Where Men Rebel

A Modern Approach to Conflict Forecasting

Institute of Political Science, University of Copenhagen
By Simon Polichinel von der Maase
Under Frederik George Hjort

Format: Article

July 2019

Character count: 117.685/120.000 (50 normal pages)

Contents

Abstract

This thesis represents a substantial step towards the construction of a reliable computational early-warning system for predicting future intra-state conflicts at a sub-national level. Specifically I examine to what extent it is possible to reliably predict the time and place of future conflicts, using the temporal and geo-spatial patterns of past conflicts in tandem with the machine learning technique of Gaussian processes.

Previous research has shown that past conflict patterns hold potential when it comes to predicting future conflicts. However, previous efforts have largely failed to create features mirroring the theoretical insights we have regarding conflict patterns. To best take advantage of the predictive potential inherent in past conflict patterns we need a tool that can capture conflict patterns in a manner which complies with theoretical expectations. In this thesis I illustrate that the machine learning technique of Gaussian processes is such a tool.

I use highly disaggregated spatial data from the years 1989 through 2012 to train my framework while data pertaining to the years 2013 through 2017 are held back as a test set to emulate the unknown feature which I can test my predictions against.

Using Gaussian processes I capture the spatial and temporal patterns of past conflicts and extrapolate these into "future" test years. To asses the extent to which my approach has succeeded in generating reliable forecasts, I use these patterns as features in a predictive framework based on xgboost. The predictions generated by this framework are then evaluated through out-of-sample prediction.

Using my approach to forecast respectively two and three years into the "future" I achieved AP scores of 0.51 (2014) and 0.52 (2015) against a baseline of roughly 0.05. For both these years I achieved an AUC score of 0.91 against a baseline of 0.50.

While direct comparison with previous efforts is difficult, the framework presented here does at least as well, if not better, than the state of the art. Notably this is done using only past patterns, and in a setting which emulates a real world forecasting scenario.

One disadvantage of the approach is that it is unable to capture truly novel conflict onsets far removed in time and space from any previous conflict. Amending this will require the inclusion of other data sources.

1 Introduction

> *[...] 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)

Civil wars, and internal conflicts in general, have plagued mankind throughout the ages. And still, this affliction has not yet been throttled. Since 1960, over half of all nations have experienced some manner of internal conflict leading to fatalities (Blattman and Miguel, 2010, 3-4). Indeed, since the end of the Second World War, intra-state conflicts have been far more common than inter-state wars and over five times as many people have died in intra-state conflicts compared to inter-state wars (Collier and Hoeffler, 2004, 563). Thus the 367 years old quote above is today as relevant as ever.

This thesis represents a substantial step towards the construction of a reliable computational framework for predicting future conflicts at a sub-national level; a so called *early-warning system*. In the short run such a system will provide international actors valuable information and time to act; mitigating the fallout of conflicts (Ward et al., 2010; Perry, 2013). In the long run such a system will generate theoretical insights into the mechanics of conflicts, potentially enabling actors to address the underlying courses (Schrodt, 2014; Chadefaux, 2017). Given these potentials, efforts to create an early-warning system have been called "conflict researchers' ultimate frontier" (Cederman and Weidmann, 2017, 474). As such, the contribution of this thesis is highly topical and can provide practical and ongoing guidance in future conflict prevention, intervention, and peace keeping efforts.

A complete early-warning-system would be constituted by a legion of sophisticated components. In this thesis I will focus on one of these components: A component capable of using patterns of past conflict to predict the patterns of future conflict. The reason why I choose this focus is, as I will demonstrate throughout the thesis, that this approach has a number of very appealing properties: it has a clear theoretical foundation, it holds high prediction power, and it takes advantage of very current data. As such, using past patterns to predict future patterns has the potential of being one of the most important and central components of any complete early-warning system.

Indeed, in this paper I show that using the machine learning technique of Gaussian process to capture and extrapolate past conflict patterns into future years carries great predictive potential. Specifically, while direct comparison is difficult, I am able to generate predictions which are at least as good, if not better, than any previous efforts of conflict prediction. Importantly I do this using solely the patterns of past conflict, a highly disaggregated unit of analysis and a setting which rigorously emulates a real world forecasting scenario. The later point is crucial since many researchers have previously ignored different aspects of real world forecasting – such as the release date of the data they use relative to the actual years they aim to forecast into.

I conclude that the the construction of an early warning system is possible and that the inclusion of past conflict patterns – correctly handled – should play a key part of such a system. However, I also conclude that we must incorporate data for other sources to create a system reliable enough to use in real world applications.

The proceeding subsections will serve as a more thorough introduction to the context, subject and objective of my thesis.

1.1 Conflict prediction so far

Three specific subjects from the broader field of conflict studies are of prime relevance for this paper: The difference between estimation and prediction, the use of sub-country disaggregated data as opposed the country level data, and the use of event data instead of structural data as the basis of conflict prediction. Each of these subjects will be briefly introduced in turn below.

Firstly, the study of internal conflicts has a long history. The sub-field of computational conflict prediction[1], however, is a rather novel offspring. Researchers of conflict prediction use modern computational tools in efforts to predict the time and place of future conflicts (Cederman and Weidmann, 2017; Chadefaux, 2017). Some of the earliest examples in academia of such a focus being Goldstone et al. (2010), Hegre et al. (2013) and Perry (2013).

There are two main reasons why the field of conflict prediction – despite its obvious usefulness – has not received more attention. First of all; the required technology is relatively new. Indeed, some of the more impressive results are the products of very recent developments in quantitative methods, computation power and data availability (O'Loughlin et al., 2010; Perry, 2013). Secondly, traditional conflict research has focused on causes, correlates and prerequisites of conflict rather than generating actual predictions (Chadefaux, 2017, 8). As such, modern predictive efforts are still regarded with high skepticism and papers devoted to prediction have been accused of lacking adequate theoretical grounding (Chadefaux, 2017, 8-9).

In reality, however, both prediction and explanation are important components of scientific inquiry, and both are needed to generate and test theories (Schrodt, 2014; Chadefaux, 2017). Indeed, in the natural sciences, accurate predictions are central for the validation of theory (Schrodt, 2014, 289). As such, the restraint against predictions in the field of conflict research has been increasingly criticized (King and Zeng, 2001b; Ward et al., 2010; Goldstone et al., 2010; Schrodt, 2014; Chadefaux, 2017). Furthermore, creating models with actual prediction-power will also provide heuristic tools and powerful policy-guides for real-world application (Ward et al., 2010, 372). As such, a focus on prediction have been increasingly encourages (Ward et al., 2010; Schrodt, 2014).

Secondly, countries have traditionally been the main focus, but since the phenomenon of interest is a sub-country phenomenon, the unit of analysis should also be disaggregated at sub-national

[1] from here just denoted conflict prediction

level (Cederman and Gleditsch, 2009, 490). Employing a disaggregated unit of analysis allows me to create a more detailed analysis of conflict patterns. It also allows me to better analyze incidents where administrative boarders provide little to no ward against conflict diffusion (O'Loughlin et al., 2010, 445-446). In turn, the derived insights allow me to generate forecastings pertaining to specific sub-national areas rather than entire countries.

Thirdly, the last subject constitutes most directly the starting point and contribution of this thesis: the use of event data over structural data to predict future conflicts. Traditional efforts have tended to use structural features as foundation. This has been true for both estimations and predictions (Chadefaux, 2017, 10). However, in a recent contribution of my own I show that patterns of conflict events in themselves appear to hold much more prediction power than the traditional structural variables (von der Maase, 2019). These findings align perfectly well with a growing strain of the literature concerning conflict diffusion; little doubt remains that conflicts cluster in time and space (Crost et al., 2015, 15) and recent efforts have shown that the clustering and diffusion of conflict is often a direct product of contagion (Buhaug and Gleditsch, 2008; Schutte and Weidmann, 2011; Crost et al., 2015; Bara, 2018). As such, when it comes to sheer prediction power, event data appears to have the advantage over structural data.

Importantly, event data is also, at present, much more current and more often maintained than data pertaining to the structural features. In practice this means that the data I use for actual forecasting is much closer in time to the forecastings I wish to generate, thus creating more reliable predictions and mitigating uncertainty.

Given these insights and conditions – and the scope of the thesis at hand – I will only utilize data relating directly to past conflict events in my prediction effort. That is, estimating the probability of future conflicts given past patterns of conflict. The next section will introduce the research question I answer, the approach I utilize and the framework I construct in this thesis.

1.2 This framework

The prerequisite for using past patterns of conflicts to predict future patterns of conflict is the ability to identify and extract salient patterns and further extrapolate these patterns into future years. In this thesis, I show how this can be done using the machine learning technique of *Gaussian processes*. I will also demonstrate how the technical properties of Gaussian processes align remarkable well with the theoretical foundation of conflict diffusion in general. The presentation, qualification, and exemplification of how Gaussian processes can turn past conflict patterns into future conflict predictions is the prime objective of this thesis. The research question motivating this thesis can be summed up accordingly:

> ***Research Question:*** *To what extent is it possible to reliably forecast and predict the time and place of future conflicts, using the temporal and geo-spatial patterns of past conflicts in tandem with Gaussian processes.*

To answer this question, I use event data pertaining to sub-country geographic grid cells. The data I use covers the years form 1989 through 2017. I split the data into a *training set* (1989-2012) and a *test set* (2013-2017). The test set is meant to emulate the unknown future, and as such is kept isolated until the final evaluation effort. I then construct a framework composed of two layers of estimation. First, Using the training data I estimate the spatial and temporal conflict patterns via Gaussian processes. Also using Gaussian processes, I extrapolate these patterns into future years corresponding to those of the test data. The product is eight features pertaining to the patterns of conflict. Secondly, from these eight features I use the values corresponding to the training years to train a predictive framework based on *Extreme Gradient Boosting*. Having trained this framework I introduce the extrapolated values corresponding to the test years. Using these extrapolations I estimate the probability of intra-state conflict in a specific geographic grid cell given a specific (future) year from the testset. To asses how well the extracted patterns are able to actually predict the time and space of future conflict these predictions are evaluated through *out-of-sample predictions* against the actual observations contained in the testset.

To be more specific; the data I use is obtained from the Uppsala Conflict Data Program (UCDP, 2017). This data denotes, among other things, conflict fatalities at specific coordinates and dates (UCDP, 2017). I use the logarithm of this measure to create a measure of *conflict magnitude* (cm). The temporal dimension spans from 1989 to 2017 and I aggregate it at a yearly level. The spatial dimension, I aggregate using a global grid constituted by 0.5×0.5 decimal degree cells; that is squares roughly measuring $50km \times 50km$ at the equator (Tollefsen et al., 2012, 367). Thus, the unit of analysis is one specific grid cell at one specific year.

I will refer to the combined years of a given cell as the *time-line* of that cell. Each time-line tells the story of conflict in the corresponding cell throughout the included years. Each time-line can also be seen as a continuous function $y = f(x) + \epsilon$ where ϵ is a noise term, y is conflict magnitude and x is years. If we can estimate this function, we can also extrapolate it into the future by introducing new years; x_{new}. This is exactly what we can use Gaussian processes for.

Now, this will give an estimate of the future magnitude of conflict in each individual cell. This is in itself a prediction. But I can do better. Using only the magnitude of the individual cell, we would fail to account for the spatial diffusion of conflicts between cells. We need a measure accounting for *spatial exposure* as well. Encouragingly, I can use Gaussian processes for this purpose as well.

For each individual year in my sample, I estimate a 2D function of the spatial pattern of conflict over all included grid cells. This gives me an estimate of the level of *static conflict exposure* (sce) each cell experiences in any given year. To bind this information together across all years, and allow for extrapolation into the future, I once more view the individual time lines as functions $y = f(x) + \epsilon$. x is still years, but now y is conflict exposure. This captures the spatial patterns of *dynamic conflicts exposure* (dce) across time[2].

[2]Effectively this constitutes a pseudo 3D space. It could also have been estimated directly in 3D if I had access to

I thus have two base functions. One of conflict magnitude and one of conflict exposure. Intriguingly, since I am effectively working with continuous functions, I can readily extract other relevant information. Specifically how much the conflict magnitude or exposure is increasing or decreasing (slope), whether the conflict magnitude or exposure of a given cell is accelerating or decelerating (acceleration), and the total mass of conflicts in the time-line of each cell (mass). Using these options on both conflict magnitude and conflict exposure, I construct six additional features. All in all, I produce eight features representing the spatial and temporal patterns of conflict. All eight feature are extrapolated into future years.

The extrapolation of these measures constitute forecastings in themselves. However, to harvest the full potential of these measures, I will use them as features in a combined predictive framework. A framework capable of predicting probability of conflict in a given cell at some future point in time. For this I will use the forecasting-framework presented in von der Maase (2019). This framework is based on an Extreme Gradient Boosting algorithm (Chen and Guestrin, 2016). The framework is further primed for rare events by employing a combination of case-cohort under-sampling (King and Zeng, 2001b, 142) and informed under-sampling (He and Garcia, 2008, 1267). I specifically designed this framework with internal conflicts in mind, thus making it an obvious choice for the evaluation effort.

To evaluate how good these predictions are I employ out-of-sample prediction. If I were to forecast into actual future years, I would not be able to conclude how reliable my approach is – at least not before a number of years have passed. To avoid this impractical latency I divide my data into a *train set* and a *test set*. My train set will encompass all years from 1998 to 2012 while the test set will encompass all years from 2013 to 2017. Importantly, neither my features nor my model will be based on data from the test set, which will be kept isolated until the evaluation effort. As such, the test set effectively represents "the future" (Goldstone et al., 2010, 199-200). Using this evaluation-scheme insures that I obtain an accurate assessment regarding my approach potential for actual forecasting.

Using my approach I am able to achieve rather reliable conflict prediction. Given the highly disaggregated unit of analysis I employ, direct comparison with previous efforts is less forward, but it is safe to say that my framework produces prediction which are at least as good if not even better then the state of the art in academia.

As such, I find there is a great potential for using Guassian processes to capture and extrapolate conflict patterns for conflict forecasting. That being said the approach does have its limits. Using only this component I will never be able to predict truly novel conflicts far removed in time and space from any previous conflict. Amending this issue will require the introduction of other components based on other data sources e.g. text or images. As such, I recommend that these subjects should be the focus of future research efforts.

Having outlined the structure, approach and conclusions of my thesis, the next section will

vastly more computational power.

serve to elaborate on how my thesis fits into the large context of conflict studies. This will serve to further qualify why I have chosen the specific approach I present in this thesis.

2 Recent and Relevant Developments

The field of conflict studies has flourished during the last two decades. Yet, during this period it has also been challenged and critiqued from numerous angles. This thesis is naturally a product of these past debates and developments. Since my starting point is the latest research in the field, I here present the two most recent and relevant developments. In the first subsection I elaborate on the development from estimation towards forecasting. In the second subsection I elaborate on the development from cross-country studies towards more disaggregated geographical units. A basic understanding of these two developments is necessary to appreciate my main challenge presented hereafter: Using past patterns as future predictors.

2.1 From estimation to forecasting

Traditionally, conflict studies have focused on explaining conflict and understanding which features facilitate conflicts. The features under investigation have mostly been structural such as poverty, natural resources or a specific political regime. Researchers used statistical tools such as linear and logistic regressions to estimate the effects of such features on the probability of conflict (Chadefaux, 2017, 8), and further whether this estimated effect could be considered *statistically significant* (Ward et al., 2010, 363-364). If a given feature was found significantly related to conflict – and the study could be considered valid and unbiased – researchers would use theory to argue why the correlation could be considered causal. This is *causal inference* (Chadefaux, 2017, 8). Knowing which features "causes" conflict could then be used for policy recommendation.

The focus on causal inference was (and still is) prevalent, and few conflict experts have ever attempted to undertake actual conflict prediction and forecasting (Cederman and Weidmann, 2017, 474). This reluctance to focus on prediction is primarily due to two assertions. Firstly, efforts devoted to prediction have been accused of lacking strong theory and as a consequence been rejected by traditional political science journals (Chadefaux, 2017, 8-9). Secondly, predictions have been considered fruitless due to "the perceived impossibility of forecasting political events" (Chadefaux, 2017, 8).

Such assertions, however, are both misleading and unproductive. First of all, prediction is not "unscientific". On the contrary, both prediction and explanation are key components of scientific inquiry, and both are needed to generate and test theories (Chadefaux, 2017, 8). In the natural sciences, accurate predictions are seen as the "epitome of validation of a theory" (Schrodt, 2014, 289). Secondly, while we do not know the full potential of conflict prediction (Cederman and Weidmann, 2017; Chadefaux, 2017), we know it is not impossible. The reason is

that we already have a few tentative, but encouraging, results available (Goldstone et al., 2010; Perry, 2013; Mueller and Rauh, 2016; von der Maase, 2019). Indeed, developments in statistical methods, computational power and data availability make such endeavours increasingly feasible (O'Loughlin et al., 2010; Perry, 2013).

Given these insights and developments, the sentiment is shifting and criticism is increasingly aimed at the explanatory approach. This approach has often been justified by assuming that high explanatory power equals high predictive power (Chadefaux, 2017, 8), however, This is not the case. Using statistical significance as the basis of policy recommendation is now considered highly imprudent, since many prominent studies using this approach have shown very poor predictive capabilities (Ward et al., 2010; Schrodt, 2014; Chadefaux, 2017). Indeed, the approach of significance testing in conflict studies has been consistently criticized for almost two decades (King and Zeng, 2001a; Ward et al., 2010; Goldstone et al., 2010; Schrodt, 2014; Chadefaux, 2017). Given this critique and my predictive goal, the focus in this thesis will be solely on prediction-power.

When employing a predictive approach I do not rely on significance levels for evaluation. Even more, since the goal is purely heuristic, concerns about bias and endogeneity are shelved. Instead, the primary concern is increasing prediction-power by mitigating both *underfitting* and *overfitting*. Underfitting, meaning creating a model which has failed to learn any salient patterns in the data pertaining to the phenomenon under investigation. Overfitting, meaning the accidental identification of patterns present in the data, yet not present in the real world (McElreath, 2018, 165-168). One failure of traditional estimation efforts is the tendency for the models to either over- or underfit (Ward et al., 2010, 364). Thus, avoiding such ills, is of prime concern in this thesis.

Testing whether a given model can generate reliable predictions can help identify overfitting, underfitting and other ills. There are many tools available for testing the prediction power of a model. One such tool recommended for conflict studies is out-of-sample prediction (King and Zeng, 2001a; Ward et al., 2010; Perry, 2013; Schrodt, 2014). This approach is simple and entails the construction of the predictive model(s) on one set of data, leaving another set of data for testing; a training set and a test set. In simplified terms, if the model fits the test set as well as the training set, overfitting has been avoided. If the predictions are also accurate and reliable, underfitting has been avoided. While out-of-sample predictions might not cure the ills, it generates honest evaluations and reveal overfitting and other compromising malaise. As such, out-of-sample prediction is the evaluation approach I use in this thesis.

Adopting similar evaluation frameworks, some scholars have started using prediction to evaluate the salience of given features along with traditional parameter estimation (Goldstone et al., 2010). Other scholars have abandoned parameter estimation and now combines modern machine learning techniques with the traditional roster of structural features to create predictive tools for policy recommendation (Perry, 2013). Lastly, some scholars abandoned both parameter estimation and the traditional roster of features all together, such as Mueller and Rauh (2016) who rely solely on

text data from news outlets to predict future conflicts.

In this thesis I will use modern machine learning tools in cohort with past conflict patterns to predict future conflicts. A "causal" discussion regarding whether conflict can be considered truly contagious is both interesting and worthwhile, but it is not the focus of this thesis. My goal is solely to examine the potential of using past conflict patterns for reliable forecasting. As such, high prediction power is my prime concern going forward into this project. This does not mean that theory does not matter. What it does mean is that the theory I employ serves to improve prediction power by informing how I construct my framework ex ante – and not to argue about causality ex post.

This concludes my section on the development from estimation to prediction. Another central development is the shift from cross-country studies to more disaggregated studies. That is, from studies taking countries as their unit of analysis to studies using much smaller geographic entities as unit of analysis.

2.2 From cross-country to disaggregated

Through out the quantitative literature on internal conflict and civil war, the unit of analysis has traditionally been *country-years*. That is; a specific country in a specific year. This has been true both when the goal was prediction (Goldstone et al., 2010; Mueller and Rauh, 2016) and explanation (Collier and Hoeffler, 1998; Fearon and Laitin, 2003; Collier and Hoeffler, 2004; Hegre and Sambanis, 2006). In this thesis, rather than countries, the unit of analysis is 0.5×0.5 decimal degree grid-cells. This subsection presents the theoretical background and motivation for this disaggregated approach.

In reality, civil war rarely encompasses entire countries, but are often confined to specific regions of a country (Cederman and Gleditsch, 2009, 487). As such, we miss important nuances and patterns when treating internal conflict as a phenomenon which is necessarily country-wide. Some regional conflicts will be presented as country-wide civil wars, while other regional conflicts will be treated as non-conflicts. Furthermore, some theoretically relevant features cannot readily be modelled when using observations aggregated at country-level. As an example, Cederman and Gleditsch (2009) argues that the many non-findings regarding the role of ethnic fractionalization in the quantitative literature[3] can be attributed to over-aggregation (Cederman and Gleditsch, 2009, 493). A point which is rather convincingly elaborated in Cederman et al. (2013) and supported by the results in Goldstone et al. (2010).

Indeed the denomination "internal conflict" itself should compel us to explore the phenomenon at a sub-country level. As formulated by Cederman and Gleditsch (2009): "If our theories are disaggregated, then our empirical analyses and research designs should reflect this" (Cederman

[3]Among the most seminal of these studies are Fearon and Laitin (2003), Collier and Hoeffler (2004), and Hegre and Sambanis (2006)

and Gleditsch, 2009, 490). Now, using sub-country units requires more computational power, better models, and not least the right data. Such obstacles have previously impeded sub-national analysis on a wider scale. Encouragingly, recent developments in statistics, technology, and data availability address these issues and make disaggregated studies evermore manageable (O'Loughlin et al., 2010, 446).

By using a disaggregated approach I am able to analyze the local temporal and spatial dynamics of conflicts at a level which both yields more practical insights and is more appropriate given the theoretical foundation (O'Loughlin et al., 2010, 446). If I were to model the spatial patterns of conflict using countries as the unit of analysis, I would be hard pressed to produce insights beyond those most trivial, such as "If country A is experiencing conflict, a neighbouring country B might be at a greater risk of also experiencing conflict". This would tell us nothing about where the conflict is located in country A; whether it is getting closer to country B; whether it engulfs the boarder between A and B; or where in country B it develops. A more disaggregated approach allows me to produce deeper and more fine-grained insights regarding the diffusion of conflict – a potential powerful policy tool indeed.

This concludes my presentation of two recent and highly relevant developments in the fields of conflict studies and conflict predictions. In the next section I will focus on the challenge at hand; using conflict patterns as conflict predictors.

3 The Present Challenge

Whether the goals have been explanation or prediction, past efforts have usually focused on structural features such as poverty, deprivation, resources, political regimes etc. (Chadefaux, 2017, 10). That being said, temporal and spatial patterns of conflict have not gone unnoticed. In the proceeding sections I will first present empirical motivation for using past patterns as opposed to structural features as the basis of my forecasting framework. Secondly, I will present some of the insights recently produced by the literature on conflict diffusion and contagion. Lastly, I show how past efforts have tried to incorporate the patterns of conflict in predictions and explanations alike.

The point is to juxtapose the theory of conflict patterns with past implementations illustrating *what not to do* while also formulating *what we should do*, thus creating clear theoretical criteria for how conflict patterns should be modelled. These criteria will then serve to qualify the tools I use in the project; particularly Gaussian processes.

3.1 Why Conflict patterns

When reliable prediction is the primary goal, we need features with substantial predictive potential. In von der Maase (2019) the predictive potential of various features was evaluated in order to map

fertile paths for future research regarding the construction of an early-warning-system. The paper showed that, while structural features such as poverty, deprivation, population size, country size etc. did contribute with prediction power, the most important features – by far – where those pertaining directly to the spatial and temporal patterns of conflict. Specifically three features: Distance from the geographic unit to the nearest conflict, all past fatalities in the geographic unit, and number of past conflict years in the geographic unit (von der Maase, 2019, 17-18). As such, the conclusion was that further predictive efforts should focus on extracting even more information from the temporal and spatial dimensions by developing more theoretically and methodically appropriate features pertaining to these dimensions (von der Maase, 2019, 21-23).

A second very important advantage when the goal is actual forecasting is that the data pertaining to the patterns of conflict is more readily available, more up-to-date and more current compared to structural data. To capture the spatial and temporal patterns of conflict I only need event data. That is, data pertaining to past conflicts themselves. Structural data such as wealth measures often take a lot of time to gather and process. The consequence being that the last observations of such data is often quite dated by the time the data is made public. Event data such as that used in this thesis is released much more frequently with the last entry being much more current. As an example, if I used structural data for the project at hand, it would be hard to get complete data more current than 2015. The event data used for this effort, however, has its last entry at 2017, with 2018 already available[4]. As such, the topicality of event data is another important advantage over structural data when forecasting is the goal.

3.2 What we know about conflict patterns

It is well known that conflicts exhibit discernible patterns. I here present four specific insights from the literature on conflict patterns. Firstly, it has been firmly established that conflicts cluster in time and space, with Crost et al. (2015) listing no less then 20 published academic papers supporting this assertion (Crost et al., 2015, 15). Secondly, the emerging consensus is that the clustering and diffusion of conflict is often a direct product of contagion rather then merely a bi-product generated through clusters of structural features such as poverty or political regimes (Buhaug and Gleditsch, 2008; Schutte and Weidmann, 2011; Crost et al., 2015; Bara, 2018). Thirdly, it has been found that conflicts often diffuse seamlessly across any administrative boundaries they might encounter (O'Loughlin et al., 2010, 442-443). Lastly Schutte and Weidmann (2011) show how patterns of internal conflict are characterized by distinctive patterns comparable across cases. Indeed, when conflict engulfs one location it is likely to expand outwards from the location over time (Schutte and Weidmann, 2011, 151). This pattern of conflict can, at its most fundamental level, be seen as exhibiting bell-curve (or Gaussian) like properties. The closer you

[4]While the data for 2018 has been available from UCDP since 03-06-2019, it was not included in this project since all the most demanding computations where completed before this release. Since I do not actually here forecast into true future years such as 2020 or 2021 the inclusion of the 2018 data would have delayed the completion of this project, without adding much value to the effort.

are to an active conflict in time and/or space, the more likely it is to spread to your location, largely unimpeded by national boarders. A simple but powerful guideline which will prove very useful going forward.

The challenge is to construct a framework which incorporates such insights into the actual modeling of conflict patterns. Unfortunately, past efforts which have aimed at taking conflict patterns into account often fall short of creating theoretically and methodologically coherent operationalizations. Indeed it would not be unfounded to call many past solutions underdeveloped. Naturally this contention requires elaboration and justification which is best done through the power of examples. This leads me to the next subsection.

3.3 How (not) to use patterns as predictors

I will start by presenting how the temporal dimension of conflict patterns have previously been modelled. A deterioration index recently proposed by Perry (2013) can serve as an example of the general state of affairs. Perry's idea is related to the phenomenon of a conflict trap and the notion of some inertia in conflict which might deteriorate over time. This time deteriorating index would be created by including the number of fatalities in some geographic unit for each of the last ten years as features. The deterioration rate is then incorporated through down-weighting these fatalities by dividing with the number of years passed since the corresponding events. Thus, a feature pertaining to fatalities two years ago will have, as its values, half of the fatalities observed that year (Perry, 2013, 14). In short, the proposed index tries to fit the insight that the closer in time a geographic region is to past conflicts, the more likely the region is to experience conflict again.

As such the heart is at the right place and the sentiment presented in Perry (2013) is also rather symptomatically for the literature at large. The problem is that it is an ad hoc and underdeveloped solution. There is no reason to cap the effort at ten years and there is no theoretical or practical reason to choose the suggested deterioration rate. Instead of dividing with years past it might be more appropriate to divide by half that; or the deterioration rate might have an altogether different functional form, such as an exponential or linear decay function. The point is that if we do not know the relevant functions, efforts should been made to estimate them rather than guess them.

Some scholars have used slightly more involved solutions. Collier and Hoeffler (2004) use a linear decay function counting years since last conflict and Hegre and Sambanis (2006) use a linear decay function counting years since last peace. Yet, while such frameworks come closer to actual estimation, the specification of these functions still require ad hoc and arbitrary decisions (Gelman et al., 2013, 501). Others such as Cederman et al. (2013) and von der Maase (2019) simply count the number of previous conflicts in the geographical unit of analysis. Surely, we can do better if we employ methods capable of estimating actual functions directly derived from the relevant data.

Turning to the spatial patterns of conflicts the challenges are, to a large extent, the same. However, the enduring focus on countries inflates the problem even further. To exemplify, a simple dummy is often used to indicate whether some predefined number of neighbouring countries are experiencing conflict or not (Hegre and Sambanis, 2006; Goldstone et al., 2010). The thresholds here often appear ad hoc as Hegre and Sambanis (2006) has a dummy denoting 1 or more "bad neighbours" (Hegre and Sambanis, 2006, 521-522) while Goldstone et al. (2010) has a threshold of 4 or more bad neighbours (Goldstone et al., 2010, 197). Surely the mechanisms determining whether conflicts spread from A to B are more complex then this and surely we can do a better job of emulating them.

Encouragingly, a few scholars have started to explore more complex patterns, taking into account the size of shared boarders, the country size of bad neighbours, the presence of trans-boarder ethnic kinship, and the explicit number of bad neighbours (Buhaug and Gleditsch, 2008; Cederman et al., 2013; Bara, 2018). This is definitely a progressive development, yet the enduring focus on countries still hampers the development of micro-level insights concerning where precisely a conflict unfolds in a given country and where it will expand and diffuse to its surroundings.

Two studies which do take a disaggregated approach are O'Loughlin et al. (2010) and Weidmann and Ward (2010). O'Loughlin et al. (2010) do a commendable job of showing the diffusion of conflict from Afghanistan to Pakistan over the Duran Line from a disaggregated perspective, while Weidmann and Ward (2010) illustrate how spatial and temporal patterns influenced the civil strife in Bosnia between 1992 and 1995. The challenge is that, using a disaggregated approach, we still have to decide how to include the effects of bad neighbours – the neighbours now simply being some sub-country unit. Indeed the uncertainty of how many bad neighbours to include only amplifies as we move from country to the sub-country level. Here 2^{nd}, 3^{th}, 4^{th} and potential n^{th} order neighbours will have to be considered, preferably with some demising influence as a function of distance, given the insight from Schutte and Weidmann (2011). Furthermore, it seems logical that the magnitude of adjacent conflicts also have an important part to play.

Unfortunately, both O'Loughlin et al. (2010) and Weidmann and Ward (2010) choose only to include a dummy for conflict in 1^{st} order neighbours. The operationalization presented in von der Maase (2019) is hardly any better; here distance to nearest conflict is used without taking into account the magnitude of this conflict, or the number of other adjacent conflicts zones. And even if the study had included the 2^{nd} nearest conflict, the 3^{th} nearest conflict etc., and scaled the distance by some factor related to the magnitude of conflict, we would still not know if these conflicts surrounded and engulfed the observation of interest or instead clustered neatly and distinctly beside it. The pattern of conflict might matter; the magnitude of conflict might matter; the magnitude of adjacent conflicts in 1^{st}, 2^{nd} and n^{th} order might matter. Again the point is clear: modeling features to capture these phenomena must be an estimation effort in and of itself.

To reiterate, what we need with regards to the temporal dimension is a tool that allows incorporation of information drawn from all past (included) years. The magnitude and volatility of

past conflicts should influence the probability of future conflicts with some decreasing influence as a function of time. Furthermore this deterioration rate should be estimated through the data itself, rather than being guessed by the researcher.

In regard to the spatial dimension, the criteria are rather similar. We need a tool capable of including information from all relevant events, taking into account both the magnitude and distance of all relevant conflicts. Notably, I do not want to decide how close a conflict has to be in order to be included. Nor do I want to decided the deterioration rate of influence. These elements should be estimated from the available data. Furthermore, the tool should be able to distinguish between different patterns of conflicts. E.g. I want encirclement to exhibit greater influence than tangency.

In the next section I present a tool capable of fulfilling all these criteria: The machine learning technique of Gaussian processes. The goal is to demonstrate the nuts and bolts of the tool before I move on to the actual implementation. The section below will also briefly introduce a number of other computational tools used to bind the whole effort together.

4 The Appropriate Tools

The main focus in this subsection will be Gaussian processes. In addition I will briefly introduce the final predictive framework and the evaluation metrics used in the final evaluations effort. First however, a small subsection will introduce the concept of *feature engineering*.

4.1 Feature engineering – capturing theory with data

Creating features from raw data is called feature engineering. In conventional machine learning this means modifying the data to maximize its predictive power. In political science this often means modeling the data to more accurately capture the theoretical mechanism connecting x to y. This could involve changing a wealth measure from absolute to relative, if one believes relative deprivation to be more important than greed or state capacity. When I talk about extracting patterns from event data, I am talking about feature engineering.

Theoretically appropriate features are naturally needed if we wish to estimate a proposed theoretical relationship between two phenomena (Blimes, 2006; Cederman et al., 2013). However, this thesis focuses on prediction and I do not estimate any relationships per se. That being said, theoretically viable features will allow us to create models which strives to emulate the true data generating process, thus increasing the predictive power of the model (McElreath, 2018, 209-211). As the criteria formulated above illustrates, theory can go a long way in guiding our feature engineering effort – also when prediction is the prime objective.

The feature engineering I present in this thesis, will be an estimation effort in itself: Not

estimating a limited number of parameters as in traditional linear regressions, but estimating complete functions pertaining to the diffusion of conflict through time and space. To this end I will employ the machine learning technique of Gaussian processes.

4.2 Gaussian Processes – capturing theory with method

The point of this subsection is to familiarize the reader with the basics of Gaussian processes and concurrently illustrate how Gaussian processes conform beautifully to the criteria formulated for modelling conflict patterns. I will not get overly technical, but some intuition is needed in order to appreciate the link between theory and method.

4.2.1 From lines to functions

To get an appreciation for how Gaussian Processes works, it is helpful to start with a simple linear relationship. In the example presented in equation 1, a linear relationship between a target y and the feature x is described by a slope parameter β and an error term ϵ. The parameter β is drawn from a Gaussian distribution ($\mathcal{N}$) with mean μ and standard deviation σ:

$$\tilde{y_i} = \beta x_i + \epsilon \tag{1}$$

$$\beta \sim \mathcal{N}(\mu, \sigma) \tag{2}$$

Now, the linear form presented above only allows the relationship between y and x to be described by a straight line with the slope given by β. We rarely see a linear development in temporal trends, – at least not in the long – and regarding spatial patterns it is hard to imagine any natural or social phenomenon taking this shape. As such, we want a way to model more complex functional forms. As a solution we might include nth order terms, which effectively would allow us to depict a function of arbitrarily large complexity:

$$\tilde{y_i} = \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \cdots + \beta_n x_{in}^n + \epsilon \tag{3}$$

However, unless we have very clear theoretical presumptions, we would not know how many orders to include. This would lead to arbitrary decisions. Furthermore, the approach of using higher order transformations is a well known source of overfitting (Williams and Rasmussen, 2006, 2). We need a framework which can accommodate complex forms and simultaneously combat overfitting.

Instead of including n^{th} order terms let us imagine some relationship where the function f is given by a Gaussian Process $\mathcal{GP}$. The Gaussian process is closely related to the Gaussian distribution, but instead of being a distribution of numbers, it is a distribution of functions (Williams

and Rasmussen, 2006, 13-15). As such, equation 5 does not denote any specific functional form for f. Instead, it denotes that the function f is drawn from a Gaussian process – just as we draw numbers from a Gaussian distributing. Rather than being defined by the parameters μ and σ a Gaussian process is given by a *mean function* m and a *covariance function* k.

$$\tilde{y} = f(x) + \epsilon \tag{4}$$

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{5}$$

The mean function m is our assumption in the absence of knowledge. Where data presents itself, the mean function adapts to the data (Williams and Rasmussen, 2006, 3-4). In Bayesian terms this mean function can be understood as a weakly informative prior and in frequentist terms as a weak regularization (McElreath, 2018, 35). Notably, only very little data is needed to overshadow the mean function. Indeed the surrounding observations and the covariance function k is much more important.

Usually the mean function is set to the constant zero, but it need not be the case – it could be any other constant such as the mean of y, a slope or any other function which might be appropriate given the specific theoretical underpinnings (Williams and Rasmussen, 2006, 28-29). In our case, the median and the mode of our target is 0 and the mean is rather close; most places do not experience conflict most of the time – and since the target is logged, even large outliers, such as the genocide in Rwanda 1994, do not overly influence the mean. As such, a constant of zero also appears as an appropriate mean function here.

Furthermore, the only situation in this project where I do not have data dictating the mean function is when I am forecasting. Proximate forecastings will be determined by the data pertaining the preceding years and the covariance function k. Forecasting to more distant years will increasingly be influenced by the predetermined mean function. That is, the further we move away from the last observed year, the less the data can tell us about future conflicts and the more influence the mean function will claim. Notably, the mean function will only dominate, when I extrapolate into a future too distant for data to reliably describe – at which point the projections should not be consulted anyhow. Thus, as presented in equation 6 a mean functions of zero will be used going forward.

$$m(x) = 0 \tag{6}$$

The covariance function k is the crux of the machinery, often referred to as the kernel. It describes the similarity between each of the observations x and x'. It effectuates the assumption that two observations with similar values on x should also have similar values on y (Williams and Rasmussen, 2006, 79). In the case of time and space this is a quite reasonable assumption.

The covariance function can be understood as a similarity measure. Many similarity mea-

sures are valid covariance functions – each appropriate for specific tasks given specific theory and assumptions (Williams and Rasmussen, 2006, 79). In this thesis I use the *squared exponential covariance function*. This covariance function is widely applied and is often a safe choice since the only assumption attached is that we are estimating smooth functions (Williams and Rasmussen, 2006, 84). As such, it is often used to model trends (Williams and Rasmussen, 2006, 119).

Naturally, if we find it unsafe to assume smoothness or safe to make more detailed assumptions, other covariance functions might be appropriate (Gelman et al., 2013, 502-503). It is beyond the scope of this thesis to explore different covariance functions, but future endeavours should test whether more appropriate kernels exist. Here, the squared exponential covariance function will serve as the covariance function of choice going forward. It is given in equation 7.

$$k(x, x') = \eta^2 exp\left(-\frac{|x - x'|^2}{2\ell^2}\right) \tag{7}$$

From equation 7 it is clear that this is simply a similarity measure. What stands out are the *hyperparameters* ℓ and η. Hyperparameters in this context meaning scale parameters, which can be estimated and determine the volatility of the functions drawn from the Gaussian process.

The ℓ is denoted the *lengthscale* (Gelman et al., 2013, 501-502) and can be interpreted as the distance we have to move across our x-axis before the functional value can change drastically (Williams and Rasmussen, 2006, 14-15). As such, in the present endeavour ℓ also serves as an appropriate rule-of-thumb-limit regarding how far into the future we should consult the extrapolated forecasting. Beyond the last observed year + ℓ our data provides little to no information, uncertainty proliferates, and the predefined mean function will start to dominate.

The η is denoted *amplitude* or *output variance*, and roughly determines the average distance the functions vary from their means (Gelman et al., 2013, 502). While η serves a central part in the covariance function, the substantial interpretation is only of secondary interest given the focus of my project. Thus, I will only devote little attention to this hyperparemater going forth.

4.2.2 Exemplifying Gaussian Processes

In Figure 1 I have drawn 15 samples from three different Gaussian processes given three different, pre-defined, sets of hyperparameters. The functions are randomly drawn from the *prior*; that is the function space before any data have been introduced to the Gaussian processes at hand.

The thick line is denoted μ and is the mean value of all 15 functions. It is routinely used as the *most-likelihood* estimate of the functional form, and without any data it defaults to the mean function (Williams and Rasmussen, 2006, 3). The volatility of μ seen in Figure 1 is only due to the small number of samples drawn. The shaded area denotes $\mu \pm 2\sigma$ where σ is the standard deviations. Without any data introduced all functions given by the mean and covariance function are valid estimates. However, introducing data makes some functions more likely than others,

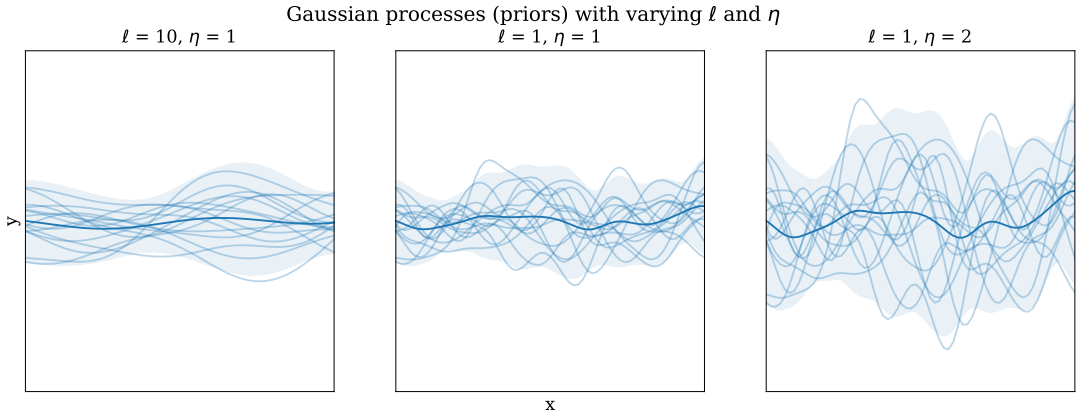

Figure 1: 15 functions drawn from three different GPs (priors) with different hyperparameters ℓ and η. All three GPs uses a squared exponential kernel function and a mean function = 0

effectively limiting the forms of the functions. This is illustrated in Figure 2 where the sampled functions now accommodate the data where it is available.

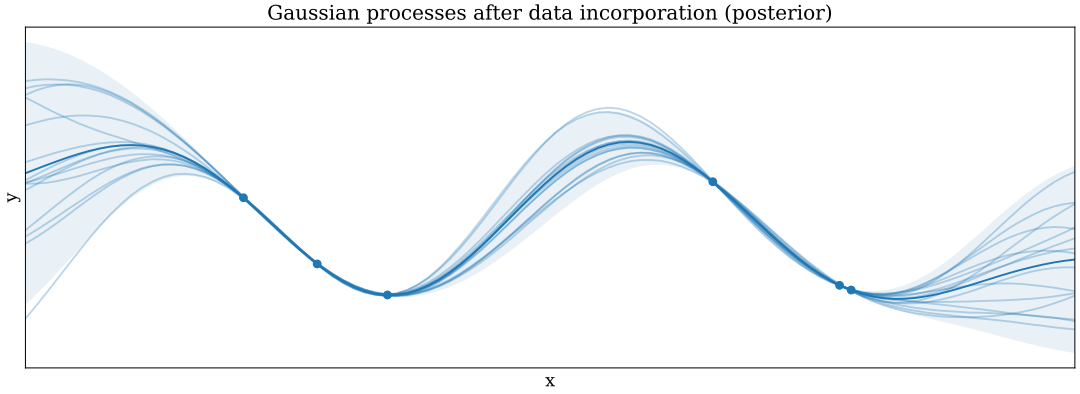

Figure 2: 15 samples from a Gaussian process after the introduction of data. The points represent the data. The thick line is μ and the shaded area represents $\mu \pm 2\sigma$.

We now see that uncertainty shrinks in regions illuminated by data, and grows the further we move away from data. Note that this Gaussian process is defined without any error term ϵ, thus forcing all drawn functions to cut directly through the data points.

It should be clear from Figure 1 and Figure 2 that if I were to determine the hyperparameter myself and omit an error term ϵ, I could fit most smooth patterns with total precision. This would lead to overfitting and furthermore, I would not have handled the issue of researchers guessing at functions rather than estimating them.

Encouragingly, this can be amended elegantly. We included an error term ϵ to the Gaussian processes, which along with ℓ and η, will be estimated through a framework which actively dis-

courages overfitting (Williams and Rasmussen, 2006, 114-115). Using the estimated error term and hyperparameters in tandem with our data will generate Gaussian processes which will capture the general patterns of the data, but not overfit to noise.

These estimates are obtained by maximizing the *marginal likelihood* of the given model (Williams and Rasmussen, 2006, 114-115). This is a technical and non-trivial mathematical operation which I shall not explore further here[5].

A toy-example, emulating the subject of conflict magnitude, can be seen in Figure 3. This could be the development in three different conflict-ridden geographic units from 1989 to 2018. Here ℓ, η and ϵ are estimated from the data. I have only included μ, as the solid lines and the shaded areas represents $\mu \pm 2\sigma$ given the underlying samples. Since we have observations for each year until 2018 there is very little uncertainty regarding μ during the observed years. Note that the inclusion of ϵ allows μ some freedom to vary from the observations, thereby discouraging overfitting to noise.

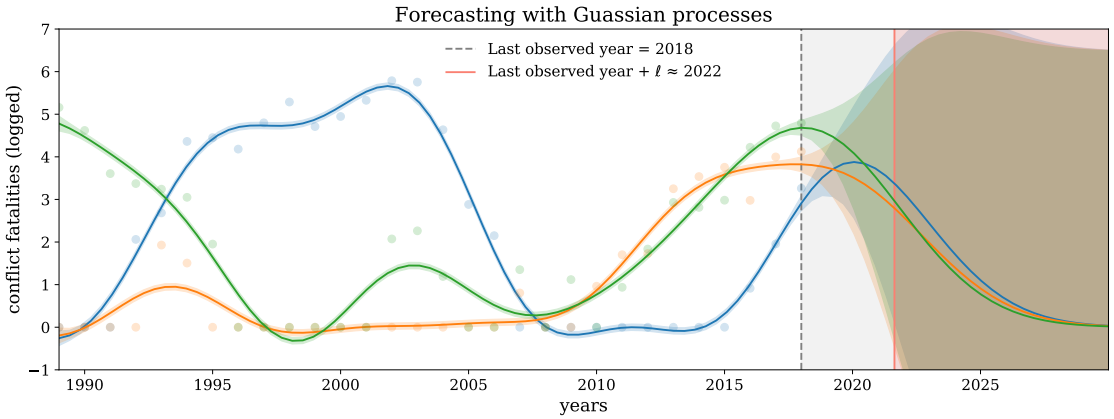

Figure 3: Three (generated) time-lines of conflict-ridden geographical locations. $\ell = 3.64$, $\eta = 2.54$, $\epsilon = 0.51$

In Figure 3 I have extrapolated 10 years ahead to illustrate how the hyperparameters determine the functional form of the functions in the absence of data and how the mean function takes over as we venture into the unknown future. Crucially, I do not estimate a separate sets of hyperparameters for each of the three time-lines, but use the information from all three time-lines to estimate shared hyperparameters for all observations – so-called *multi-task learning* (Williams and Rasmussen, 2006, 115).

While I do use the same hyperparameter for all time lines in my implementation, I will not use all of the time lines when I estimate the hyperparameters. The reason is that most time lines do not experience conflicts and as such do not tell us anything about conflict patterns. Including such "flat-lines" would yield a misleading estimate of ℓ. This is because ℓ tells us how far we

[5]Curious readers should consult Williams and Rasmussen (2006) section 5.4.1 for a complete overview and implementation of the operation.

have to move on x to see changes in y and time lines with no conflicts have a consistent conflict magnitude y of 0 no matter which year x we look at. The result would be an estimated ℓ which would reflect that conflict magnitude is largely constant across time [6]. Naturally, this is only the case for time lines experiencing no conflicts and not time lines actually experiencing conflicts. To accommodate this issue the hyperparameters are only estimated using time lines with at least two years of conflicts. While two might seem an arbitrary number, it stands to reason that you need at least two none-zero observations to constitute a pattern.

As such, I assume that the broader patterns of conflict across time (and later space) are comparable across the observations[7]. While the results from Schutte and Weidmann (2011) do support the use of the same covariance functions across all cases, it would be too bold to claim that the results conclusively support the use of the same hyperparameters across cases. The validity of this assumptions should be scrutinized in future endeavours. For now, however, this will do.

In the toy example ℓ is estimated to be 3.64 with 95% probability of being between 2.46 and 4.56. A so called 95% *credibility interval*, which is similar to a traditional *confidence interval* but based on the actual samples rather then estimated (McElreath, 2018, 54). Thus, in this example it would be imprudent to extrapolate beyond year 2022 as indicated by the red line and shade in Figure 3. Importantly, this hard limit is only a rule-of-thumb, here given by the mean estimate of ℓ. The specific uncertainty should **always** be consulted in practice.

To model the spatial pattern, I simply use latitude and longitude instead of years as features. Here, the point is not to forecast into some unknown geographical space, but rather to let each observed cell incorporate information from all other cells in the data to estimate how exposed the given cell is to conflict. Naturally, the spatial and temporal patterns can also be combined into 3D space including two spatial and one temporal dimension, thus allowing for extrapolation of conflict exposure into future years.

Having both estimated temporal patterns of conflict magnitude and conflict exposure, forecasting is simply a matter of introducing new future years x_{new} to the time lines and generate Gaussian processes over these elongated time lines.

Having functions extending into future years also allow for easy extractions of measure pertaining to *slopes*, *acceleration* and *mass*. In practice, the estimate I produce is not an actual formula for the function f but merely the values generated by the function f. As such equation 11, 12 and 13 denotes what I derive and extract, but not the actual mathematical or computational operation used. The values from f', f'' and F are simply measures derived from the function mu and can also be used as individual features in a greater predictive framework.

[6] Indeed, preliminary tests showed that the models would often not converge at all, if I did not add a minimum of noise to the "flat lines".

[7] Using separated hyperparameters for all observations would be going too far and would mean that we do not think conflict patterns share any similarity across cases. Yet, less demanding assumptions could be made if we used a hierarchical structure which allowed the hyperparameters to be similar without being identical. Notably, that would entail a substantial increase in the computational resources needed.

$$\text{slope} = f'(x) \tag{11}$$

$$\text{acceleration} = f''(x) \tag{12}$$

$$\text{mass} = F(x) \tag{13}$$

Now, revisiting the criteria I presented earlier, we see that Gaussian processes fulfill them all. In regards to the temporal pattern, they allow incorporation of information drawn from all past (included) years with decreasing influence (McElreath, 2018, 410-419). Furthermore, the rate with which past years influence the forecastings is estimated through the data itself. These properties are the reason why Gaussian processes are often used to model functions over time (Williams and Rasmussen, 2006, 13). Furthermore, no arbitrary "splines" or "knots" needs defining, which is one reason it has been recommend as a substitute to linear decay functions (Gelman et al., 2013, 501). In regards to the spatial dimension, utilizing Gaussian Processes on a disaggregated geographical grid allows us to analyze and model spatial patterns at arbitrarily high complexity levels only hampered by the resolution of the data and computational power available. We need not define the number of neighbours, since all observations are considered[8]. Furthermore, we do not have to guess the deterioration rate of influence since this is estimated. Lastly, using Gaussian processes to model spatial diffusion, a given geographical cell will be influenced directly by the pattern and magnitudes of conflict around it, in a manner similar to the conflict patterns identified by Schutte and Weidmann (2011). High magnitude will translate to higher influence, as will encirclement compared to tangency. Indeed, these properties are part of what compelled Gelfand and Schliep (2016) to declare the union of spatial data and Gaussian processes a "beautiful marriage" (Gelfand and Schliep, 2016, 86).

Notably however, it should also be clear from Figure 2 that the framework presented here is somewhat misspecified. Since my base measure for conflict is conflict fatalities (logged), one could argue that I should treat the data as count data. Since I will not be using the extrapolations directly as estimations of death tolls, getting fractions of deaths is not the problem. Instead the issue is that I will incidentally estimate negative conflict magnitude and exposure. This will happen proceeding steep dives in temporal or spatial patterns.

The issue could be handled by formulating the regression presented in equation 4 as a *Poisson regression*. This is conceptually trivial, but unfortunately computationally expensive. Somewhat encouragingly however, I ran initial tests which showed little to no difference when using a Poisson framework. The estimated μ's rarely dips below zero, and when they do, it is only marginally. As

[8]That said, we could choose some appropriate subset to improve computation massively (Gelfand and Schliep, 2016) – but, this is a project for an other time.

such, I allow this misspecification to persist for the purpose of simplicity and efficiency[9].

The next section will briefly present the theoretical underpinnings of the predictive framework which will transform the patterns estimated into predictions.

4.3 The predictive framework

Gaussian processes can be used as predictive models in themselves; after all we forecast expected values into future years. However, to combine these forecasted values of conflict magnitude, conflict exposure and the derived measures of slops, accelerations and mass into unified predictions of conflict probability, we need another tool. Indeed, Gaussian processes are often utilized as subcomponents in larger models (Gelman et al., 2013, 505). A suitable predictive framework has been presented in von der Maase (2019). Since the nature of this framework is not a prime concern in this thesis, I will keep this presentation very brief and concise. A more in-depth presentation can be found in the original paper (von der Maase, 2019, 9-12).

The framework consists of a number of individual elements best introduced in turn. First of all we need some algorithm capable of using the patterns extracted to estimate the probability of an event. We know that conflicts are complicated phenomenas ridden with *interactions* (Cederman and Weidmann, 2017, 474). Thus, the predictive algorithm needs to be able to identify and generate such interactions. Furthermore, we know that conflicts are rare events. As such, the algorithm also needs to be suitable for handling rare events. For the sake of both policy recommendations and theoretical discussion it is also important that the algorithm is not a *black box*; we must be able to asses which "decisions" facilitate the results (Cederman and Weidmann, 2017, 476). Secondly, even an algorithm tuned for rare events can have difficulty handling the rarity of conflicts. As such I need tools which can mitigate some of the ills of imbalanced data. Finally, we need appropriate metrics for evaluating the potential of the endeavour as a whole. Metrics which again take into account the rarity of the event give us an honest evaluation of the predictive potential of the approach. In the following subsections I show how the extreme gradient boosting algorithm, undersampling and the precision-recall curve can fulfill these roles.

4.3.1 Extreme Gradient Boosting

The algorithm I use to estimate both the probability of conflict is called extreme gradient boosting or simply *xgboost* (Chen and Guestrin, 2016). It can be constructed as a regressor to predict a number, similar to a linear regression, or as a classifier to predict probabilities, similar to a logistic regression. In this thesis, I will use it as a classifier. Particularly three characteristics of the xgboost classifier deserve attention; it is a *boosting* algorithm; it consists of *regression trees*;

[9] A choice which is not unheard of (Williams and Rasmussen, 2006, 123). A similar example can be found in Gelman et al. (2013) chapter 21. Effectively, it is the same choice made every time a linear regression is used on count data such as death tolls, births, employment and indeed most data revolving around items or people.

and it is *self-regularizing*.

Boosting is used in a variety of machine learning algorithms and entails aggregating a lot of weak classifiers to create a single strong classifier (Friedman et al., 2001, 337). To this end, we evaluate the results from the weak classifiers and on basis of these results we distribute different weights across our observations. We now run a new batch of classifiers but this time, the observations which were correctly predicted in the first round will carry less weight while observations which were incorrectly predicted will carry more weight. This procedure is iterated until some predefined criteria is met. Lastly, all classifiers are weighted according to their performance and used as a predictive ensemble to estimate the probability of some event (Friedman et al., 2001, 338-339). This procedure ensures a continuing focus on "hard to classify"-observations, and as such it is a particularly useful approach when working with rare events.

Naturally, we need to decide which weak classifiers to use for our boosting. Xgboost utilizes regression trees – which strictly speaking are regressors and not a classifiers but in the xgboost algorithm they are used as classifiers in non-the-less[10] (Chen and Guestrin, 2016, 786). We use these regression tress to partition our data according to the split-points which best sort our data according to a predefined target (Chen and Guestrin, 2016, 786) In this project it means using my eight features to split grid cells experiencing conflict from cells not experiencing conflict. The features and specific split-points to be used are automatically found by identifying the splits that minimizes a specific *objective function*[11](Chen and Guestrin, 2016, 786). We keep splitting partitions until we fail to minimize the objective function further (Chen and Guestrin, 2016, 786). What is crucial about this procedure, is that each split (apart from the very first) effectively constitutes an interaction. Thus the xgboost algorithm is able to automatically identify and generate the most salient interactions.

An another important feature of the specific objective function is that is is self-regularizing thus mitigating overfitting. Specifically the objective function penalizes complex tree structures. When the algorithm decides whether to create a new split, it compares the potential improvement in prediction power to the added complexity. If the added complexity is to great compared to the improvement in prediction power the split is not effectuated (Chen and Guestrin, 2016, 787-791). Thus, the regression trees used in the boosting effort searches for relevant patterns in the data, but automatically stops before overfitting commence.

Since the xgboost algorithm is based on tree structures, we can readily asses which splits generated most prediction power, and thus which features are most *important* in regards to the combined prediction effort. Importance here meaning how much prediction power the specific feature contribute with, compared the the other features (Chen and Guestrin, 2016, 787-788).

[10]The reason the xgboost algorithm uses regression trees rather then decision trees, even for classification, is that each tree produces a continues score as oppose to a binary classification at the end of each branch. These score holds more information then a hard classifications and they can be added together across all trees before being normalized to probabilities and used for classification

[11]I will not go into the mathematics of this function, but the curious reader can find it in Chen and Guestrin (2016) page 786.

As such, feature importance should be understood as a relative measure comparing how much information a specific features bring the the effort compared to the other features (Friedman et al., 2001, 367-368). While this tells us nothing about causation, it is still highly relevant for guiding future prediction frameworks, informing theoretical debates pertaining to explanations and for formulation of concrete policy recommendations.

Naturally, this is a very superficial introduction to the xgboost algorithm; nevertheless it serves to illustrate why this algorithm is especially suited for the endeavour at hand. The boosting element makes the algorithm suitable for rare events and the fact that is is based on regressions trees both allows for automatic generation of interactions and for assessment of feature importance. Importantly, the specific trees used in this algorithm are self-regularizing thus discouraging overfitting. However, I can still do more to handle the imbalance of the data. The next section will introduce a number of *undersampling* techniques which I implement to further tackle the rarity of conflict events.

4.3.2 Undersampling

When our data is highly imbalanced, the predictive algorithm will favor the classification of the majority class. In my case this would mean that the framework would be focused on predicting the absence of conflict and not the presence of conflict. A simple solution is to undersample the majority class. As such in our train set we drop a portion of non-events to even the ratio between events and non-events. This is called undersampling (He and Garcia, 2008, 1266-1267).

Specifically I combine two procedures; *case-cohort sampling* and *informed undersampling*. Case-cohort sampling implies, that I train my model using all available events in the training set together with a randomly drawn and equally sized set of non-events also from the training set (King and Zeng, 2001b, 142). This procedure is usually justified by arguing that more information is stored in the events than in non-events (King and Zeng, 2001b, 139). While this may be true, I still discard a lot of potentially relevant information this way. To amend this, I also use variant of informed undersampling. Instead of just using one model with one random subset of non-events I use a large *ensemble* of models each using a new random subset of non-events (He and Garcia, 2008, 1267). Specifically I run the xgboost algorithm 1000 times; each time with a new randomly drawn subset of non-events. As such I am effectively estimating a distribution of conflict probabilities for each cell thus taking advantage of all the information in events and non-events alike. I can then use the mean of these distributions as a maximun likelihood point estimate for the actual probability. Furthermore, given that I can also asses the variance of these distributions, I am also able to infer how certain I am of this point estimate by assessing the credibility interval.

One issue with this undersampling approach is that the specific probabilities produced will be somewhat inflated. All models are trained to believe that conflict is more common than it is. This can easily be amended using a *Bayesian prior correction* akin to what is presented in King and Zeng (2001b,a) and implemented in Goldstone et al. (2010). I use the overall probability of conflict

in the last observed year in my training set: 2012. This is simply the ratio between events and non-events. I denote the share of events $Pr(E_{2012})$ and the share of non-events $Pr(NE_{2012})$. Then, denoting the estimated probabilities of an event in a specific cell at a specific year $Pr(E_{estimated})$; the corresponding estimated probabilities of a non-event $Pr(NE_{estimated})$; and the corrected probabilities of events as $Pr(E_{corrected})$ the correction can be expressed as follows:

$$Pr(E_{corrected}) = \frac{Pr(E_{estimated}) \times Pr(E_{2012})}{Pr(E_{estimated}) \times Pr(E_{2012}) + Pr(NE_{estimated} \times Pr(NE_{2012})} \tag{14}$$

Together, the xgboost algorithm and the undersampling approach do a good job of priming the framework for rare events. However, both efforts do so through the model training; the test data that we feed our model will still be imbalanced. We cannot undersample this data since its role is to emulate an unknown future. Undersampling it would imply that we already know which cells that will experience conflict. Unfortunately conventional evaluation matrices tends to judge models used on such imbalanced data too favorable (He and Garcia, 2008, 1264). The next section introduces the evaluation metrics which I use for this project, and qualify why these are appropriate for evaluating unbalanced data.

4.4 Evaluating the Predictions

Given the predictive scope of my project, I employ out-of-sample prediction to evaluate the performance of my approach. Out-of-sample prediction is implemented by using a subset of my data to train my predictive framework and another subset to test the framework. Given that my goal is practical forecasting I use the last five years (2013-2017) of my data as test set. As such, I use a ratio of roughly 20% for the test set, which is rather conventional (Friedman et al., 2001; Ward et al., 2010). To train the predictive framework I use the target $y_{cmBinary}$ from the training set which is a binary measure simply denoting whether any conflict was present in the specific cell some specific year. The features I use are those constructed via Gaussian processes $X_{patterns}$ also pertaining to the training set. After the training is complete the extrapolated values $X_{patternsEX}$ covering the test years are introduced into my predictive framework to produce out-of-sample predictions $\tilde{y}_{prob}$. These prediction can then be evaluated against the empirical observations $y_{cmBinary}$ contained in the test set.

Now I need some metric which can capture how well $\tilde{y}_{prob}$ fits $y_{cmBinary}$. There are a plethora of such metrics available. The metric I will primarily use is the *precision-recall* (PR) curve and the three related metrics *recall*, *precision* and the *average precision* (AP) score. However, to put these metrics into perspective I will also introduce the metrics *accuracy* and the *Receiver Operating Characteristic* (ROC) curve along with the corresponding metric the *Area Under the Curve* (AUC) score. The reason I include more then one metric well become apparent as I introduce each measure.

The metrics used must fit the challenge; not least the fact that I am dealing with imbalanced

data. Furthermore, many metrics require that I set some hard threshold partitioning my predictions into predicted events and predicted non-events. The framework I use, however, does not produce binary results but probabilities. As such $y_{cmBinary}$ might by binary, either 0 or 1, but $\tilde{y}_{prob}$ can take any value between 0 and 1. These probabilities are in themselves far more informative than a simple binary classification. As such, the metrics I use should not depend on the formulation of some arbitrary threshold.

The most commonly used metric for classification is accuracy. Accuracy simply denotes the proportion of correctly predicted observations. As such accuracy is easy to interpret and is intuitively appealing. However, there are two problems. Firstly, the measure requires that we choose a hard threshold. Secondly, when the data is imbalanced we can achieve a rather high accuracy just by predicting in favor of the majority class every time. If conflict only happens 5% of the time we can get an accuracy of 95% by predicting that conflict never happens. Thus accuracy is not suited for imbalanced data (He and Garcia, 2008, 1264). As such, even though it is the most commonly known measure, I will not be using this metric going forward.

The metric most commonly chosen to avoid the problems afflicting accuracy is the ROC-curve and the summarizing measure the AUC scores (He and Garcia, 2008, 1277-1278). The ROC curve circumvents the issue of a hard threshold by evaluating each possible threshold. Specifically the ROC curve denotes the trade-off between the *true positive rate* (TP_{rate}) and the *false positive rate* (FP_{rate}) by plotting the TP_{rate} over the FP_{rate}. This curve can then be interpreted visually or summarized by the area under it; the AUC score (He and Garcia, 2008, 1277-1278). Denoting the actual number of negatives N_C and the actual number of positives P_C, the rates can be expressed by equation 15.

$$TP_{rate} = \frac{TP}{P_C}; \quad FP_{rate} = \frac{FP}{N_C} \tag{15}$$

As such the ROC curve does not require me to specify a hard threshold and it illustrates clearly the trade-off between true positives and false positives at all possible thresholds. Given these attributes it has been widely used in conflict studies (Chadefaux, 2017, 14), and even been coined as the "gold-standard" in this field (Perry, 2013, 366).

However, while better suited for imbalanced data than accuracy, the ROC curve and AUC score tends to judge model-performance on highly imbalanced data too favorable (He and Garcia, 2008, 1278). Looking at equation 15 it is clear that if we can increase N_C without increasing the FP rate we will get a better score. As such, including Antarctica would probably improve my results substantially. Antarctica would constitute a large number of non-events (N_C) and given my focus on past conflict patterns my framework would never produce a false positive (FP) here, leading to a lower false positive rate (FP_{rate}).

Despite this weakness I will still report the AUC score. The simple reason is that the metric is widely used and rather commonly known. The ROC curve and AUC score are points of reference

for most political scientist working with advanced quantitative methods and including the measure will make it easier to compare my project with past efforts. That being said, we should move on to more appropriate metrics better suited for highly imbalanced data. Therefore, my main focus will be on the PR curve.

The PR curve shares many similarities with the ROC curve. It also denotes the trade off between two measures, specifically recall and precision. Notably recall is just another name for the true positive rate TP, while precision denotes the rate of true positive out of all positives. As such, recall and precision can be expressed as in equation 16.

$$precision = \frac{TP}{TP + FP}; \quad recall = \frac{TP}{TP + FN} \tag{16}$$

It is clear that since $TP+FN = P_C$ recall is indeed simply the true positive rate. In substantial terms precision denotes the share of classified conflicts which were actual conflicts, while recall denotes the share of all actual conflict I manage to capture. These two measures are in themselves valid metrics and are both relevant when dealing with imbalanced data, but since these metrics rely on components such as TP's, FP's, TN's and FN's, both metrics require a hard threshold. Encouragingly, Combining them into a curve for all possible thresholds eliminates the need for any hard threshold and concurrently summarizes both metrics neatly (He and Garcia, 2008, 1287).

The PR curve can be interpreted visually and also be summarized in a single measure called average precision AP. This summarizing metric denotes a weighted mean of precision at each possible threshold. The weighting is the increase in recall from the previous threshold. Denoting recall R, precision P, and the different thresholds as n the AP score can be expressed as seen in equation 17.

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{17}$$

The AP score has a number of similarities with AUC and can be thought of as the area under the PR curve (Su et al., 2015, 349-350). The key difference is that AP places more emphasis on identifying high probability events than identifying low probability events (Su et al., 2015, 350). The interpretation also differs from AUC. An AUC score of 0.5 denotes a classifier which is no better than random. This is not the case with AP where 0.5 can be a decent score depending on the context (Su et al., 2015, 350-351). Specifically, an approximation of the "random" baseline of AP is given the share of events (Bestgen, 2015, 132). If the data is precisely balanced the baseline for the AP score is 0.5 just like the AUC score. If the data is imbalanced, say 5% events versus 95% non-events, the baseline for the AP score will be 0.05. As such, AP is a good summarizing measure for judging and comparing the predictive performance of models on imbalanced data without setting a hard threshold.

The downside of AP is that it is not particularly interpretable on a substantive level. As such,

to best present the potential of my framework in substantial terms I will use the AP score in tandem with recall and precision. In the same manner I will also use the components TP, FP, FN and TN for visualizations. This will naturally demand that I set some hard threshold. I use a threshold which generates a number of predicted events which roughly matches the number of actual events in the last observed year: 2012. Together these measures will ensure both honest evaluation and interpretability of my framework.

Having presented the methods which I will use to construct my predictive framework, I will now move on to present the specific data which I use. As such the next section serves as the last primer before I present the implementation and the results.

5 The Chosen Data

In this paper I use two data sources. A spatial grid which divides the world into cells is obtained from the PRIO grid database (Tollefsen et al., 2012) while the Upssala Conflict Data Program (UCDP) (Sundberg and Melander, 2013; Croicu and Sundberg, 2017) provides all substantial data used. I will present each source in turn below.

5.1 The PRIO grid

The PRIO grid is a global grid dividing the world – excluding Greenland and Antarctica – into grid cells of 0.5×0.5 decimal degrees, which corresponds to roughly $50km \times 50km$ at the equator (Tollefsen et al., 2012, 367). It is constructed as geo-spatial data and primed for collaboration with the data from UCDP. As such merging and handling these two data sources is a trivial task. However, due to computational limitations I have chosen only to use a subset of the full globe. This subset is presented in Figure 4.

In Figure 4 I have plotted all conflict fatalities (logged) classified by UCDP in 2017 aggregated at PRIO grid cell level in the chosen subset. While the exact boarders of the subset are rather arbitrary, I have aimed to capture as many conflicts as possible across the surveyed timespan in one continuous geographic area.

Naturally, this is selection on the dependable variable, and will lead to bias in any estimation effort (King et al., 1994, 129-130). However, I have no traditional parameters to estimate and no claim of effect or causality to make. Thus, since accurate prediction is the goal, selection on the dependable variable is less of a problem. Furthermore, any ills could easily be handled through yet another Bayesian correction (King and Zeng, 2001b, 627-628). That being said, it would still be imprudent to extrapolate any insights I produce – such as the estimated hyperparameters – beyond the geographical scope of this project. Correspondingly, the probabilities I produce should be seen as conditional on the geographical subset under scrutiny.

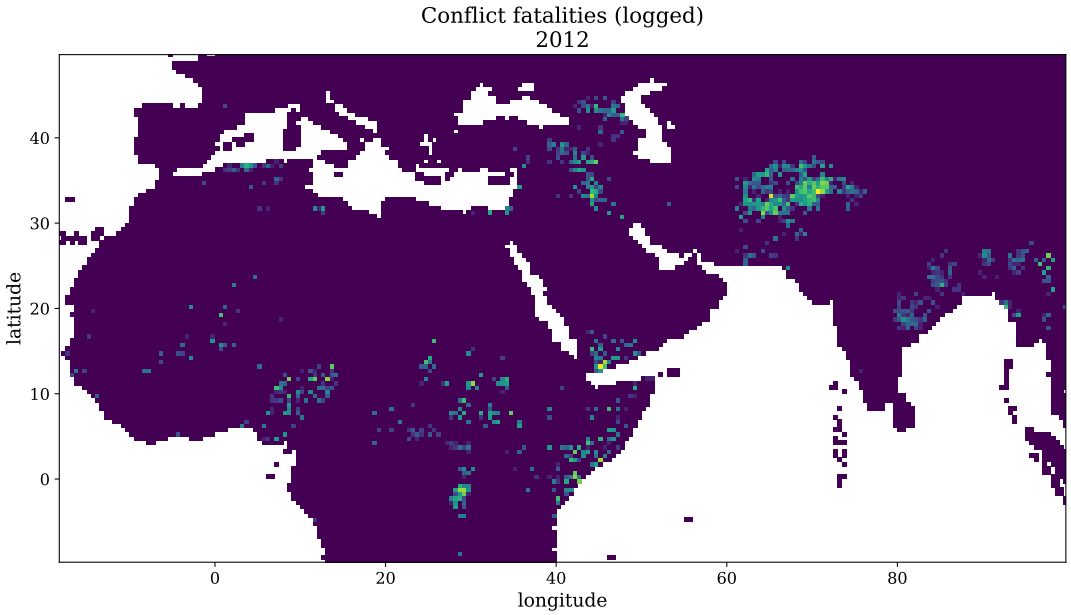

Figure 4: Conflict fatalities (logged) from 2017 according to UCDP aggregated at the level of PRIO grid cells. Encompasses the geografical area given by the following coordinates; north: 50° , south: −10° , east: 100° , west −20°

With this call for prudence I now introduce the substantial data which I will use to capture the patterns of conflict magnitude and exposure.

5.2 The UCDP data

My framework consists of two layers of estimations. First, I estimate the patterns of conflict magnitude and conflict exposure. Then I use these estimated patterns to estimate the probability of conflict in a cell at some future year. The only data I need for these two endeavours is data pertaining directly to the history of conflict patterns. For this I use data from the UCDP. Specifically I utilize the UCDP Georeferenced Event Dataset (GED) Global version 18.1 (UCDP, 2017). The data contains records of conflict fatalities and the corresponding coordinates and dates. I aggregate this data to yearly fatalities in grid cell and utilize data from 1989 through 2017.

I define conflict fatalities using the minimal definition from UCDP. They include in their data all fatalities 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, 9).[12]

Since I limit my project to intra-state conflict, I only include estimates from incidents which do not include two different nations as the organized actors. What is left are all conflict fatalities

[12]See definitions regarding "armed force" and "organized actor" page 10 and forth of Croicu and Sundberg (2017).

induced by internal conflicts, civil strife and terror.

There are a number of interesting features included in the UCDP data. I only use the feature *best*. This feature is defined as: "The best (most likely) estimate of total fatalities resulting from an event" (Croicu and Sundberg, 2017, 7). To get my raw measure of conflict magnitude, used as a target in my pattern estimation efforts, I take the log of this measure.

$$\text{conflict magnitude} = log(\text{conflict fatalities}) \tag{18}$$

The log-transformation is warranted since the data is highly skewed; across all years the majority of the cells experience no conflict fatalities, while a few cells experiences a lot of conflict fatalities. Furthermore a logged transformation mitigates the challenges of extreme outlets such as Rwanda 1994.

This measurement, conflict magnitude, aggregated at a yearly level and distributed across the PRIO grid is the basis of my whole framework. In the first estimation layer it will be the target y_{cm} as I estimate the patterns of conflict magnitude and years will constitute the single feature x_{year}. As I estimate the static conflict exposure for each year it will also be the target y_{cm}, while longitude and latitude will be the features X_{ll}. In the second layer, a binary transformation of conflict magnitude will constitute the target y_{binary} when I estimate the probability of conflict, while the features will be the eight features derived from the first estimation layer will constitute the features $X_{patterns}$.

Having presented the context of my project, the tools I employ and the data used, I now move on to the actual creation of my framework.

6 Compiling the framework

The following section is divided into two subsections corresponding to the two layers of estimation I employ. Firstly, I estimate the temporal and spatial patterns of conflict using my test set and Gaussian processes. These patterns will then be extrapolated into future years corresponding to those of the test set. From these elongated time lines I will derive the slope and acceleration pertaining to each year, along with the total conflict mass inhabiting each time line. This yields eight features. Secondly, using xgboost and undersampling, these eight features will be combined in my final predictive framework to estimate the probabilities of conflict in each individual grid cell given the test years from 2013 through 2017. An overview of the two layers of estimations is presented in Table 1.

Estimations
Overveiw

	Target (y)	Features (x/X)	Estimate ($\tilde{y}$)
First layer	y_{cm}	x_{year}	$\tilde{y}_{cm} = f_{cm}(x_{year}) + \epsilon_{cm}$
	y_{cm}	$X_{ll} = x_{long}, x_{lat}$	$\tilde{y}_{sce} = f_{sce}(X_{ll}) + \epsilon_{sce}$
	$f_{sce}(X_{ll})$	x_{year}	$\tilde{y}_{dce} = f_{dce}(x_{year}) + \epsilon_{dce}$
Second layer	$y_{cmBinary}$	$X_{patterns} = f_{cm}(x_{year}), \quad f'_{cm}(x_{year}), \quad f''_{cm}(x_{year}), \quad F_{cm}(x_{year}), \quad f_{dce}(x_{year}), \quad f'_{dce}(x_{year}), \; f''_{dce}(x_{year}), \; F_{dce}(x_{year})$	$\tilde{y}_{prob}$

Table 1: Targets, features and estimates pertaining to the various estimation efforts in the two next sections. For brevity I omit the extrapolations in the table

6.1 First layer: Patterns as features

To construct these eight features, I start by estimating and extrapolating the functions pertaining to conflict magnitude f_{cm}. I then move on to estimate and extrapolate the function pertaining to conflict exposure f_{dce}. Finally I derive the slope, acceleration and mass corresponding to each of these two base functions.

To estimate the function pertaining to conflict magnitude via Gaussian processes I use conflict magnitude as my target y_{cm} and the years in my training set as the sole feature x_{year}. The mean function m_{cm} is simply a constant 0 and the covariance function k_{cm} is the squared exponential function. Mathematically this is expressed in the equations 19, 20, 21 and 22.

$$\tilde{y}_{cm} = f_{cm}(x_{year}) + \epsilon_{cm} \tag{19}$$

$$f_{cm}(x_{year}) \sim \mathcal{GP}_{cm}(m_{cm}(x_{year}), k_{cm}(x_{year}, x'_{year})) \tag{20}$$

$$m_{cm}(x_{year}) = 0 \tag{21}$$

$$k_{cm}(x_{year}, x'_{year}) = \eta^2_{cm} exp\left(-\frac{|x_{year} - x'_{year}|^2}{2\ell^2_{cm}}\right) \tag{22}$$

I first estimate the hyperparameters η_{cm} and ℓ_{cm} along with the error term ϵ_{cm} using time lines with two or more years of conflict. I then introduce $x_{yearNew}$ which is a vector of all years; training

and test alike. Using the estimated hyper parameters and the data contained in the training set I now generate Guassian processes for each time line across all years included in $x_{yearNew}$. The estimated hyperparameters can be found in Table 2 and an illustrative sample of 15 estimated functions and corresponding time lines can be seen in Figure 5.

Hyperparameters
Conflict magnitude

	Point estimate (mean)	Standard deviation	95% Credibility interval
ℓ_{cm}	3.56	0.24	3.08 - 3.99
η_{cm}	1.36	0.04	1.26 - 1.39
ϵ_{cm}	0.95	0.02	0.91 - 0.98

Table 2: The tabel contains the estimated hyperparemeters pertaining to $\tilde{y}_{cm}$.

The most informative entry in Table 2 is the lenghtscale ℓ_{cm}. Given the data and my model specifications there is a 95% probability that ℓ_{cm} lies between 3 and 4 with my point estimate being 3.6. As such, the patterns I extract should be somewhat reliable until three years past my last observation. Thus, given that my training set ends at 2012, I expect my prediction power to drop substantially beyond 2015.

The error term ϵ and the amplitude η are most informative when assessed in tandem. Given the ratio between η and ϵ it appears that my estimated functions are able to explain well over half of the variations in the data. This is an early indication that conflict patterns can indeed be captured. On the other hand it also shows that there is still a lot of variation not captured by my estimated functions.

Turning from the hyperparameters to the estimated functions; Figure 5 shows a sample of 15 randomly drawn time lines and their corresponding functions f_{cm}. These functions have been extrapolated to cover all years; $x_{yearNew}$. As such, the estimated functions continue beyond the training years 1989-2012 into the test years 2013-2017. Thus, what I illustrate with the small sample in Figure 5 is how conflict has waxed and waned in the geographical grid cells over the course of the training years, and how this pattern can be expected to continue into the test years given the data and model specification employed.

The samples in Figure 5 illustrates in practice how the criteria for theoretical coherent (temporal) conflict patterns are fulfilled by using Gaussian processes. Information is drawn from all past included years; the volatility of past conflict impact the magnitude of estimated future conflicts; the influence of past observations decline over time; and the deterioration rate of influence is estimated, as opposed to guessed at.

The functions f_{cm} only capture the temporal patterns. To capture the spatial patterns more explicitly I need f_{dce}. But to get this function I must first estimate a 2D Gaussian process for each

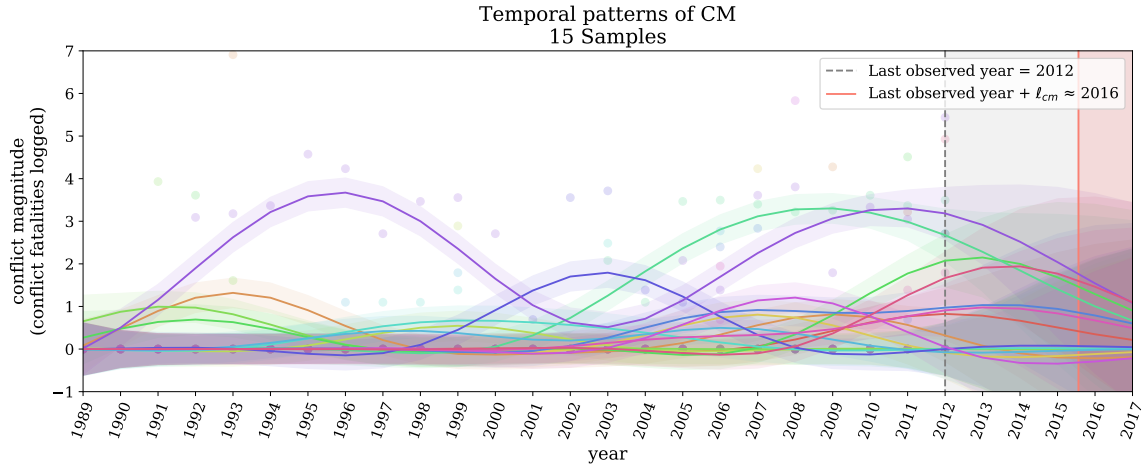

Figure 5: The estimated temporal patterns of conflict magnitude from fifteen randomly drawn time lines. For visualization purposes all fifteen samples has minimum one event over the course of the train years. The solid line is the maximum likelihood point estimate of μ_{cm} and the corrosponding shading denotes $\mu_{cm} \pm 2\sigma_{cm}$. At such, given the model and the data there is roughly a 95% that μ_{cm} is within this zone. The scatter points are the actual observed y_{cm} values from the test set. The estimated hyperparameters are $\ell_{cm} = 3.6$, $\eta_{cm} = 1.1$, $\epsilon_{cm} = 0.9$

separate year in the training set using longitude and latitude as features (x_{lat}, x_{long}: X_{ll})[13] and conflict magnitude (y_{cm}) as target. This is the measure static conflict exposure (sce). It is "static" since I have yet to take time into account – the measure simply tells us how exposed a given cell is to conflict in a given year, without taking into account any temporal patterns. Secondly, I simply estimate a 1d Gaussian process similar to that already implemented to estimate $\tilde{y}_{cm}$. Once more I use x_{year} as the sole feature, but instead of conflict magnitude (y_{cm}) as target, I use the values obtained through f_{sce} as my target y_{dce}. The estimate I derive is one of dynamic conflict exposure $\tilde{y}_{dce}$. Instead of only taking into account the past patterns of the individual cell, this measure takes into account the past patterns of all cells in the relevant proximity.

Starting with the first step I estimate the shared hyperparameters and error term ℓ_{sce}, η_{sce} and ϵ_{sce} for a 2D Gaussian process pertaining to the spatial conflict patterns pertaining to each specific year in the training set. Conflict magnitude is my target y_{cm} and I use longitude x_{long} and latitude x_{lat} as the features X_{ll}. This gives me an estimate of static conflict exposure $\tilde{y}_{sce}$. The mathematical notation can be found in equations 23, 24, 25 and 26.

$$\tilde{y}_{sce} = f_{sce}(X_{ll}) + \epsilon_{sce} \tag{23}$$

$$f_{sce}(X) \sim \mathcal{GP}_{sce}(m_{sce}(X_{ll}), k_{sce}(X_{ll}, X'_{ll})) \tag{24}$$

[13] While X_{ll} actual represents a $2 \times n$ matrix and not vector, I will still keep all mathematical notations as vector notations for simplicity.

$$m_{sce}(X_{ll}) = 0 \tag{25}$$

$$k_{sce}(X_{ll}, X'_{ll}) = \eta^2_{sce} exp\left(-\frac{|X_{ll} - X'_{ll}|^2}{2\ell^2_{sce}}\right) \tag{26}$$

As before, I first estimate the hyperparameters and error term η_{sce}, ℓ_{sce}, ϵ_{sce}. Since I have only estimated the function over spatial dimensions I cannot extrapolate into future years $x_{yearsNew}$ (yet) and I have no intention of extrapolating the patterns into unknown geografical regions so I do not introduce any X_{llNew}. The estimated hyperparameters and error term can be found in Table 3. To visualize the output, the static conflict patterns pertaining to 2012 are plotted in Figure 6.

Hyperparameters
Static conflict exposure

	Point estimate (mean)	Standard deviation	95% Credibility interval
ℓ_{sce}	1.33	0.02	1.26 - 1.39
η_{sce}	0.20	<0.01	0.19 - 0.20
ϵ_{sce}	0.48	<0.01	0.47 - 0.48

Table 3: The tabel contains the estimated hyperparemeters pertaining to $\tilde{y}_{sce}$.

Looking at the hyperparameters, the most informative is once again the lenghtscale ℓ_{sce}. There is a 95% probability that ℓ_{sce} is between 1.2 and 1.4 with my point estimate being 1.3. In substantial terms this tells us that conflict magnitude is relatively stable within a radius of roughly 130 kilometers. To illustrate the actual spatial patterns estimated I have plotted the obtained $\tilde{y}_{sce}$ values pertaining to 2012 in Figure 6. Similar, but unconnected, output is estimated for all years in the training set.

The ratio between η and ϵ tells us that there is a lot of variation not explained by our 2D function. Indeed, with ϵ being twice as large as η it appears my function only captures 1/3 of the total variation in conflict exposure. There are two simple reasons why this is not overly surprising. First of all there are many unknown/unincluded factors separating the different cells from each other. One cell might include a major city, another a great lake. One might be inhabited by a wealthy elite and another by a marginalized and/or poor minority. Secondly, just because a cell did not experience conflict, it does not mean that there was no risk of conflict. What the four features pertaining to the spatial patterns of conflict are meant to capture is not whether or not a given cell experience conflict, but simply whether or not it is exposed to conflict. Thus this ratio between η and ϵ poses no apparent problem.

What I illustrate in Figure 6 is simply an estimated 2D function over the conflict magnitude of all cells in 2012. We see that the spatial patterns estimated conforms neatly to the theoretical

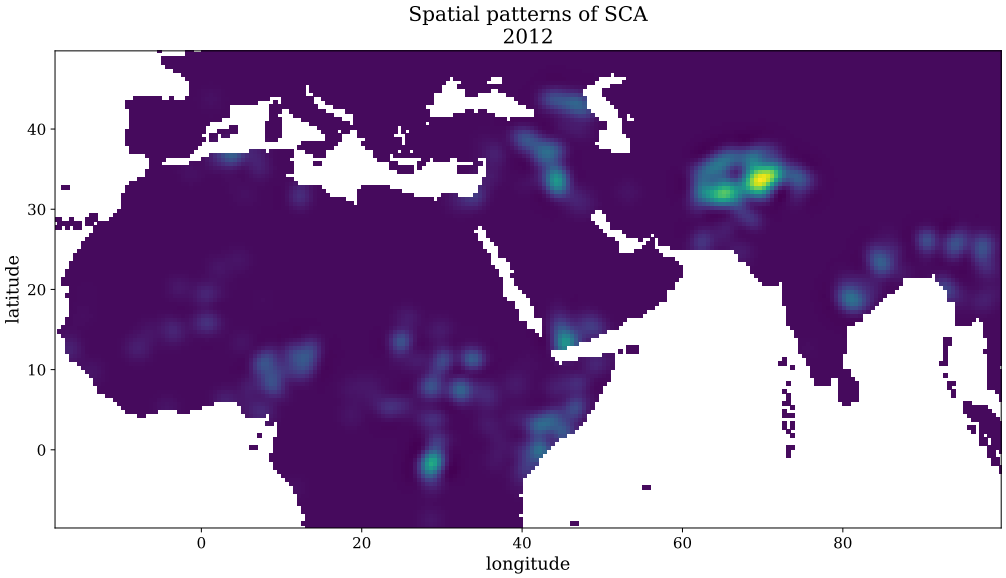

Figure 6: The estimated 2D function $f(X_{ll})_{sce}$ pertaining to the year 2012.

expectation regarding spatial conflict patterns. Information is drawn from all relevant events in the specific year, taking into account both the magnitude and the distance of said events, the deterioration rate of influence is estimated from the available data; conflict diffusion appears fundamentally bell-shaped and radiates out from specific centers. Further more, looking at examples such as Nigeria, Somalia and especially Afghanistan we also see how patterns such as circulation produces different results compared to tangency.

To connected this measure across time and facilitate extrapolation into future years the next step is to estimate $\tilde{y}_{dce}$. To this end, I simply repeat the steps taken to estimate $\tilde{y}_{cm}$. Now, instead of using y_{cm} as target I use the estimated values obtained from $f_{sce}(X_{ll})$ as my target. I denote this estimate $\tilde{y}_{dce}$. Beyond that small change, the procedure is identical. The mathematical specifications are presented in equation 27, 28, 21 and 29.

$$\tilde{y}_{dce} = f_{dce}(x_{year}) + \epsilon_{dce} \tag{27}$$

$$f_{dce}(x_{year}) \sim \mathcal{GP}_{dce}(m_{dce}(x_{year}), k_{dce}(x_{year}, x'_{year})) \tag{28}$$

$$m_{dce}(x_{year}) = 0 \tag{21}$$

$$k_{dce}(x_{year}, x'_{year}) = \eta^2_{dce} exp\left(-\frac{|x_{year} - x'_{year}|^2}{2\ell^2_{dce}}\right) \tag{29}$$

I estimate the relevant hyperparameters η_{dce} and ℓ_{dce} and the error term ϵ_{dce}. The patterns are then extrapolated into the test years using $x_{yearNew}$. I present the estimated hyperparemeters in Table 4. To illustrate the estimated patterns I draw 15 adjacent time lines which are plotted in Figure 7.

Hyperparameters
Dynamic conflict exposure

	Point estimate (mean)	Standard deviation	95% Credibility interval
ℓ_{dce}	3.23	0.13	2.99 - 3.50
η_{dce}	0.59	0.01	0.55 - 0.62
ϵ_{dce}	0.23	0.01	0.22 - 0.23

Table 4: The tabel contains the estimated hyperparemeters pertaining to $\tilde{y}_{dce}$.

We see that ℓ_{dce} has a 95% probability of being between 3 and 3.5 with a point estimate of 3.2. As such, this result is very similar to that of ℓ_{cm}. Again, I only expect my results to be reliable three years into the future. A sample of the estimated patterns is presented in Figure 7.

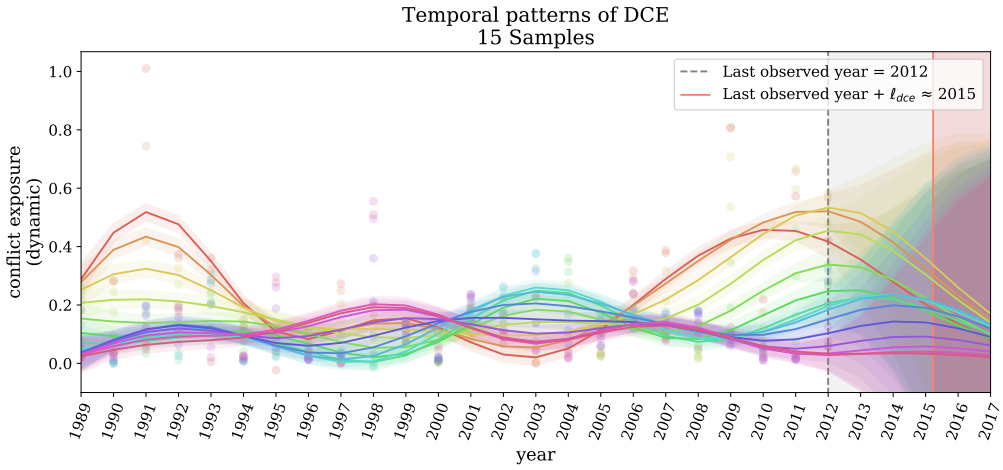

Figure 7: The estimated temporal patterns of conflict exposure from fifteen randomly drawn adjacent time lines. The solid line is the maximum likelihood point estimate of μ_{dce} and the corrosponding shading denotes $\mu_{dce} \pm 2\sigma_{dce}$. At such, given the model there is roughly a 95% that μ_{cde} is within this zone. The colors of the lines denotes how close the time lines are in space. More similar colours equals more proximate cells. The scatter points are the target values obtained through f_{sce}. The hyperparameters are $\ell_{dce} = 3.2$, $\eta_{dce} = 0.59$, $\epsilon_{dce} = 0.2$

By drawing adjacent samples, the spatial dimension of conflict exposure becomes easily discernible. As such, even though I have not explicitly plotted the spatial dimension in Figure 7, it is still clearly visible. Each time line is influenced by all proximate time lines with declining impact

as distance grows. Again, the product aligns gracefully with our theoretical criteria regarding conflict patterns.

I now have my two base measures f_{cm} and f_{dce}. Both these measures are estimated using my training set and have then been extrapolated into the "future" test years. As such, it is now a trivial task to further extract the corresponding slopes (f'_{cm} and f'_{dce}), accelerations (f''_{cm} and f''_{dce}) and masses (F_{cm} and F_{dce}). With this done I have created all eight features which I will use to capture the patterns of conflict.

The last step here is to split the features according to the training and test set. Therefore, all feature values pertaining to the training years from 1998 through 2012 ($X_{patterns}$) are appended to the training set, while the extrapolated feature values from 2013 through 2017 ($X_{patternsEX}$) are appended to the test set. Importantly, the target values in the test set have still not been used for any estimations, computations or model training. As such, the process still emulates a real world forecasting scenario where we are able to extrapolate patterns based on past observations, but not use any information pertaining to future results. With all features constructed and the data-split completed, I now move on to test the combined predictive power of these features using the final predictive framework.

6.2 Second layer: Predicting conflicts

To train the models, I employ the features $X_{patterns}$ and a binary conflict indicator y_{binary} as the target – naturally only the set pertaining to the training years. As such, $X_{patterns}$ contains the values of all eight features from 1989 through 2012 and y_{binary} is a binary indicator denoting whether or not a given cell experienced any conflict in some given year between 1989 and 2012. This constitutes the training data. The test data, which holds the extrapolated $X_{patternsEX}$ and the out-of-sample prediction target y_{binary} values from 2013 through 2017 are kept isolated from the model construction effort and not used before the evaluation effort.

I train 1000 xgboost models using the training data. Every model uses the complete set of events from the training set along with an equally sized set of randomly drawn non-events, also from the training set. Every time a single model is finished training I use the model to generate out-of-sample predictions $\tilde{y}_{pred}$ by introducing the $X_{patternsEX}$ from the test set.

The genereated predictions are expressed as probabilities. Specifically, probabilities that some given grid cell will experience conflict the specific (test) year. As such the full outputs are probability distributions of probabilities. The individual probabilities are adjusted via Bayesian correction to account for the undersampling technique, and a maximum likelihood estimate is obtained from each individual distribution of probabilities. Given that the distributions appears approximately Gaussian, the maximum likelihood estimate is simply the mean and represents my best guess of what the probability of conflict is in a given cell, in a specific year.

To evaluate to what extent my framework, and as such the features I have created, have

successfully captured the patterns of conflict, I will now compare these probabilities to the actual observations contained in the test set y_{binary}. The results will be presented in the next section.

7 The Results

In the preceding sections I have shown how we can estimate and extrapolate the temporal and geo-spatial patterns of past conflicts using modern machine learning techniques. However, I have yet to demonstrate whether these estimated patterns hold any predictive power – whether they can actually by used to forecast the time and place of future conflicts. In the following section I will evaluate the predictive performance of my approach in order to assess the extent to which I have succeeded at reliably predicting the time and place of future conflicts, using the temporal and geo-spatial patterns of past conflicts in tandem with Gaussian processes.

7.1 The predictive potential

To best convey the potential of my framework, I will evaluate it from a number of angles starting with the PR-curve, the AP score, and the ROC-AUC score. Then, I will set a threshold which will be used to convert my predicted probabilities to binary predictions. From these binary predictions I derive the rates of true positives, false positives, true negatives and false negatives as well as the related measures precision and recall. Following this, I will evaluate the relative importance of the eight features used in the framework.

The AP score is the prime evaluation metric of interest. Since I have a distribution of predictions, I naturally also have distributions of results. In Figure 9 I show the distributions of estimated out-of-sample AP scores for each of the test years used with point estimates simply being the mean. As expected the performance of my framework changes as I move into the future. Specifically – and as foretold by the lenghtscales ℓ_{cm} and ℓ_{dce} – the performance drops substantially beyond 3 years. This is easily seen in Figure 8 and Figure 9 as the AP scores fall subtantially after 2015.

Figure 8 shows how the AP score change over the test years from 0.55 in 2013 down to 0.42 in 2017. The orange line is the maximun likelihood estimate of the AP given all the models and blue shade is the individual estimates. This is further illustrated in Figure 9.

As such it is clear that there is a substantial difference between my prediction power before 2015 and after. Indeed, Figure 9 show practically no overlap between 2013, 2014 and 2015 on one side and 2016 and 2017 on the other. Looking at the actual numbers, out of all the models run, no model from 2013, 2014 and 2015 did as bad or worse then any model from 2016 or 2017. This leaves us a probability of $< 0.001\%$ that our framework will perform as well after 3 years as before. For comparison there is a 2.4% probability that $AP_{2015} \geq AP_{2013}$ and a 89.3% probability that $AP_{2015} \geq AP_{2014}$. The point is that the lenghtscale indeed does provide powerful guidance

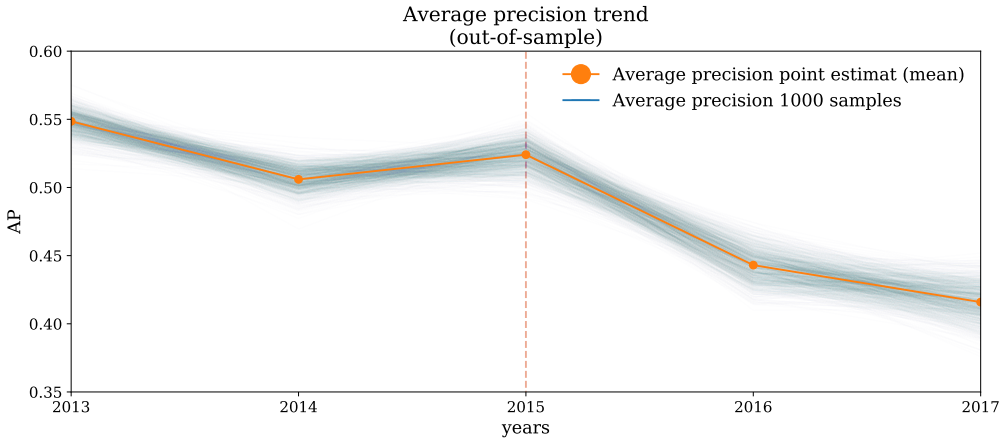

Figure 8: The trend in average precision as it changes over the test years. The orange line denotes the maximum likelihood estimate (mean). The blue shade denotes the actual 1000 estimates per year. The red line denotes that last year $2012 + 3 = 2015$

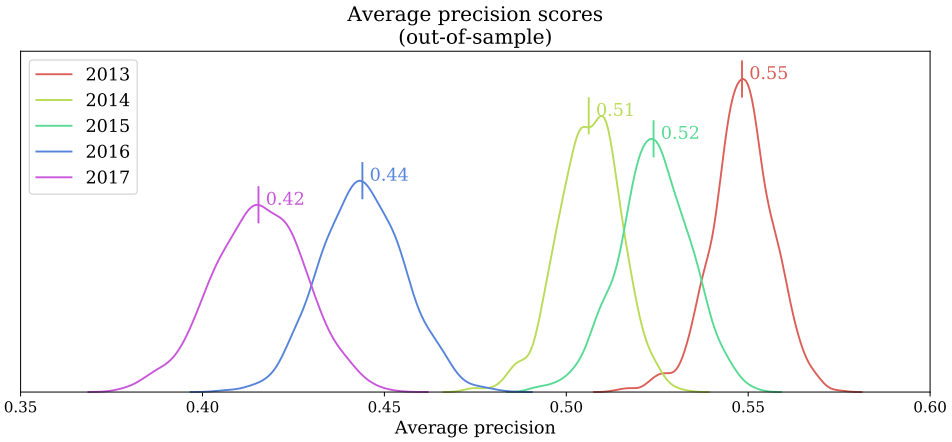

Figure 9: Average precision distributions across all test years. Each distribution is base on the results from 1000 models

regarding how far we can look into the future, and it should be taken quite seriously doing any real world application.

Naturally, it might seem a bit odd that my model seems to do better in 2015 than in 2014. Intuitively the predictive power of the framework should always decrease somewhat when going into the future. As I will comment on below this is neither a flaw nor a mystery. It is simply due to the fact that the patterns of 2015 coincidentally aligns more with the patterns of 2012 – the last train year – than 2014 does. As such, this is not a pattern I would expect to generalize beyond these particular years.

Now, given these results, and to keep the evaluation effort concise, I will not spend more time evaluating the two years 2016 and 2017. To mirror the challenges of real world applications I

also ignore the first test year 2013. After all, the data used to train the models would not have been available before spring/summer 2013, and thus predictions regarding conflicts in 2013 would have been of limited use. Further more, any relevant actor needs some reaction time to take any forecasting into account. As such I focus on the years 2014 and 2015, respectively 6 and 18 months into the future from when the data would first have been available for analysis.

The PR-curve and the AP score are presented in Figure 10 while the ROC-curves and AUC scores are presented in Figure 11.

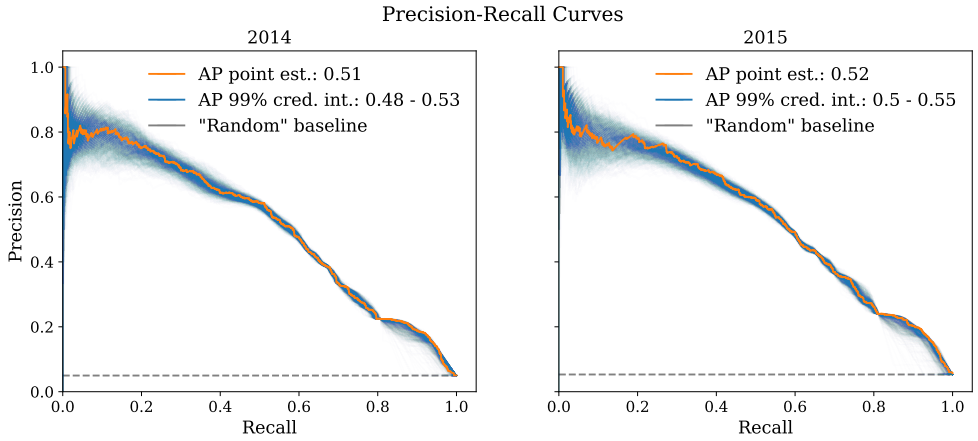

Figure 10: The precision-recall curve for the years 2014 and 2015. The orange line is the maximun likelihood point estimate while the blue lines represents the individual samples

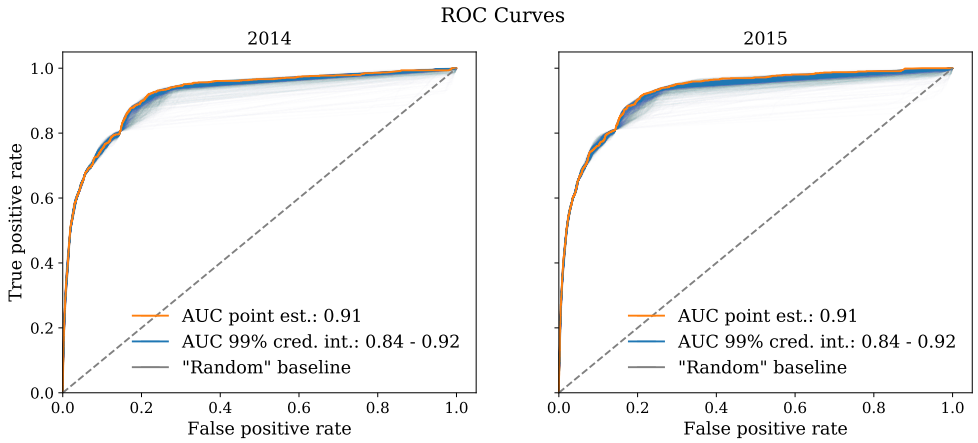

Figure 11: The receiver operating characteristic curve for the years 2014 and 2015. The orange line is the maximun likelihood point estimate while the blue lines represents the individual samples

The PR-curve in Figure 10 illustrates the trade-off between recall and precision at various thresholds, while the ROC-curve in Figure 11 denotes the relationship between the false positive rate and the true positive rate at various thresholds. The share of events in both 2014 and 2015 is

respectively 0.050 and 0.053. As such an AP score around 0.05 and AUC score at 0.5 is equivalent to random guesses. Thus, with an AP scores of 0.51 and 0.52, and AUC scores of 0.91 and 0.91 (point estimates) it is clear that the features I produced exhibit considerable predictive power. The question now is how these results hold up compared to other efforts using different approaches.

Intriguingly, the results presented here are better than those produced in von der Maase (2019). Specifically von der Maase (2019) achieved an AP score of 0.47 (von der Maase, 2019, 14). This is interesting for a number of reasons. First of all, since the results which I have produced for this thesis rely solely on conflict patterns, they reaffirm the conclusion from von der Maase (2019) stating that patterns are more informative than structural features when it comes to predicting future conflicts. Secondly, the framework presented in von der Maase (2019) also included features pertaining to the temporal and spatial patterns of conflicts. These features, however, were vastly less theoretically and methodologically developed compared to the eight features I use in this thesis. Point being, using only the past patterns of conflict, in a theoretically and methodologically coherent manner, here generates better predictions, than using a broad roster of more traditional structural variables in cohort with poorly operationalized patterns.

Beyond the results presented in von der Maase (2019), direct comparison is less forward. My framework produces and AUC of 0.91 for both 2014 and 2015. The state of the art prediction efforts in academia achieve an AUC at approximately 0.80 (Chadefaux, 2017, 14). Importantly however, most of these past studies deal with country-level conflicts and only include novel onsets (Chadefaux, 2017, 14). This naturally makes any direct comparisons with my framework somewhat misleading. On one hand it could be argued that predicting onsets is the most challenging part of conflict prediction. On the other hand, I am both predicting conflict onsets, continuation of conflict and conflict termination; and indeed at a much more disaggregated level than any efforts before this, which should make prediction harder. Point being, while direct comparison is unfeasible, I find it safe to say that the framework I have produced generates predictions at least as good – if not better – than what previous efforts have accomplished. Crucially, I achieve these results in a setting which emulates a real-world scenario. Conversely, most previous efforts have relied on structural data, ignoring the fact that this data would already be dated by the time of release. In a real-world setting this would surely lower the predictive capabilities of frameworks relying on structural features greatly compared to what is presented in published papers.

It should also be noted that when we analyse conflicts on a highly disaggregated level and treat conflict as a function of past temporal and spatial patterns, the concept of a "hard" onset appears somewhat theoretical blurred. It does not make much sense to say that one single cell experienced an actual conflict onset if the conflict simply spread from one or more neighbouring cells – possible only a few kilometers away. Remember these cells are nothing but analytically constructs aligning with no structural or geographical features. And even when conflict diffuses across political boundaries such as country boarders – e.g. the Duran line between Afghanistan and Pakistan – the conflict might well be the same and the notion of a new onset still misleading. Novel onsets do happen but we need to think about the phenomenon differently then we used

to when employing highly disaggregated data. Further more, it should be noted that for policy purposes, predicting both termination and continuation of conflicts is arguably just as relevant as predicting conflict onsets.

Given the challenge of comparison and the fact that measures such as AP or AUC are hard to derive substantial interpretations from, I now move on to present the results in a different light. To better convey the actual real world potential of the framework, I set a hard threshold denoting whether or not I predict that a given cell will experience conflict or not. That is I convert the probabilities into a binary measure. A lot of very useful information is discarded this way, but it does serve to present the results in a more intuitively appealing manner. For real world applications, however, the actual probabilities should naturally always be consulted.

I choose to set a threshold, which insures that I predict roughly the same total number of conflicts as was observed in the last year of the training set (2012). Setting the threshold at 0.10 fulfills this criteria. This means that every cell which has a probability of conflict above 10% is classified as "expected to experience conflict", and all other cells as "not expected to experience conflict". Having actual classifications allows us to assess the generated number of true positives, false positives, true negatives, and false negatives. Given the threshold employed I have plotted these metrics in the two maps presented in Figure 12 and Figure 13.

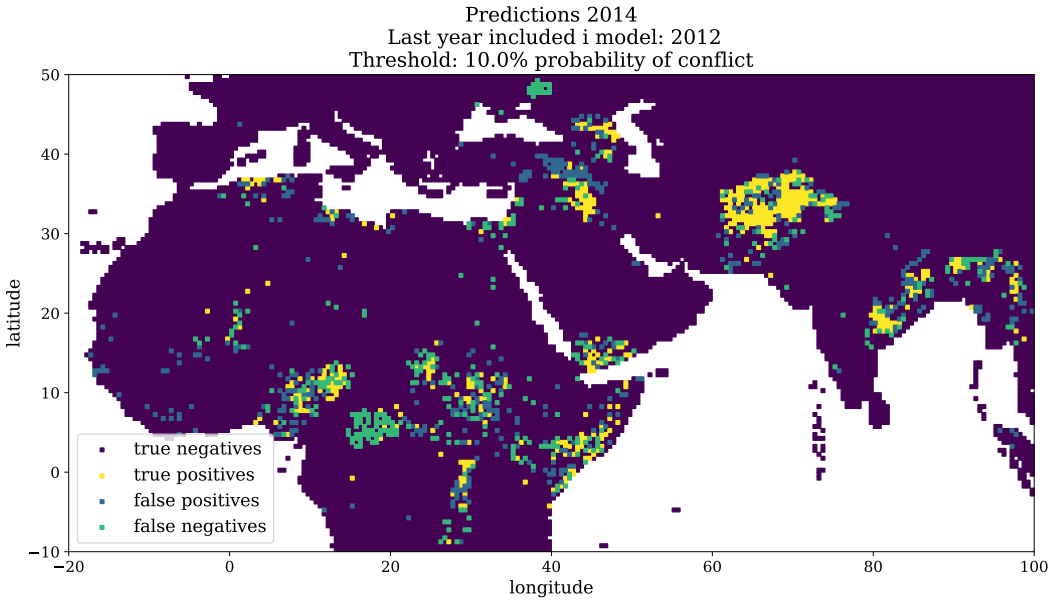

Figure 12: The map is created by binarizing the estimated probabilities of conflict in 2014 and comparing this to the actual binary observations from 2014. Given the threshold of 0.10, all cells with a probability of conflict above 10% are classified as conflicts, while all cells with a probability of 10% or below are classified as non-conflicts.

Interpreting these maps, we get a more substantial evaluation of the performance of the framework. The first thing to notice is that most of the classifications are true negatives. Most cells do not experience conflict most of the time, and my framework captures this. Secondly the framework

is able to generate a large amount of true positives. Remember these predictions are generated on the basis of respectively two and three years old data and each cell is only approximately $50km \times 50km$. Point being, this is an extremely hard classification task – not least considering that I only incorporate past conflict patterns as predictors. As such, the performance of the framework does inspire optimism.

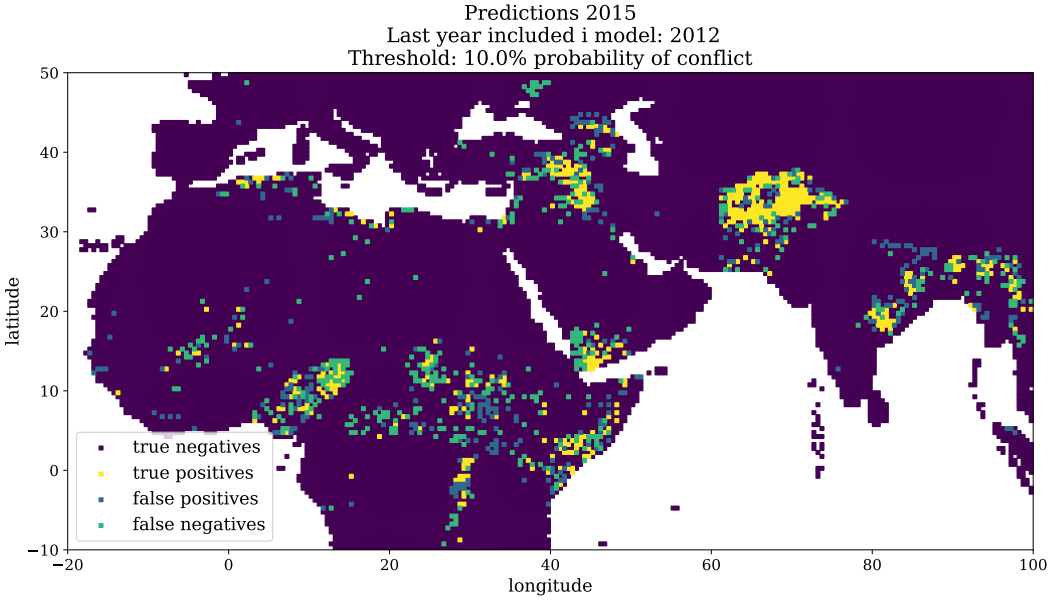

Figure 13: The map is created by binarizing the estimated probabilities of conflict in 2015 and comparing this to the actual binary observations from 2015. Given the threshold of 0.10, all cells with a probability of conflict above 10% are classified as conflicts, while all cells with a probability of 10% or below are classified as non-conflicts.

However, it should also be clear from Figure 12 and Figure 13 that there is still a lot of room for improvement. It is especially clear in 2014 where my framework completely misses two major conflicts. A novel conflict in south-east Ukraine and a great surge of conflict in the Central African Republic. While it is unfortunate that my framework misses these conflicts, it is not surprising. The conflict in Ukraine is far removed in both time and space from any other conflict, thus no past patterns were there to alert the framework. The Central African Republic, on the other hand, has a long history of conflict and is surrounded by other conflict prone regions. As such, my framework also predicts some sporadic conflicts scattered around the country, but not nearly as widespread a conflict as was actually the case. The reason is simply that a peace agreement had been finalized in the last years of my test set, but said agreement fail spectacularly in the beginning of the test years (BBCNews, 2018). As such, the patterns I have extrapolated are those of a declining, not an increase or ongoing, conflict. These two examples clearly show the limits of my framework. Sudden and abrupt changes in the political landscape will always elude a framework solely based on the past spatial and temporal patterns of conflicts.

It is also clear from Figure 12 and Figure 13 that I do generate some false positives. These,

however, are somewhat less problematic than false negatives for a number of reasons. First of all, just because conflict did not manifest in a given cell, it does not mean that the risk of conflict in that particular cell was not real. As an example, a large region from Eastern Turkey over northern Syria, northern Iraq and into western Iran is classified as conflicts in 2014, yet no conflict ensued that year. This region corresponds in large to the geographical region of Kurdistan (Dahlman, 2002, 271-272), which has historically been the sight of many conflicts (Dahlman, 2002). Indeed, we also see conflict erupt in the region again in 2015 where many of my false positives change to true positives. This does not prove that there was a risk of conflict in 2014, but it surely does make it rather plausible. The danger of false positives is that resources might be allocated to prevent conflict when no conflict is likely to happen. However, if the false positives tend to indicate high risk zones where conflict simply failed to erupt due more or less to chance, allocating resources here might not constitute any huge waste after all.

Given the specific threshold I here use, I obtain a recall of respectively 0.51 and 0.57 and a precision of respectively 0.56 and 0.52 for the years 2014 and 2015. Thus, in substantial terms; my framework is able to correctly classify slightly over half of all conflicts cells, but in the process I will roughly produce one false positive for each true positive. Naturally this ratio is highly dependent on the specific threshold I employ, but these results can still do a lot to convey the potential of my framework. Again, given the imbalanced and the highly disaggregated nature of the data and the sole focus on patterns, this is a decent result – but it also does leave a lot of room for improvement. Naturally, such improvement might need to come from the inclusion of more features from other data sources. This paper, after all, only present one piece of the puzzle for creating a complete early warning system.

As such, future research should survey both the potential of including more data and even better handling of past conflict patterns. To inform such efforts the next subsection will briefly touch upon the subject of feature importance. That is, assessing how much prediction power each of my eight features brings to the final predictive framework.

7.2 Feature importance

Regarding feature importance, I find no difference between 2014 and 2015. As such I will only here present the results for 2014. The results from 2015 are vitually indentical and can be found in the appendix. subsection 11.2. The results from 2014 can be seen in Figure 14.

From Figure 14 it is clear that the majority of the prediction power comes from the base feature of conflict magnitude f_{cm}. Interestingly this is the most basic feature in the framework simply denoting the expected future conflict magnitude in some cell given said cells history. It is easily understood and I only use a 1D Gaussian process to derive this measure. As such, future efforts aimed both at prediction and estimation might gain a lot from simply including this measure, if measures of the spatial dimensions are deemed to resource demanding. That being said it is still clear from Figure 14 that all the features do bring some information to the effort. As such I

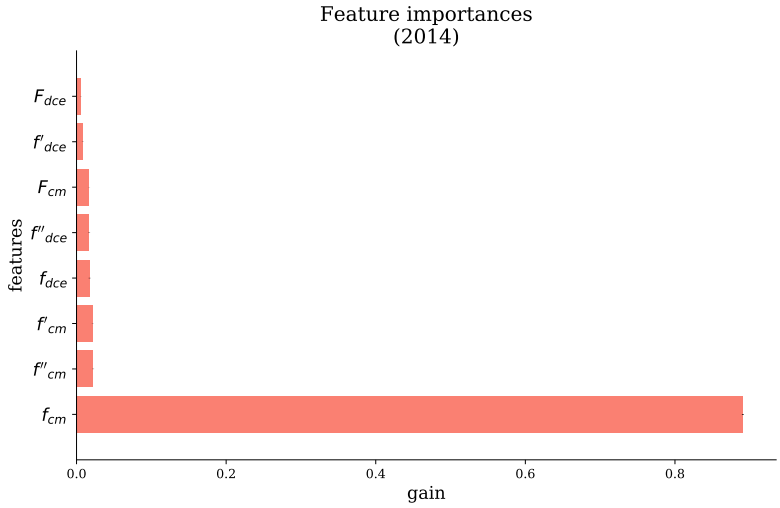

Figure 14: The relative importance of each feature in the prediciton effort

recommend all eight being included in any complete early warning system.

This finalizes my evaluation of the predictive framework I created for these thesis. In the next section I will concluded the entire endeavour by summing up all major insights and results which I have presented throughout this thesis. The conclusion is followed by a prospective look in the future where I discuss how we should proceed from here to create a fully functional early warning system for conflict forecasting.

8 The Conclusion

Previous research has shown that past conflict patterns hold a lot of potential when it comes to predicting future conflicts. Therefore, if we are to create an early warning system for conflict forecasting, such system should include a component focused on extracting prediction power from past temporal and spatial patterns of conflict. However, previous efforts – estimations and predictions alike – have to a large extent failed to create features mirroring the theoretical insights we have regarding conflict patterns. To best utilize the predictive potential of past conflict patterns we need a tool that can capture conflict patterns in a manner which complies with the general theoretical expectations of conflict patterns. In this thesis I illustrate that the machine learning technique of Gaussian processes can be such a tool. Specifically my goal was to assess the extent to which it is possible to use Gaussian processes in tandem with past spatial and temporal conflict patterns to predict the time and place of future conflicts.

To this end I used Gaussian processes to create eight features each pertaining the spatial and temporal patterns of conflicts. Also using Gaussian processes I extrapolated these features into

future years and then used these extrapolations to predict the probability of future conflicts in the geographical grid cells used as unit of analysis.

To asses the extent to which my approach had succeeded in generating reliable forecasts, I used a predictive framework based on xgboost and undersampling along with out-of-sample prediction to evaluate the predictive potential of the extracted and extrapolated conflict patterns. Using my approach to forecast respectively two and three years into the future I achieved AP scores of 0.51 (2014) and 0.52 (2015) against a baseline of roughly 0.05. For both years I achieved an AUC score of 0.91 against a baseline of 0.50. In more substantial terms; if I set a probability threshold at 0.10 I am able to correctly predict half of all conflicts, but doing this will also generate one false positive for each true positive. Naturally I am able to capture more true positives, at the price of relatively more false positive.

Given the highly disaggregated unit of analyses, the great imbalance of the data and the subject at hand these are encouragingly results which definitely shows that past conflict patterns do hold a lot of predictive potential – not least when they are captured in a theoretical and methodological coherent manner.

Unfortunately my results are somewhat hard to compare to previous efforts – not least due the highly disaggregated unit of analysis. The only effort which is directly comparable is von der Maase (2019) which my approach here outperforms despite that fact the von der Maase (2019) includes both simple measures of conflict patterns and structural features. While comparison with other efforts is less forward it is clear that the framework presented here does at least as well, if not better, than the state of the art. Notably this is done using only past patterns, and in a setting which emulates a real world scenario. The later point contrast to many previous efforts where researches ignores the fact that the data their used would be rather dated by the time it is actually made available.

That being said it is also clear that I still have work to do before an actual reliable early warning system can be compiled. specifically my approach are unable to identify truly novel conflict onsets far removed from any past conflicts in time and space. That being said this thesis do present a substantial step forward, towards the creation of a actual early-warning-system for conflict forecasting.

As presented, using the patterns of past conflicts as a predictor only constitute one of several components needed to create a complete early warning system. Naturally, there are no final answers concerning how many or which specific components are needed to create a fully operational early warning system. As such, in the next and final section I will present a number of components which might ammend the weaknesses of the approach presented above.

9 Futher Perspectives

Using past conflict patterns as predictors does result in some natural limitations. Most prominently, we will never be able to predict truly novel conflicts far removed in time and space from any previous conflicts using such approach alone. We might be able to achieve even better results with Gaussian processes than what I have presented above. This might be done by distinguishing between long term trends and short term trends or by using hierarchical hyper priors. Furthermore other even more involved methods such as artificial neural networks might be able to capture the patterns of conflict in even more effective ways. But while these opportunities should be explored in future research, they will never amend the fundamental limitation of using past patterns to predict future patterns. As such, while patterns should be a central component in any early warning system, they should also only be one of many.

One component which also deserves attention is composed of structural data and theory on the structural prerequisites of intra-state conflict, such as poverty, resources, regime ect. The advantages of such a component is the large catalog of theory following it (Collier and Hoeffler, 1998; Fearon and Laitin, 2003; Collier and Hoeffler, 2004; Hegre and Sambanis, 2006; Kalyvas, 2007; Goldstone et al., 2010; Cederman et al., 2013; Perry, 2013). Such insights inform feature engineering, feature selection and model selection (Cederman et al., 2013, 30).

As noted, however, one clear disadvantage of this approach is dated data. An other disadvantage is high inertia (Chadefaux, 2017, 10). Structural features are slow to change while political unrest can evolve into actual intra-state conflict relatively fast. Intriguingly however, given the this inertia we can address the disadvantage of dated data simply by use Gaussian processes to extrapolate the structural patterns into future years. As such, I would exploit the inertia of structural patterns to ameliorate the dated state of the data. To a large extent this effort would mimic the approach presented in this thesis, simply using different data.

Yet, while the inertia inherent in structural data might justify extrapolating these patterns across five or even ten years, the inertia still means that we might identify fragile regions but not necessarily which of these fragile regions will experience conflict (Chadefaux, 2017, 10). Combining a structural component with one of these patterns will reduce this problem especially if we include data on other forms of events such as civil or political unrest.

Following this logic, another component should be one which uses text or image data obtained from news/social/political sources. Such data would effectively constitute its own kind of event data, but one specifically tailored for the task at hand. Using text data is still a novel approach in social science (Grimmer and Stewart, 2013) and image data even more so (Williams et al., 2019). Yet, the use of text data has already shown promise in conflict predictions (Chadefaux, 2014; Mueller and Rauh, 2016). Using image data for conflict prediction has to my knowledge not been done yet, but it could hold great promise not least since it circumvents the problem of varying norms and languages which challenges the use of text data.

Text and image data has two advantages on structural data. Firstly, the data has the potential to be very current. In theory, it could be updated in real-time (Cederman and Weidmann, 2017, 474). Secondly, like more traditional event data, text and image data would be able to sort fragile but peaceful regions from fragile regions on the verge of conflict. Importantly, such data would also holds an advantage over the event data used in this thesis; text and image data could help predict onsets far removed in time and space from any previous conflicts.

Naturally, using text or images also has its weaknesses. Gathering text or images from the entire globe is no trivial effort. Local sources is hard to come by, and legal barriers might prevent the use of any data obtained. As such, it is not realistic to produce data on a disaggregated level comparable with that of the other components which I have presented. Thus, to ameliorate the individual weaknesses of each component they must all be combined into a single unified predictive framework. A unified early warning system for conflict prediction.

With these paths for future research presented I conclude this thesis on an optimistic note. Reliable conflict prediction might be the conflict researcher's final frontier, but given the results presented above, it does by no means appear a fruitless nor impossible endeavour.

10 Bibliography

References

Bara, C. (2018). Legacies of violence: Conflict-specific capital and the postconflict diffusion of civil war. *Journal of Conflict Resolution*, 62(9):1991–2016.

BBCNews (2018). Central african republic profile - timeline. Accessed 22-09-2019.

Bestgen, Y. (2015). Exact expected average precision of the random baseline for system evaluation. *The Prague Bulletin of Mathematical Linguistics*, 103(1):131–138.

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.

Buhaug, H. and Gleditsch, K. S. (2008). Contagion or confusion? why conflicts cluster in space. *International Studies Quarterly*, 52(2):215–233.

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. and Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, 355(6324):474–476.

Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, 51(1):5–18.

Chadefaux, T. (2017). Conflict forecasting and its limits. *Data Science*, 1(1-2):7–17.

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.

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*.

Crost, B., Felter, J., et al. (2015). Is conflict contagious? evidence from a natural experiment. Technical report, Households in Conflict Network.

Dahlman, C. (2002). The political geography of kurdistan. *Eurasian Geography and Economics*, 43(4):271–299.

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:.

Gelfand, A. E. and Schliep, E. M. (2016). Spatial statistics and gaussian processes: A beautiful marriage. *Spatial Statistics*, 18:86–104.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis - third edition*. Chapman and Hall/CRC.

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.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

He, H. and Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 9:1263–1284.

Hegre, H., Nygård, H. M., Karlsen, J., Strand, H., and Urdal, H. (2013). Predicting Armed Conflict, 2010–2050. *International Studies Quarterly*, 57(2):250–270.

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

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.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

King, G. and Zeng, L. (2001a). Explaining rare events in international relations. *International Organization*, 55(3):693–715.

King, G. and Zeng, L. (2001b). Improving forecasts of state failure. *World Politics*, 53(4):623–658.

McElreath, R. (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Mueller, H. F. and Rauh, C. (2016). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 2(112):358–375.

O'Loughlin, J., Witmer, F. D., and Linke, A. M. (2010). The afghanistan-pakistan wars, 2008-2009: Micro-geographies, conflict diffusion, and clusters of violence. *Eurasian Geography and Economics*, 51(4):437–471.

Perry, C. (2013). Machine learning and conflict prediction: a use case. *Stability: International Journal of Security & Development*, 56(2(3)).

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

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

Su, W., Yuan, Y., and Zhu, M. (2015). A relationship between the average precision and the area under the roc curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 349–352. ACM.

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

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

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

von der Maase, S. P. (2019). A modern computational approach to conflict prediction. *Unpublished Exam Paper, Free Assigment 12.5 ECTS, UCPH*.

Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.

Weidmann, N. B. and Ward, M. D. (2010). Predicting conflict in space and time. *Journal of Conflict Resolution*, 54(6):883–901.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA.

Williams, N. W., Casa, A., and Wilkerson, J. D. (2019). An introduction to images as data for social science research.

11 Appendix

11.1 Python scripts

The following is a list of all scripts used for this thesis. They are all written for python 3 in jupyter notebooks (.ipynb). the list is ordered by order of execution. All the notebooks can also be accessed via my github: https://github.com/Polichinel/Master_Thesis. For simplicity they are all contained in the .zip file thesis_apps_simon_vonderMaase.zip

Note that it will take days if not weeks to execute all notebooks even on a high-end PS.

Order of execution:

1. GP_presentation.ipynb
2. Pickle_full_data.ipynb
3. Pickle_restricted_data.ipynb
4. Pickle_test_train.ipynb
5. Pickle_hp_sce_all_years.ipynb
6. Pickle_hp_dce_mu_and_var.ipynb
7. Pickle_dce_mu_predictions.ipynb
8. Pickle_hp_cm_mu.ipynb
9. Pickle_cm_mu_predictions.ipynb
10. Pickle_slope_acc_mass.ipynb
11. Plot_hp_n_samples.ipynb
12. Pickle_predictions_n_metrics.ipynb
13. Plot_results.ipynb

The Prio grid shape file can be found at https://www.prio.org/Data/PRIO-GRID/ or directly from file.prio.no/ReplicationData/PRIO-GRID/priogrid_shapefiles.zip

The UCDP shape file can be found at https://www.ucdp.uu.se/downloads/#d1 or directly from http://ucdp.uu.se/downloads/ged/ged181-shp.zip

The paper von der Maase (2019) can also be found at my github page:

https://github.com/Polichinel/Conflict_Prediction

11.2 Feature importance 2015

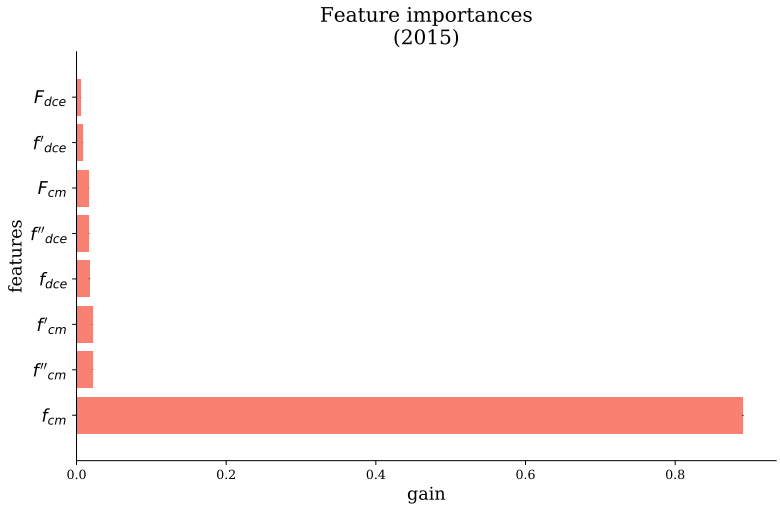

Figure 15: The relative importance of each feature in the prediciton effort