific, and yet others denotes the difference between specific cell and country values. While it would be satisfying to present the features through a comprehensive literature review thus also presenting the theoretical justification, the scope and focus of this endeavour does not allow such academic gluttony[3]. The appendix, subsection 8.3 presents more comprehensive theoretical justifications and background context for all features below along relevant mathematical definitions. Here, in the main text the reader will have to settle for following summary.

Wealth and Capacities: Reflecting the arguments put forth in Collier and Hoeffler (1998); Fearon and Laitin (2003); Collier and Hoeffler (2004) but here with light night emission as a proxy for wealth as proposed by Elvidge et al. (2009), Chen and Nordhaus (2011) and Cederman et al. (2013). I have included both a cell-specific and country specific measure both divided by the populations count of the specific country/cell.

Inequality and deprivation: Reflection the classic argumentation put forth by Gurr (1970) which has been revitalized by Cederman et al. (2013). Here i also utilize night light emission as a proxy for wealth and I included both the specific measure used in Cederman et al. (2013) and a more simple measure simply capturing whether a not a given cell is worse of than the median cell of the corresponding country.

Ethnicity and Exclusion: A dichotomous feature denoting whether or not the cell is inhabited by one or more excluded ethnic groups (e.i. discriminated or powerless) at any given year. Capturing the argumentation put forth in Cederman et al. (2011, 2013) the measures are originally from the GeoEPR/EPR data by Vogt et al. (2015).

[3]Readings which together do serve as a rather comprehensive review are Hegre and Sambanis (2006), Kalyvas (2007), Cederman and Gleditsch (2009) and Blattman and Miguel (2010)

Population size and density: The correlation between the size of a country's population and the risk of civil war have been found rather robust (Collier and Hoeffler, 1998; Fearon and Laitin, 2003; Collier and Hoeffler, 2004; Hegre and Sambanis, 2006). Furthermore (Fearon, 2004, 287) also finds that country population is correlated with longer civil wars. I include measure of both absolute population count for both cell and country.

Geography and Accessibility: Fearon and Laitin (2003) argues that rough terrain and mountains are natural obstacles hindering effective projection of state power. Hegre and Sambanis (2006) concludes that a feature for rough terrain is found to robustly positively correlated with civil war across a large number of model specifications.(Hegre and Sambanis, 2006, 526-529)[4]. The PRIO Grid includes a readily available feature measuring the proportion of mountainous terrain within the cell based on Blyth et al. (2002) which I utilize.

Distance to Power Center: Another natural hindering for projecting state power is sheer distance (Fearon, 2004; Buhaug et al., 2009; Cederman et al., 2009; Buhaug, 2010). To capture phenomenon this I have included is the distance to the nations capital[5], the travel time to the nearest major city, and the total size of the country (Tollefsen and Buhaug, 2015).

Trans-boarder Influences: A number of different mechanism have been proposed and explored (Blattman and Miguel, 2010, 29-30), the one explored here is rather simple and follows Hegre and Sambanis (2006). This is simply the distance to the nearest (land) boarder shared with another country.

Prime Commodities, Oil and the Recourse course: A lot have been written on this subject, but the empirical results have varied a lot (Collier and Hoeffler, 1998; Fearon and Laitin, 2003; Fearon, 2004; Ross, 2004; Collier and Hoeffler, 2004; Fearon, 2005; Buhaug, 2010; Hegre et al., 2009). In this endeavour I simply include a dichotomous feature denoting whether or not a given cell is known to hold oil deposits. Thus a cell is not denoted as having oil before oil is actually discovered.

Inertia, dispersion, traps and time trends: Conflict traps and inertia have been modelled be both Collier and Hoeffler (2004), Hegre and Sambanis (2006) and Cederman et al. (2013) while dispersion have been used by Goldstone et al. (2010). I included a number of features which aims to capture the pattern of conflict as it moves through space and time as a function of it

[4]Though see Goldstone et al. (2010)

[5]The measure utilized by Buhaug (2010)

self. Distance from cell center to nearest conflict, number of past conflict and fatalities in cell and number of fatalities in the country as a whole for the given year.

A lot more features and specifications where tried out[6] and even more could have been scrutinized. Furthermore some of the included features could be even better specified. Never-the-less, the features presented above will do for now. As mentioned the curious reader will find more elaborate subsections in the appendix corresponding to each paragraph above. There, some light theory is presented along the features thus introducing some much needed context. Here I proceed to presenting the model, the estimations and the results.

4 The Predictive Framework and the Results

In the follow sections I present the predictive framework and evaluate the performance of the framework through various metrics. One of the greater challenges pertaining to the project at hand is the imbalance of the data; most of the time most grid cells does not experience any conflict fatalities. The data as such is imbalanced and only around 1% of the data constitutes "events" - that is actual conflicts. I take a number of steps handle imbalanced data: I use an suitable estimation procedure, I under-sample the non-events and I utilized the appropriate evaluation metrics. These approaches will all be presented in the sections below

The evaluation will naturally be out-of-sample, and to mimic the forecasting element of a real world scenario the training set will be constituted by all observations from 1990 through 2005. All observations after 2005 through 2009 will serve as test-set. That is; observations after 2005 will not be used to create the models used; this data is saved for evaluation of the frameworks predictive capabilities.

4.1 Predictive Framework

The predictive framework is is made up by a subset of approaches. The subsections below present each step at a time, before moving on the the evaluation effort in the next section.

4.1.1 Extreme Gradient Boosting

As already mentioned, I utilize an Extreme Gradient Boosting (xgboost) algorithm as the basis of my estimations. This method is quite advanced and as such I shall not dig deep into the

[6] See the appendix, subsection 8.3

technicalities of the algorithm. However, some fundamentals serve to explain why this framework is especially well suited for the problem at hand. Three characteristics are especially worth noting; it is a boosting algorithm, it is consists of regression trees and it it self-regularizing (or self-pruning).

Boosting means using a lot of weak classifiers to create a strong classifier. Imaging that we begin with one classifier with which we try to classify geographic locations which will experience conflict fatalities next year. The classifier does rather poorly not least due to the fact that the events are few and and hard to classify. Thus we decide to run a new classifier, but now we give less weight to the observations which was correctly predicted and more weight to the observations which was incorrectly predicted. We reiterate this procedure until some measurement criteria is met. Then all our classifiers are weighted according to their performances and used as a weighted ensemble to predict which geographic locations will experience conflict fatalities the following year (Friedman et al., 2001, 338-339). Since the procedure ensures continuing focus on "hard to classify" observations, it is a particularly fruitfully approach when dealing with imbalanced data.

The Next question is naturally which classifiers to build our boosting framework on. Xgboost utilizes regression tree. These closely resembles decision tress where the features are used to split the observations into categories according to which splits yields the most information according to some predefined criteria. Thus the first split would here be the split which best sorted conflict events from non-events and so on. But unlike decision trees, each regression tree holds a continuous score on each of the leaf, which can be converted into probabilities rather then binary classifications (Chen and Guestrin, 2016, 2). To put it simply, xgboost utilizes N number of regression tress and iterative uses the predictions of one tree to update the weights of the next tree such that 'harder to classify' observations a given more and more attention.

The last question here is then how to decide to number of times we shall allow a given tree to split. To few times and the tree fails to learn any pattern of the data and we underfit. To many splits and the trees start to learn idiosyncratic artifacts of the training data and we overfit. The xgboost algorithm handles this be penalizing complicated tress. Thus when the algorithm evaluates whether or not to make a split the information gain obtained by the potential split is down-weighted according to the increased complexity induced by the new split (Chen and Guestrin, 2016, 4-7). Thus the algorithm search's for any pattern which might help the prediction effort but stops before such pattern becomes overly complicated.

Naturally there is more to the xgboost algorithm then presented here, not least the fact that it has been optimized for sparse data in a number of more technical ways making it even more suited for large and unbalanced data-set then other boosting algorithms(Chen and Guestrin, 2016, 5). But the small introduction above should suffice to outline why the approach works well for the problem at hand, but also why there still are some challenges ahead.

An important property induced by the approach being based on regression trees, is that I can

readily asses how important each included feature is in the prediction effort. Given the exploratory nature of the present endeavour it is paramount that I be able to evaluate which features might present the most fertile research areas in future prediction efforts.

4.1.2 Undersampling

An other instrument in combating the imbalancedness of the data is simply to only utilize a portion of the non-events; undersampling. On could naturally undersample to the extent that the dataset becomes completely balanced. In the present case this leads to a number of problem, not least to many false positives. Naturally one could just raise the threshold for classifying an observation as an event. However, I found that to many observations where given very high probability of conflict leading to a bimodal distribution where I would have expected something more akin to a power distribution.

Some trail and error lead me to use 20 times as many non-events as events. This means that instead of events being 1% of the data, events now constitute around 5% of the data. Not a big difference, but enough to do the trick. The portion of non-events is randomly chosen which of course leads to a lot of information not being used in the estimations. The next subsection illustrate how I amend this.

4.1.3 Predictive Sampling

Now, the undersampling step above leads to a random element in the process; I do not choose which non-events to utilize for the prediction effort. Further more the xgboost framework does itself include a number of stochastic elements; some randomness goes into choosing the exact candidates for splits and thus you will not get the exact same model each time unless you specify a random seed for the model. An approach which utilizes both these elements of randomness to our advantage is to create a distribution of predictions rather the simply a single prediction for each observation in the test-set. As such, I estimate 1000 xgboost models, each including the same events - all events in the training set - and a random sample of non-events 20 times bigger also drawn from the trainingset. Thus for all results derived from the model, estimations, predictions and metrics, I get a distribution of results. That is, I do not get one prediction for each cell; I get 1000 together forming a distribution of predictions for each cell. Likewise I do not get one final metric, e.g. AUC, recall or AP, score but 1000 forming distribution for each metric. for both the predictions and the metrics the mean of these distributions constitutes a more reliable estimate of respectively the "true" risk of conflict and the "true" performance of the framework. Furthermore I also get estimates regarding the uncertainty of the predicted conflict risk and the variance of the frameworks performance.

At the moment only 100 are used for the plots below

4.1.4 Bayesian Correction

BAYESIAN CORRECTION TO COME - you're are still getting to many false positives...

4.2 Evaluation

For the evaluation I use a number of metrics; AUC, AP, recall, precision and accuracy. Presenting this rather broad roster of metrics serves neither elegance nor simplicity. It does however serve transparency, honesty and highlighting of the frameworks strengths and weaknesses.

Naturraly more text will accompany this section. For now here are just some plots of the results to discuss Thursday

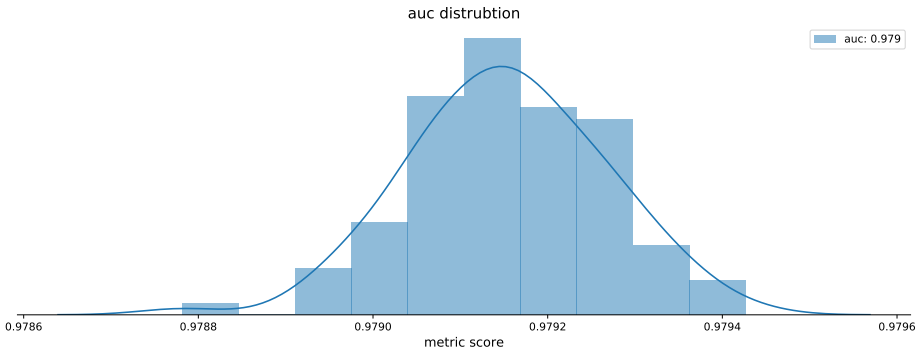

Figure 1: auc from 100 models

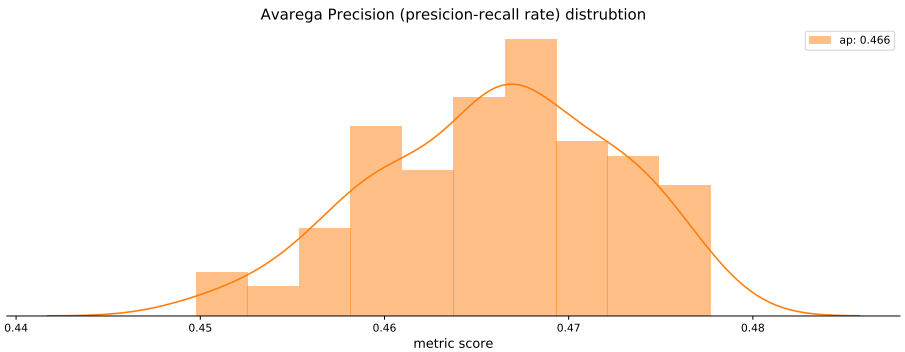

Figure 2: Avarege Precision (presicion-recall rate) from 100 models

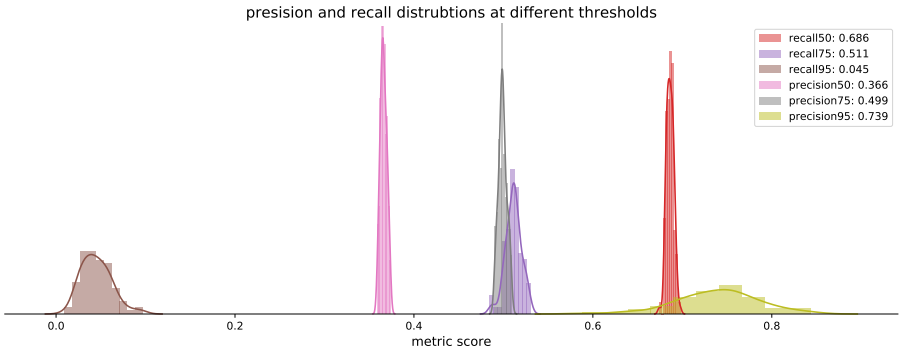

Figure 3: Precision and recall at various thresholds from 100 models

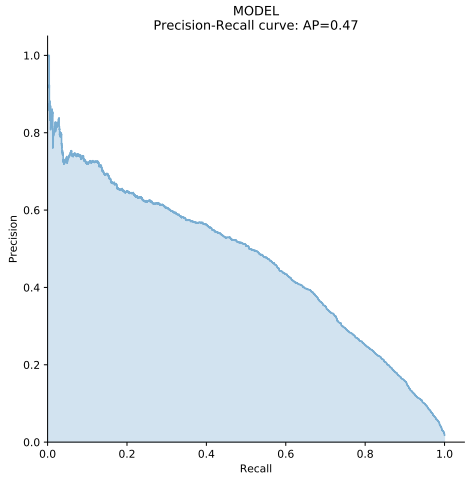

Figure 4: Precision/Recall curve from a random model

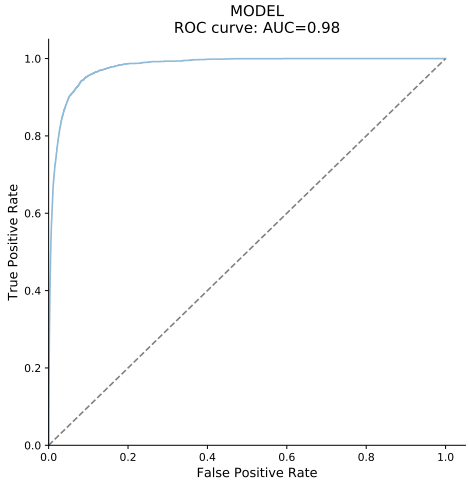

Figure 5: ROC curve from a random model

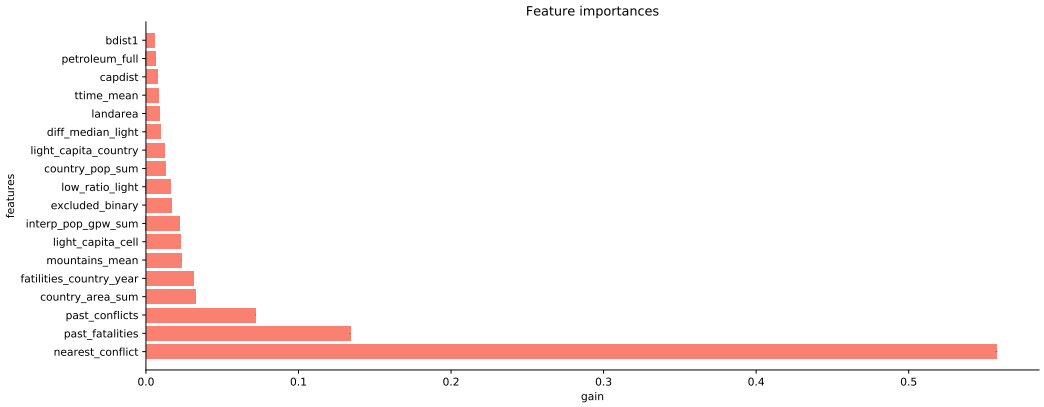

Figure 6: Feature importance across the 100 models

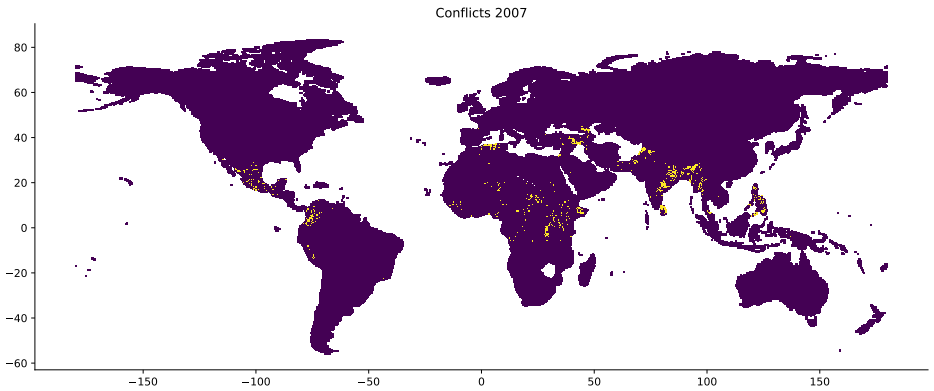

Figure 7: Conflicts (binary) 2007

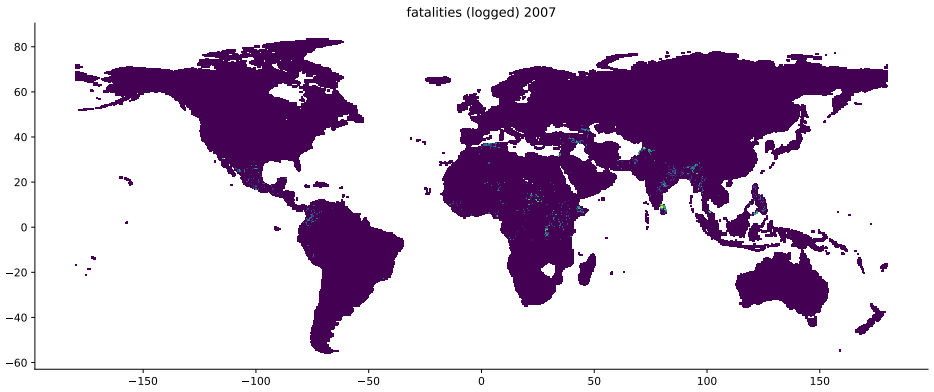

Figure 8: Fatalities (logged) 2007

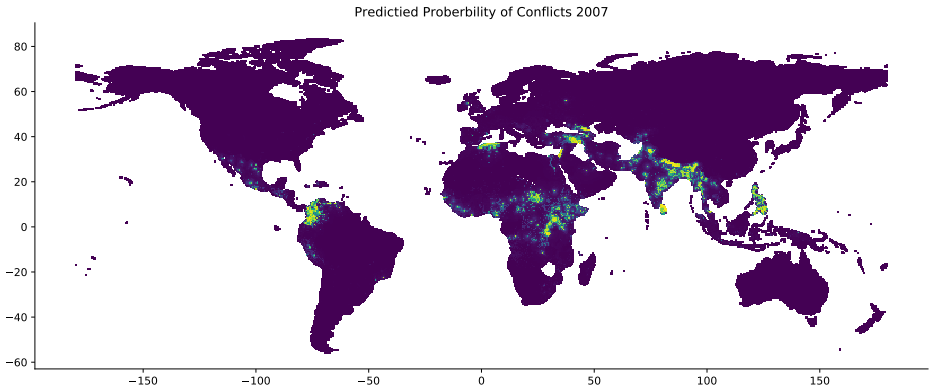

Figure 9: predicted probability of conflict 2007 - as seen we still need the bayesian correction or set the htreshold higer then 50..

5 Future Challenges and Improvements

TO COME

6 Conclusion

TO COME

7 Bibliography

References

Blattman, C. and Miguel, E. (2010). Civil war. *Journal of Economic Literature*, 48(1):3–57.

Blimes, R. J. (2006). The indirect effect of ethnic heterogeneity on the likelihood of civil war onset. *Journal of Conflict Resolution*, 50(4):536–547.

Blyth, S., Groombridge, B., Lysenko, I., Miles, L., and Newton, A. (2002). Mountain watch: environmental change & sustainable development in mountains. *Cambridge, UK: UNEP World Conservation Monitoring Centre*.

Buhaug, H. (2010). Dude, where's my conflict?: Lsg, relative strength, and the location of civil war. *Conflict Management and Peace Science*, 27(2):107–128.

Buhaug, H., Gates, S., and Lujala, P. (2009). Geography, rebel capability, and the duration of civil conflict. *Journal of Conflict Resolution*, 53(4):544–569.

Cederman, L.-E., Buhaug, H., and Rød, J. K. (2009). Ethno-nationalist dyads and civil war: A gis-based analysis. *Journal of Conflict Resolution*, 53(4):496–525.

Cederman, L.-E. and Gleditsch, K. S. (2009). Introduction to special issue on "disaggregating civil war". *Journal of Conflict Resolution*, 53(4):487–495.

Cederman, L.-E., Gleditsch, K. S., and Buhaug, H. (2013). *Inequality, grievances, and civil war*. Cambridge University Press.

Cederman, L.-E., Weidmann, N. B., and Gleditsch, K. S. (2011). Horizontal inequalities and ethnonationalist civil war: A global comparison. *American Political Science Review*, 105(3):478–495.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Chen, X. and Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594.

Collier, P. and Hoeffler, A. (1998). On economic causes of civil war. *Oxford Economic Papers*, 50(4):563–573.

Collier, P. and Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595.

Croicu, M. and Sundberg, R. (2017). Ucdp ged codebook version 18.1. *Department of Peace and Conflict Research, Uppsala University*.

Diamond, J. M. (1998). *Guns, germs and steel: a short history of everybody for the last 13,000 years*. Random House.

Elvidge, C. D., Hsu, F.-C., Baugh, K. E., and Ghosh, T. (2014). National trends in satellite-observed lighting. *Global urban monitoring and assessment through earth observation*, 23:97–118.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., and Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences*, 35(8):1652 – 1660.

Fearon, J. D. (2004). Why do some civil wars last so much longer than others? *Journal of Peace Research*, 41(3):275–301.

Fearon, J. D. (2005). Primary commodity exports and civil war. *Journal of Conflict Resolution*, 49(4):483–507.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York, NY, USA:.

Gartzke, E. (1999). War is in the error term. *International Organization*, 53(3):567–587.

Girardin, L., Hunziker, P., Cederman, L.-E., Bormann, N.-C., and Vogt, M. (2015). Growup-geographical research on war, unified platform. *ETH Zurich. http://growup. ethz. ch*.

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

Gurr, T. R. (1970). *Why men rebel*. Routledge.

Gurr, T. R. (1995). Minorities at risk- a global view of ethnopolitical conflicts. *UNITED STATES INSTITUTE OF PEACE PRESS, ARLINGTON, VA 22210(USA). 1995.*

Hegre, H. and Sambanis, N. (2006). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50(4):508–535.

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

Hobbes, T. (1991[1651]). *Leviathan. Cambridge texts in the history of political thought*. New York: Cambridge University Press. edited by Richard Tuck.

Kalyvas, S. N. (2007). Civil wars. In Boix, C. and Stokes, S. C., editors, *The Oxford handbook of comparative politics*, volume 4, chapter 18, pages 417–434. Oxford Handbooks of Political.

Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American political science review*, 53(1):69–105.

Nordhaus, W. D. (2006). Geography and macroeconomics: New data and new findings. *Proceedings of the National Academy of Sciences*, 103(10):3510–3517.

Ross, M. L. (2004). What do we know about natural resources and civil war? *Journal of Peace Research*, 41(3):337–356.

Skocpol, T. (1979). *States and social revolutions: A comparative analysis of France, Russia and China*. Cambridge University Press.

Sundberg, R. and Melander, E. (2013). Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Tollefsen, A. F., Bahgat, K., Nordkvelle, J., and Buhaug, H. (2016). Prio-grid codebook v.2.0. *Oslo: PRIO*.

Tollefsen, A. F., Strand, H., and Buhaug, H. (2012). Prio-grid: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374.

Tollefsen, Andreas Forø, K. B. J. N. and Buhaug, H. (2015). Prio-grid v.2.0 codebook. *Oslo*.

UCDP (2017). Ucdp georeferenced event dataset (ged) global version 18.1. http://www.ucdp.uu.se/downloads/. Accessed: 2018-11-13.

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P., and Girardin, L. (2015). Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family. *Journal of Conflict Resolution*, 59(7):1327–1342.

8 Appendices

All scripts can be found on : https://github.com/Polichinel/Conflict_Prediction

8.1 Handling of random missing values

8.2 Linear interpolation from 5-years intervals

8.3 The Included Features - Expanded

8.3.1 Wealth and State Capacities

Easily one of the most robust findings in country level studies of civil wars is GDP per capita[7] has a negative effect on the probability of civil war onset (Collier and Hoeffler, 1998; Fearon and Laitin, 2003; Collier and Hoeffler, 2004; Hegre and Sambanis, 2006; Blattman and Miguel, 2010) and also to some extent the conflict duration(Fearon, 2004; Hegre et al., 2009).

A number of mechanisms have been proposed linking GDP to conflict, Two have been especial prolific. The first is championed by Collier and Hoeffler (1998, 2004) sees GDP per capita as a proxy for opportunity-cost. that is what i given citizen have to loss by engaging in conflict. The second story draws on some insight from Skocpol (1979) and also echoes the gospel of modernization theory (Lipset, 1959). Regarding the context at hand it has most prominently been presented by Fearon and Laitin (2003). Here GDP per capita is seen as as proxy for state capacities. Simple put; weak or fragile states have low GDP per capita and these states are more conflict prone(Fearon and Laitin, 2003, 88).

A measure for GCP (Gross cell product) per capita (ppp) is included in the PRIO GRID fomr the Gecon dataset (Nordhaus, 2006) and this measure could be aggregated creating a feature for GDP per capita (ppp) (Tollefsen and Buhaug, 2015). However the data available from the directly from the PRIO GRID only extent to 2010. Fortunately Elvidge et al. (2009), Chen and Nordhaus (2011) have shown that Night light emission can serve as an proxy for economic activies - especially for countries of areas with low-quality statistical systems and few or no recent population or economic censuses. (Chen and Nordhaus, 2011) - an approach explicitly proposed by (Cederman et al., 2013, p. 101) in regards to conflict studies. A ad hoc illustration of the high correlation between the two features can also be found in the appendix, subsection 8.4[8].

[7] Often logged and adjusted for purchasing power parity (ppp)

[8] Futhermore initial test showed very limited difference between the derived features and interactions, thou the features constructed on the basis of Night Light Emission was picked more often by the sequential feature selection

The measure of Nigth Light Emmision in the PRIO GRID is borrowed from Elvidge et al. (2014)it extent all the way up to 2015 and calibrated as to better accommodate time series Tollefsen and Buhaug (2015).

Thus I proceed from here with two features:

- Cell-year specific Night Light Emission (mean)
- Country-year specific Night Light Emission (aggregated mean)

Naturrally one could argue that wealth is a relative concept, which leads us to the next section.

8.3.2 Inequality and depravation

If we - for the time being - leave the Strong State proposition behind and focus on the satisfaction of the individual citizens it is only natural to argues this satisfaction should be considered a function of *what we have* and *what we believe we rightfully should have*. This, indeed, is the crux of Robert Gurr's (1970) Relative Deprivation Theory. Perhaps one of the most seminal[9] takes on inequality and conflict, Gurr (1970) defines relative deprivation:

> *"[...] as actor's perception of discrepancy between their value expectation and their value capabilities. Values expectation are the goods and conditions of life to which people believe they are rightfully entitled. Value capabilities are the goods and conditions they think they are capable of getting and keeping."* (Gurr, 1970, 24).

While intuitively appalling, Gurr's theory was award little credit doing the haydays of comparative corss country conflict studies. Supporting statistical results failed to materialize, and the axplanation eccoed the of (Skocpol, 1979, 11); injustice and misery is simply to widespread to account for the rarity of major conflicts (Collier and Hoeffler, 1998, p. 22, Collier and Hoeffler, 2004, p. 22 Fearon and Laitin, 2003, p. 44(FLERE?)). (så lige styr på de 22 side tal der?)

However, Cederman and Gleditsch (2009); Cederman et al. (2013) have noted that the aggregated country level features conventional used as indicators for inequality might lead to misspecifications; that is, they do not properly measure the theoretical concept of relative deprivation or the correct mechanisms through which inequality affects conflict-propensity (Cederman et al., 2013, XX). Acknowledging this critique I utilized the operationalization put forth in (Cederman et al., 2013, p. 104-105)[10]:

process in **??**.

[9] At least after Marx

[10] These scholars alos argues the higher conflict propensity might be fund in the other end of the inequality spectrum.

$$y_g = \text{country year mean} \quad , \quad y_c = \text{cell year value}$$

$$\text{low\_ratio} = y_c/y_g \quad \text{if} \quad y_c < y_g, \quad 1 \quad \text{otherwise}$$

Thus cells which are relatively well of compared to the mean of country takes the value 1. Cells worse of the the country mean takes a value above 1, with the magnitude of this value indicating *how* severe the deprivatione is. Cederman et al. (2013) Uses GCP per capita (ppp) but as mentioned also suggests using night light emission, Thus I here produce ratio feature solely on Night Light emission (cell-year mean).

While I do appreciate this operationalization I also construct my own relative deprivation feature, which differs in a number of small but relevant ways. First we know that wealth distributions in general are highly skewed, and this is no different when we use Nigt Light Emission as indicator (se Appendix XXXX). Thus, given Gurr's original conceptualization I find it more realistic that individuals should compare themselves to the median - not the mean. That is citizens in one cell compare their living standard to the most common living standard in the country over all. In the same vein, instead of a fraction, I simply calculate the difference between the cell-year values and the country-year median. Lastly I let the value start at 0. This I do to insure that interactions create later only "activates" if the cell is actually depraved. In the case of Cederman et al. (2013) when a cell is not depraved, the value of a interaction takes the naked value of the other interacted feature, which is arguably somewhat imprudent. Mathematical the features is formulated as such:

$$y_g = \text{country year median} \quad , \quad y_c = \text{cell year value}$$

$$\text{low\_diff\_median} = y_g - y_c \quad \text{if} \quad y_c < y_g, \quad 0 \quad \text{otherwise}$$

Thus the feature takes the value 0 if the cell is on level or above with the rest of the country. If the cell is deprived the value correspond to the difference between the country-year median and the cell value. Naturally - and as we I shall return to later - this measure and the measure by Cederman et al. (2011) are highly correlated, nut that does not change the fact that this last feature appears somewhat closer to the notion of relative deprivation, and more impotently it handles interactions somewhat more appropriate or at least conventional. As before the feature uses on Night Light emission (cell-year mean) as indicator for wealth.

That is; one could imagine a mechanism akin to relative deprivation in one end, while simultaneous at the other end, one might see well-off people wanting to secede from or take over a country if they fear to much redistribution. However, they find little statistical backing for this proposition, neither did my initial tests. Thus I only use the measures corresponding to actual Relative Deprivation

8.3.3 Ethnicity and Exclusion

Denotes the number of excluded ethnic groups (e.i. discriminated or powerless) in a given cell at a given year. The measures are originally from the GeoEPR/EPR data by Vogt et al. (2015). To better suit the theoretical argumentation laid forth in Cederman et al. (2013)(side) (Tollefsen and Buhaug, 2015). I also create a dummy (excluded_binary) which simply denote *if* there are any excluded ethnic groups.

As with inequality the link between ethnic diverse societies and conflict propensity have been ridden with disagreement and controversies. In the quantitative literature results have remained somewhat inconsistent (Blattman and Miguel, 2010, 23-24). A number of studies have fund different - and sometime quite convoluted - relationships between ethnicity or discrimination and conflict (Collier and Hoeffler, 1998; Fearon, 2004; Blimes, 2006; Hegre and Sambanis, 2006; Goldstone et al., 2010), while other studies have fund little or no trace of the connection (Fearon and Laitin, 2003; Collier and Hoeffler, 2004)(more? CH 2004 right?).

As with the problem with inequality the lack of discernible results have often been attributed to poor feature specifications, a framework not capturing the proposed theoretical mechanism and not least country aggregated data (Blimes, 2006; Blattman and Miguel, 2010; Cederman et al., 2013). Mirroring they effort concerning inequality Cederman et al. (2013) have recently made use of new desegregated data (Girardin et al., 2015) to closer model the theoretical mechanism proposed. Without going to much in-depth here the feature Cederman et al. (2013) utilizes aims to capture the effect of *horizontal inequalities* as defined by (Cederman et al., 2013, 31-35). That is the systematic discrimination or political exclusion of and coherent ethnic group. Thou not framed in the theoretical context of horizontal inequalities Goldstone et al. (2010) finds results supporting the argument using the Minority at Risk data from Gurr (1995).

Conveniently, a measure from Girardin et al. (2015) of how many excluded ethnic groups reside in each PRIO GRID cell is readily aviable in the PRIO GRID. To mimic the theoretical proposition lead foth in Cederman et al. (2013) I binarize this variable to simply indicate whether or not a given cell is inhabited by a politicly excluded or discriminate ethnic group[11].

Not surprisingly Cederman et al. (2013) futher finds that the properbilty of conflict gets even larger if political exclusion is followed by group deprivation, wich leads to the next feature:

[11] Also initial test did not show much prospect of incorporating the full count

8.3.4 Deprivation and Exclusion - an Interaction

This interaction does not need much justification; the general idea is simply that groups of humans (here specificly ethnic groups) which are both cut of from political influence and find themselves to be generally worse of then other - more influential - groups, tends to accumulate a lot of grievances. Given that few to no political solutions are available for these groups, they are reletively likely to experience conflict(Cederman et al., 2013, 103-111).

Given that I have utilized to measures of relative deprivation I also construct two interaction to be evaluated by the systematic feature selection to come:

$$\text{excluded\_b\_low\_ratio\_nlights} = \text{excluded\_binary} \times \text{low\_ratio\_nlights}$$

$$\text{excluded\_b\_low\_median\_diff\_nlights} = \text{excluded\_binary} \times \text{low\_median\_diff\_nlights}$$

One very relevant difference here is that the feature from Cederman et al. (2013) takes the plain value of 'excluded' $\in 0, 1$ if the cell is not deprived while the interaction only takes the value of 0 if 'excluded' takes the value zero. As mentioned I find this somewhat messy and statistically it makes it harder to determined the effect of 'excluded' in it self and the interaction. Naturally this is an issue I return to in **??**.

8.3.5 Population size and density

Returning to a variable with a rather flawless record regarding conflict onset we have "country population size" (Collier and Hoeffler, 1998; Fearon and Laitin, 2003; Collier and Hoeffler, 2004; Hegre and Sambanis, 2006)[12]. Furthermore (Fearon, 2004, 287) also finds that country population is correlated with longer civil wars.

Lastly when it comes to disaggregated grid data, the "grid population size" also seems a rather robust predictor (Buhaug, 2010; Cederman et al., 2013)[13]. Thus, whether the features influence conflict duration is more disputed (Collier and Hoeffler, 1998; Fearon, 2004). The most common implementation is to use a log transformed version of the country or grid population count[14] which is also implemented in the project at hand.

[12] though see Goldstone et al. (2010)

[13] Though see Hegre et al. (2009)

[14] Though Collier and Hoeffler (1998) use the plain count in 10.000's

One very simple explanation might be the various definitions of conflict and civil war used throughout the litterature. These definitions almost always refer to some minimum fatalities count [eks and refs]. Naturally high population counts makes these threshold relativly less restrictive.

There are however also more theoretical propositions for the relationship. One being that conflict mediation becomes inherently more difficult as social systems and societies grow (Diamond, 1998, p- 271-272).

Initial I include features for cell-year population, aggregated country-year population and corresponding population densities:

$$\text{grid\_pop\_dens} = \frac{\text{grid\_pop}}{\text{grid\_area}}$$

$$\text{country\_pop\_dens} = \frac{\text{country\_pop}}{\text{country\_area}}$$

Which leads to the last included interaction concerning excluded ethinic groups.

8.3.6 Density of the Excluded - Interaction

(excluded_pop, excluded_b_pop)

One more drawn directly from Cederman et al. (2013) disaggregated study conserns the interaction between population size and group exclusion. the Auothors construct an interaction between the size of the cell population and their feature for excluded ethnicities (Cederman et al., 2013, 73-78). Assuming that the size of the excluded population is at least positively correlated with the total cell population this feature captures the density of the excluded population.

$$\text{excluded\_b\_pop} = \text{excluded\_binary} \times \text{cell\_pop}$$

The theoretical argument is simply that it is not larger population but larger excluded populations that drives the relationship between conflict and population size (Cederman et al., 2013, 69-74).

8.3.7 Geography and Accessibility

Accessibility As noted earlier, a strong state have often been presented as the prime inhibitor of internal conflicts. Naturally though, the strength of state must be considered relative to the territory over which it claims sovereignty - both in regards to size and permeability. Fearon and Laitin (2003) pointed to rough terrain and mountains as natural obstacles hindering effective projection of state power. Following this example Hegre and Sambanis (2006) concludes that a feature for rough terrain is found to robustly positively correlated with civil war across a large number of model specifications.(Hegre and Sambanis, 2006, 526-529)[15]. The Prio Grid includes a readily available feature measuring the proportion of mountainous terrain within the cell based on Blyth et al. (2002) wich I utilize.

Another natural hindering for projecting state power is sheer distance (Fearon, 2004; Buhaug et al., 2009; Cederman et al., 2009; Buhaug, 2010). The argument is straight forward:

> *"The projection of power across distance comes at a cost. [...] In particular, large hinterlands and isolated peripheries are favorable to insurgency. In sum, this suggests that large countries are relatively more exposed to intrastate conflict"*(Buhaug, 2010, 113-114)

A number of interesting features could be derived from Buhuag's assertion above (and the paper in general). What I have included is the distance to the nations capital[16], the travel time to the nearest major city, and the total size of the countryTollefsen and Buhaug (2015).

8.3.8 Trans-boarder Influences

A number of different mechanism have been proposed and explored (Blattman and Miguel, 2010, 29-30), the one explored here is rather simple and follows Hegre and Sambanis (2006) [TJEK IGEN!]. blabla

Der er vel også lidt en tråd til Conflict dispersion....

de to mål og hvorfor du inkoorporere begge.

bdist1

[15]Though see Goldstone et al. (2010)

[16]The measure utilized by Buhaug (2010)

bdist3 This is a very rough measure and is arguably a bit far removed from any specific theoretical concept, but for this preliminary project the measures will have to do.

8.3.9 Urban contra Rural Theater

Ja, hvad er din teoritiske begrundelse her? går du tilbage til litteraturen om democratiske sammenbrud? Noget med at der er mere undertrykkelse i land områder hvilket leder? (Boix, Robinson, Acemogul...) Til grievences? Eller omvendt er det lettere at mobilisere i byerne?

Det trækker også lidt på det accesability du har snakket om tidligere... Son of the soil..

interp_urban_ih her henter du også viden fra democratic breakdown lit.

interp_agri_ih her henter du også viden fra democratic breakdown lit. og Skocpol

8.3.10 Prime Commodities and the Recourse course

først: prime comoditeis in general: Collier and Hoeffler (1998, 2004)

find no impact of prime comm. (Fearon and Laitin, 2003, 76) does find inpact of oil (Fearon and Laitin, 2003, 84-86)

Kikker videre på prime comm. Fearon (2005), (or Natural reasources) Ross (2004)

Og sons of the soils Fearon (2004), men så skulle du lave nogle interactioner med excluded..

Buhaug (2010) Oil disaggregated

As such, the soil feature here included is a binary indicator of whether oil is know to be available for extraction in a given cell at a given year.

Hegre et al. (2009) : also a bit on prime Comm

8.3.11 Inertia, dispersion, traps and time trends

Conflict, inertia, trap, conflict dispersion, time trends ect

Collier and Hoeffler (2004) : linear DECAY term since last conflict.

Goldstone et al. (2010) : conflict ridden neighbourhood.

Hegre and Sambanis (2006) : which we model as a DECAY function of time at peace

"Among others, we include a decay function proxregc of the Polity durable variable, which measures the number of years since an institutional change that leads to a minimum of three points' change on the Polity index."

"In a review of the quantitative literature on civil war, Sambanis (2002) identified the following three core variables that are almost always included in models of civil war onset: the natural log of population (Inpop), the length of peacetime until the outbreak of a war (pt8, which we model as a decay function of time at peace), and the natural log of per capita gross domestic product (GDP) in constant dollars (Ingdp)."

Cederman et al. (2013) Number of previous conflicts

8.3.12 Notable Absentees

Notable missing: infant mort, political system and intra elite conflict... (Goldstone et al., 2010) Past conflicts (contry and cell level..) (conflict trapp, Coiller ect) ... Trans boarder thing.. Downgraded... (Cederman et al., 2013)

(Cederman et al., 2013, 119-142): Du har border, men måske er mekanismen ikke rigtigt specificeret. det er trans-boarder-ethnic-kin

Goldstone et al. (2010) : Infant mortality deviant from global mean, factionalisation (political system), state led discrimination (nej, den e rjo lidt med igennem excluded), conflict ridden neighbourhood(er med på celle basis gennem nearest ish). p:

8.4 GCP and Night Light Emission