

# **Machine learning og borgerkrig**

**Prædiktion af borgerkrige og voldelige intrastatslige  
uroligheder**

Frederik Tom-Petersen \*

December 29, 2018

## 0.1. Abstract

In this thesis, I investigate the predictive potential of machine learning on civil war. Comparing the Random Forest and Extreme Gradient Boosted Tree algorithms with simple baseline models I have sought to predict the outbreak of civil war, independent instances of internal conflict and the intensity of a conflict. Furthermore, I introduce a new dataset on the political landscape based on the auto-coded GDELT database.

The analytic framework builds a bridge between competitive theoretic motives and preconditions for civil war in a synthesis, that focusses on the intersection of the theories and the independent value of each included variable. I find that the machine learning algorithm *Extreme Gradient Boosted Trees* outperforms the introduced *Random Forest* algorithm and the baseline model enabling the correct prediction of 50% of the civil war outbreaks. Furthermore, the analysis shows that data on political events serve as a good predictor, by utilizing information on the political landscapes that can be derived from the development of events.

Predicting civil war is without much meaning if the prediction model cannot be implemented in an appropriate environment. For an easy cost-efficient implementation, I introduce my analysis as a full-automatic process based on publicly available data and open-source software. The implementation is however also limited by the inherent structure of the international system and society and their tools for conflict prevention and de-escalation.

**Keywords:** Machine learning, prediction, civil war, insurgencies, big data, political events

## INDHOLDFORTEGNELSE

0.1. Abstract . . . . .	2
0.2. Forord . . . . .	6
<b>1. Indledning</b>	<b>7</b>
1.1. Forskningsdesign . . . . .	8
1.2. Casevalg og begrebsafklaring: Borgerkrig . . . . .	9
1.3. inferens vs. prædiktion . . . . .	11
1.4. Specialet opbygning . . . . .	11
<b>2. Litteratur-review</b>	<b>13</b>
2.1. Fra konkret risikovurdering til generisk model . . . . .	13
2.2. Prædiktion af borgerkrige . . . . .	14
2.3. Kritikken og centrale indsigter . . . . .	16
<b>3. Teoretiske perspektiver på borgerkrig</b>	<b>17</b>
3.1. Vækst, velfærd & borgerkrig . . . . .	18
3.1.1. Vækst og konflikt nexuset . . . . .	18
3.1.2. Betydning af institutioner for vækst og fordeling . . . . .	20
3.1.3. Ressourceforbandelsen . . . . .	21
3.1.4. Geografiens betydning for vækst . . . . .	24
3.2. Muligheden for oprør og det økonomisk motiverede oprør . . . . .	24
3.2.1. Finansiering af oprør . . . . .	26
3.2.2. Rekrutteringspotentialer . . . . .	28
3.2.3. Omkostningsreducerende geografiske fordele . . . . .	28
3.3. Gruppeidentitet: Politisk og økonomisk utilfredshed . . . . .	29
3.3.1. Individuel politisk- og økonomisk utilfredshed . . . . .	30
3.3.2. Polarisering og Horisontale ulighed . . . . .	31
3.4. Samspillet mellem de teoretiske tilgange . . . . .	33
<b>4. Machine learning og prædiktionsmodeller</b>	<b>34</b>
4.1. Machine learning og algoritmer . . . . .	34
4.1.1. Prædiktionsmodeller i praksis . . . . .	34
4.1.2. Prædiktionsmodeller som funktion . . . . .	35
4.1.3. Bias-varians trade-off og krydsvalidering . . . . .	36
4.2. Evaluering af prædiktionssevner . . . . .	38
4.3. Valg af algoritmer . . . . .	42
4.3.1. Klassifikationstræer . . . . .	43
4.3.2. Random Forest . . . . .	44
4.4. Extreme Gradient Boosted Trees . . . . .	45
<b>5. Data</b>	<b>47</b>
5.1. Responsvariablene . . . . .	47
5.2. Det politiske landskab: Politisk tilfredshed og utilfredshed . . . . .	49
5.2.1. De konstruerede variable . . . . .	51
5.3. Validiteten af nyhedsdata som datakilde . . . . .	52
5.4. Behandling af manglende data . . . . .	53

<b>6. Analyse</b>	<b>55</b>
6.1. Modellernes performance . . . . .	55
6.1.1. Udbrud af borgerkrig . . . . .	56
6.1.2. Eksistensen af borgerkrigs-lignede tilstande . . . . .	60
6.2. Intensiteten af konflikter . . . . .	62
6.2.1. Antallet af døde . . . . .	62
6.2.2. Antallet af kamphandlinger . . . . .	64
6.2.3. Sammenhængen mellem resultaterne . . . . .	65
6.3. Modellernes validitet og resultaternes pålidelighed . . . . .	67
6.4. Den bedste model . . . . .	68
6.4.1. De vigtigste variable i prædiktionsmodellerne . . . . .	68
6.4.2. Tuning af modellen . . . . .	73
<b>7. Diskussion</b>	<b>78</b>
7.1. Resultaternes anvendelse og implementering af prædiktionsmodellen . . . . .	78
7.2. Teoretiske implikationer . . . . .	81
<b>8. Konklusion</b>	<b>84</b>
<b>Litteraturliste</b>	<b>87</b>
<b>A. Appendiks</b>	<b>i</b>
<b>B. Dannelsen af Global Database of Events, Language and Tone datasættet</b>	<b>i</b>
B.1. Udvælgelse af hentning af nyhedsdata . . . . .	i
B.2. Efterbehandling og Geokodning . . . . .	i
B.3. Tekstanalyse og dannelsen af observationer . . . . .	ii
<b>C. Gruppering af gdelte data</b>	<b>iv</b>
<b>D. GDELT begivenhedstyper</b>	<b>v</b>
<b>E. Kodning af analyse i R</b>	<b>xi</b>
E.1. Analyse: Udbrud af borgerkrig . . . . .	xi
E.2. Analyse: Borgerkrigs-lignende tilstande . . . . .	xiv
E.3. Analyse: Intensiteten af borgerkrig . . . . .	xviii
E.3.1. Antal Døde . . . . .	xviii
E.3.2. Antal kamphandlinger . . . . .	xix

## LISTE OVER FIGURER

1.	Kort over interne tilfælde af kamphandlinger 1989-2017 . . . . .	10
2.	Oversigt over ressourceforbandelsen dimensioner . . . . .	22
3.	Hovedkomponenterne i greed-tilgangen der tilsammen udgør mulighederne for oprør . . . . .	25
4.	Eksempel på to-dimensionel polarisering . . . . .	31
5.	Opdeling af data i træningssæt og testsæt . . . . .	35
6.	illustration af fleksibilitetens betydning for fejl . . . . .	37
7.	Densitetshistogram over prædiktionsstyper - opstillet eksempel . . . . .	40
8.	Opstillet eksempel på ROC-kurve . . . . .	41
9.	Logikken bag segmenteringen i klassifikationstræer . . . . .	43
10.	Oversigt over borgerkrige og borgerkrigs-lignende uroligheder . . . . .	49
11.	ROC-kurver: Udbrud af borgerkrig . . . . .	57
12.	Densitets histogrammer: Udbrud af borgerkrig . . . . .	58
13.	ROC-kurver: Eksistensen af borgerkrig . . . . .	60
14.	Densitets histogrammer: Eksistensen af borgerkrig . . . . .	61
15.	Prædiktio n af antal døde . . . . .	63
16.	Forudsagte antal kamphandlinger mod faktiske antal kamphandlinger . . . . .	65
17.	Gennemsnitlige <i>variable importance</i> . . . . .	70
18.	Gennemsnitlige <i>variable importance</i> - uden konfliktvariable . . . . .	71
19.	Indledende tuning af XGBtree modellen for udbruddet af borgerkrig . . . . .	74
20.	Antallet af iterationers betydning for AUC . . . . .	75

## LISTE OVER TABELLER

1.	Gennemsnitlige BNP pr capita og økonomisk vækst . . . . .	19
2.	Confusion-matrix. . . . .	39
3.	Oversigt over data . . . . .	54
4.	Udbrud af borgerkrig: Optimale værdier for $\delta$ og omkostning . . . . .	56
5.	Falske positive i EGBT-algoritmen med høj sandsynlighed af borgerkrig . . . . .	59
6.	Eksistensen af borgerkrig: Optimale værdier for $\delta$ og omkostning . . . . .	60
7.	Sammenligning af modellernes performance: Antal døde . . . . .	63
8.	Sammenligning af modellernes performance: Kamphandlinger . . . . .	64
9.	Sammenligning af modellernes performance . . . . .	66
10.	Variablenes importance vurderet efter deres rang . . . . .	72
11.	EGBtree-modellen med systematisk og tilfældig tuning . . . . .	76
12.	Confusion-matrix . . . . .	77

### 0.2. Forord

Jeg har valgt at skrive dette speciale ud fra et ønske om at afsøge den *hype*, der omkring *big data*, *machine learning* & prædiktionsproblemer indenfor samfundsvidenskaben. Applikationen af *machine learning* til at løse prædiktionsproblemer indenfor samfundsvidenskaben har mange facetter og kan bidrage til mange forskellige områder. Jeg har valgt at fokusere på borgerkrig, da det i min optik er et område, hvor løsningen af det prædiktionsproblem som borgerkrige kan anskues som, kan gøre en reel forskel for de berørte menneskers liv og de samfund de lever i.

I udarbejdelsen af specialet skylder jeg først og fremmest tak til Jesper Svejgaard Jensen for at have udbudt seminarkurset *Political Data Science*, der i foråret 2018 introducerede mig til kodning, datascience og *machine learning* og som pirrede min nysgerrighed for at kaste mig ud i projektet. Sidenhen har Kristian Urup Larsen og Jakob Jul Elben været en enorm metodisk støtte. Til dem skal der lyde et stort tak for at de løbende har stillet deres viden til rådighed, givet kommentarer på projektet og deres tekniske sparring i forhold til særligt kodningen af projektet. Christoffer Cappelen skal have tak for at introducere mig til borgerkrigsfeltet og de diskussioner vi har haft om de teoretiske tilgange til borgerkrig. Sidst men ikke mindst fortjener min vejleder, Uffe Jakobsen, en stor ros for kyndig vejledning gennem hele forløbet.

Specialets analyse er bygget op omkring open-source programmeringssprogets R. I dettes ånd vil jeg efter specialets bedømmelse stille al kodning, der ligger til grund for specialet til rådighed på [www.Github.com/FrederikTomPetersen](https://www.Github.com/FrederikTomPetersen)

---

## 1. INDLEDNING

Samfundsvidenskaben har historisk beskæftiget sig med at etablere forklaringer på fænomener frem for at forsøge at forudsige dem (King et al., 1995a; Breiman, 2001). Dette har haft betydning for de forskningsspørgsmål, der formuleres i samfundsvidenskaben. Her har de spørgsmål der stilles, koncentreret sig om at forklare et *outcome* på baggrund af en række variable med det formål at drive inferens (Bøgh Andersen et al., 2010, s. 72-73). Denne tendens er langsomt ved at vende grundet en erkendelse af, at mange af de samfundsvidenskabelige spørgsmål, kan ses som prædiktionsproblemer. Her er målet ikke at forklare *outcome*, men derimod at prædiktere det. I løsningen af de samfundsvidenskabelige prædiktionsproblemer står *machine learning* som metode centralt.

Et af de områder hvor forskningsspørgsmålene i stigende grad stilles som prædiktionsproblemer og *machine learning* anskues som løsningsmodellen er i det internationale samfund. Her er prædiktionen af afstemningsresultater, effekten af politiske tiltag eller flygtningestrømmes størrelse blandt andet er genstand for opmærksomheden (United Nations Global Pulse, 2018). Her har præcise og meningsfulde prædiktioner førhen været set som et fatamorgana; et billede der er blevet ændret med voksende datamængder & udbredelsen af og kendskab til *machine learning* i samfundsvidenskaben. *Machine learning* har så og sige revitaliseret prædiktioner ved at gøre forskere i stand til at finde komplekse mønstre i data, der ikke har kunnet opnås gennem klassiske statistiske metode (James et al., 2013, s. 6).

At anskue samfundsvidenskabelige spørgsmål som prædiktionsproblemer er ikke bare nyttigt, men også kærkomment. For hvad nytter samfundsvidenskaben egentlig, hvis man ikke er i stand til at prædiktere skelsættende politiske begivenheder, som for eksempel udbruddet af borgerkrig? Netop dette spørgsmål er omdrejningspunktet i dette speciale, hvor jeg gennem en afsøgning i borgerkrigslitteraturen vil opstille en række superviserede prædiktionsmodeller, der forsøger at besvare problemstillingen:

Hvilket potentiale rummer machine learning i tilretningen af tiltag mod borgerkrig, hvis man anskuer dette som et prædiktionsproblem?

Hovedelementet i besvarelsen af denne problemstilling er selvsagt et fokus på den metodiske afprøvning af *machine learning* algoritmernes potentiale for korrekt at prædiktere forhold omkring borgerkrig. Dette potentiale kan ikke udelukkende evalueres metodisk. Evalueringen må også læne sig op at den potentielle praksisanvendelse, som prædiktionerne kan have i det internationale samfund i form af politiske-tiltag såsom nødhjælp, fredsmægling og interventioner.

I *machine learning* og prædiktionsmodeller står *dataet* centralt. Her vil jeg som et

bidrag til borgerkrigslitteraturen introducere et nye datasæt. Datasættet bygger på 250.000.000 nyhedsartikler omhandlende politiske begivenheder og formålet er få et konkret billede på det *politiske landskab* i alle verdens lande måned for måned. Håbet er, at dette nye datasæt kan bidrage til besvarelsen af problemstillingen ved at skabe mere præcise prædiktioner, og samtidig nuancere det teoretiske billede, der er på borgerkrig i litteraturen.

I min besvarelse af problemstillingen vil jeg nærmere bestemt afprøve potentialet gennem 2 *machine learning* algoritmer på fire forskellige mål for borgerkrig, der kunne have en praksis relevans: (1) *udbruddet af borgerkrig*, (2) *eksistensen af borgerkrigslignende tilstande* (3) *antallet af døde* (4) *antallet af kamphandlinger*. De 2 *machine learning* algoritmer jeg vil introducere er henholdsvis *Random Forest* & *Extreme Gradient Boosted Trees* algoritmerne. Modellerne vil jeg evaluere op mod baseline modeller der bygger på simple regressionsmodeller.

### 1.1. Forskningsdesign

Centralt i al forskning står forskningsdesignet, der bør ses som et generisk og essentielt element. Dette speciale er ingen undtagelse. Forskningsdesignet har ikke kun betydning for valget af teori og metodik, men tjener også som den røde tråd, der skal sikre at undersøgelsen faktisk belyser den problemstilling, der er opstillet og evaluere resultaterne på en meningsfuld måde (Bøgh Andersen et al., 2010, s. 66-68)

I King, Keohane og Verbas artikel "The Importance of Research Design in Political Science" (1995) understreger forfatterne vigtigheden af et velbegrundet og belyst forskningsdesign i den samfundsvidenskabelige forskning. Forskningsdesignet bør ses som den ramme, der sikrer at man undersøger det der er formålet og er eksplicit omkring de til-og fravalg man løbende fortager, samt en åbenhed omkring de begrænsninger der er ved forskningsdesignets besvarelse af problemstillingen. Uagtet om man ser forskningsdesignet som en iterativ proces givet ved *the wheel of science* (Bøgh Andersen et al., 2010, s. 67) eller som et statisk element, kan man i etableringen af forskningsdesignet stille sig selv tre vejledende spørgsmål:

1. Hvad vil du finde ud af?
2. Hvordan vil du finde ud af det?
3. Hvordan kan du vide om du har ret eller om du er forkert på den?

Som allerede præsenteret er formålet med specialet at afsøge *machine learnings* potentiale til at prædiktere emner indenfor borgerkrig. Selve genstandsfeltet for specialet er således borgerkrige og borgerkrigslignende tilstande. Borgerkrige er interessante at studere da de er et eksistentielt problem i det enkelte land, såvel som i det internationale



samfund, grundet eksternaliteter såsom flygtningestrømme, regionale uroligheder og afledte spørgsmål om regeringsmagts suverænitet. Borgerkrige er derfor et interessant og betydningsfuldt fænomen i sin essens.

Teorien er det andet centrale element i et forskningsdesign (Bøgh Andersen et al., 2010, s. 69). Dette speciales teoretiske fundament kan opdeles i to distinkte områder. Det første område udspringer af litteraturens søgen efter svar på spørgsmålet *Hvorfor opstår borgerkrig?* Som svar på dette spørgsmål er der i litteraturen dannet et stærkt teoretisk fundament, der skriver og belyser feltets kausalhistorier, forestillinger og forventninger. Specialets andet teoretiske ben er at finde i selve metodikken. For hvorfor er det egentlig, at det forventes at *machine learning* algoritmer er i stand til at prædiktere bedre end traditionel statistisk læring?

Sidst men ikke mindst står datagrundlaget tilbage, som det bærende element for undersøgelsen. Datagrundlaget for dette speciale er i høj grad guidet af de teoretiske perspektiver på borgerkrig. Dataet i sig selv giver ingen mening uden en forståelse for borgerkrig. En uddybning af borgerkrig som case er derfor formålet med følgende afsnit.

### 1.2. Casevalg og begrebsafklaring: Borgerkrig

Hvorfor borgerkrige overhovedet studeres? Svaret er simpelt. Borgerkrige kan, uafhængigt af tid og sted, anses som den type af politiske begivenheder, der har allerstørst indvirkning på de berørte individer, samfund og stater. For individet er borgerkrig forbundet med en fundamental eksistentiel usikkerhed, mens det for samfundet betyder et ophør af normale tilstande, og samfundsaktiviteten vil være afledt af konfliktens karakter. Indenfor politologien anses staten i sin grundessens som garant for borgernes sikkerhed, gennem den samfundskontrakt, der tegnes mellem borgere og stater (Hviid Pedersen, 2015, s.123). Krige og borgerkrige udgør dermed en eksistentiel trussel på staten som enhed, da det er en udfordring af dennes mest grundlæggende funktion.

Krige og konflikter kan have forskellige kendetegn i deres udformning, både hvad angår krigens parter<sup>1</sup>, krigens genstandsfelt<sup>2</sup>, og måden hvorpå der føres krig<sup>3</sup>. I dette speciale vil jeg se nærmere på den gruppe af krige, der bredt benævnes borgerkrige. Den klassiske ordbogsdefinition på en borgerkrig, er en konflikt mellem to organiserede parter indenfor et land, hvor den ene part er staten og hvor der kæmpes om den politiske magt i et land (McLean and McMillan, 2009). Selvom borgerkrige i sin essens er afgrænset til en stat, har borgerkrige ofte internationale dimensioner i form af

---

<sup>1</sup>Eksempelvis traditionelle hære, løst organiserede oprørsgrupper, transnationale grupperinger

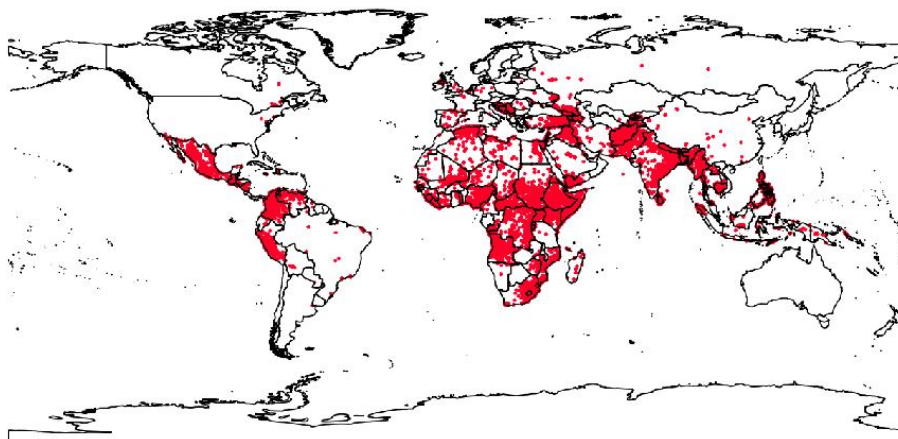
<sup>2</sup>Eksempelvis om der kæmpes over naturressourcer, territorier, den politiske magt

<sup>3</sup>Eksempelvis traditionel interstatslig krig, terrorisme, guerillakrig

flygtningestrømme, internationale interventioner, humanitærhjælp eller ved at krigens partere opererer fra flere lande.

Konflikter der har denne interne karakter er genstandsfeltet for dette speciale. Som det ses af kort 1 er dette fænomen hverken smalt afgrænset ej heller lille i sit omfang. Interne konflikter har i perioden 1989-2017 fundet sted på stort set alle kontinenter og i et omfang, hvor der årligt dør 68.243 i gennemsnit som direkte følge af konflikterne. Borgerkrige udgjorde for perioden 1945-1999 83% af alle krige og resulterede i omtrent 16,2 millioner borgerkrigsrelaterede døde (Fearon and Laitin, 2003, s. 75). Borgerkrige som sybtype af konflikter fortjener derfor et særskilt fokus.

**Figur 1:** Kort over interne tilfælde af kamphandlinger 1989-2017



**Figur 1:** Dataet fra Uppsala Conflict Data Program (Croicu and Sundberg, 2017; Sundberg and Melander, 2013), illustrerer enkeltstående tilfælde af interne konflikter i perioden 1989-2017

Hvis man anskuer borgerkrige ud fra et politisk teoretisk perspektiv, ligger der i borgerkrig et paradoks grundet den samfundskontrakt, der er tegnet mellem stat og borgere. Staten der er givet magt til at beskytte sine borgere, bekriger selvsamme og borgere gør vold mod deres egen stat. Paradokset består altså af, at statens suverænitet i borgerkrige ikke er truet udefra, men indefra af sine egne borgere. Et land i borgerkrig er kan på mange måder derfor karakteriseres som et land uden en stat og et land, hvor borgere kæmper mod borgere og stat. Som et symbol på denne tilstand af borgerkrig vækkes der associationer til Thomas Hobbes tankeeksperiment for naturtilstanden, der beskriver menneskets liv uden en beskyttende stat. Selv beskriver Hobbes naturtilstanden som en tilstand, hvor at menneskets sikkerhed kun er givet af sin egen styrke, hvor fjender findes overalt og hvor at hverken samfund, kultur eller viden kan eksistere (Hobbes, 2012, s.124)

Naturtilstanden, som her skal symbolisere en tilnærmelse af livet under borgerkrig, vidner om en eksistens, der for mennesket er reduceret til en kamp om overlevelse. For at undgå borgerkrig og de katastrofale følger disse har for menneske og samfund, symboliseret ved naturtilstanden, er det nødvendigt at forebygge og etablere afværge-mekanismer inden at borgerkrige bryder ud. Netop her finder nærtliggende speciale sin anvendelse, da anskuelsen af borgerkrige som et prædiktionsproblem, netop åbner op for diskussion af, om borgerkrige helt kan undgås. I den forbindelse er forholdet mellem litteraturens fokus på kausal estimationer og inferens kontra prædiktion en central problemstilling, som jeg kort vil belyse i næste afsnit.

### 1.3. inferens vs. prædiktion

I den samfundsvidenskabelige kvantitative forskning har normalvidenskaben, for nu at låne Thomas Kuhns begreb (Kuhn, 2012), beskæftiget sig med at afsøge kausale sammenhænge og effekter med det formål at drive inferens. Denne søgen efter sammenhænge og generaliseringer gennemsyrrer den kvantitative samfundsvidenskab og er blevet kodificeret af blandt andre King, Keohane & Verba i deres værk *Designing Social Inquiry* 1995, der ofte fremhæves som et forbillede for forskning. Postulatet om, at der er tale om ét dominerende paradigme indenfor samfundsvidenskaben kan diskuteres (Klemmensen et al., 2010, s. 19), men ikke desto mindre er det dette fokus på kausalestimation og inferens i samfundsvidenskaben, som dette speciales metode stiller sig overfor, ved at fokusere på at prædiktion.

Pointen med overstående overvejelse om forskningsparadigmer og formålet med forskning er ikke at sige, at denne type forskning er forkert eller *outdated*. Uden de centrale pointer og indsigter som forskning inden for dette paradigme har bidraget med, ville dette speciale mangle et teoretisk fundament til at guide forventninger og indsamlingen af data, der er fuldstændig essentiel. Pointen er derimod, at endemålet med forskning i borgerkrige ikke må være kausalestimation og inferens, da dette ikke er en løsning på det problem som borgerkrig er, men derimod blot en forståelsesramme for fænomenet. Samtidig må det derfor erkendes, at prædiktionsstudier ikke kan være underlagt de samme evalueringskriterier for god forskning i form af evnen til at drive inferens. I stedet må prædiktionsstudier vurderes og evalueres ud fra deres evne til at prædiktere fremtidige begivenheder.

### 1.4. Specialet opbygning

For at besvare problemstillingen vil jeg i nærliggende speciales **kapitel 2** redegøre for eksisterende studier, der anskuer borgerkrig som et prædiktionsproblem. Kapitlet har til formål at skildre de erfaringer som min undersøgelse står på skuldrene af, samt opridse den kritik som disse studier har modtaget. Opbygningen af en prædiktionsmodel indbe-

fatter en udvælgelse af faktorer, der skal udgøre kernen i prædiktionsmodellerne. Dette er temaet for **kapitel 3**. Her vil jeg udpensle de tre teoretiske tilgange til borgerkrig, der dominerer litteraturen, for at retfærdiggøre mit valg af data og sandsynliggøre faktorernes prædiktionsværdi. Centralt i specialet står begrebet *machine learning*. For at dette begreb ikke bliver et uforklaret *buzzword* gennem specialet, vil jeg i **kapitel 4** lægge det teoretiske fundament for *machine learning* og introducere de algoritmer, der er relevante for mine prædiktionsmodeller.

I **kapitel 5** vil introducere de faktorer indenfor borgerkrig som jeg vil forsøge at prædiktere. Derudover vil jeg i dette afsnit ligge vægt på det datasæt, som jeg selv introducere til litteraturen og som bør ses som et selvstændigt bidrag. Kapitlet vil afslutningsvist opsummere det samlede datasæt. I **kapitel 6** foretager jeg selve min analyse og evaluerer resultaterne individuelt og samlet på tværs af modellerne. I **kapitel 7** vil jeg diskutere prædiktionsmodellernes potentiale som løsningen på borgerkrig som prædiktionsproblemer, med fokus på implementering af modellerne. Afslutningsvist vil afledt af analysen, diskutere implikationen af analysens resultater på den eksisterende teori.

---

## 2. LITTERATUR-REVIEW

Borgerkrig er som forskningsfelt ganske velbelyst, hvilket blandt andet kan illustreres med de over 2.000.000 millioner hits, der gives på Google Scholar, hvis man søger på "civil war". I det følgende litteraturreview vil jeg ligge vægt på den delmængde, der beskæftiger sig med prædiktation af borgerkrig. Her vil jeg optegne de store streger i litteraturen og fremhæve enkelte studier, der har en særlig relevans for dette speciale og som har haft betydning for feltet som helhed. Når man ser på litteraturen, der beskæftiger sig med borgerkrig som et prædiktionsproblem, er der særligt 2 distinkte metodiske fremgangsmåder, der træder frem: *One Shot Decision or Event & Probability and Prediction of Outcome*. For at retfærdiggøre min fremgangsmåde, vil jeg i afsnit 2.1 introducere afgrænsningen mellem disse. Herefter vil jeg i afsnit 2.2, give et overblik over nogle af de fremtrædende studier. Formålet er her at afdække belyste vinkler, drage erfaringer fra eksisterende studier og imødegå den kritik som disse studier har modtaget.

### 2.1. Fra konkret risikovurdering til generisk model

Indenfor traditionen af prædiktation af borgerkrig, er det muligt at identificere flere distinkte perioder, hvor forskellige formål, metoder og datakilder har været dominerende (Brandt et al., 2011). Den mest centrale skelen mellem disse er forholdet mellem *One Shot Decision or Event & Probability and Prediction of Outcome*.

Førstnævnte tilgang er som navnet antyder, en meget snæver tilgang til prædiktation. Her har prædiktationsmodellen til formål at guide beslutningstagen omhandlende en enkelt begivenhed, ved at finde den handling, der optimerer den forventede nytteværdi. Grundmodellen og principperne for denne type af prædiktationer er udviklet af Bueno de Mesquita (Mesquita, 2011, s. 66). Denne type modeller, bygger på et nyttemaksimerende *rational choice* fundament, samt appliceringen af spilteori. Modellen har vist sig i stand til ikke kun at kunne forudsige et *outcome*, men også vejen dertil Mesquita, 2011, s. 70-73. Brugen af spilteori i modelleringen, giver samtidig mulighed for, at modellerne løbende kan reevalueres i takt med at de faktiske forhold udspiller sig. Denne type model kan ses anvendelig i tilfælde, hvor man ønsker at prædiktere udfaldet af en enkelt begivenhed, samt evaluere hvilke policy, der i den konkrete kontekst vil give den største nytte. For borgerkrig må dette ses som særdeles relevant, når det skal vurderes om nødhjælp, intervention eller fredsmægling er den foretrukne handling. Fremgangsmåden har ikke stået ukritiseret hen og der kan blandt andet påpeges problemer med afrapportering af usikkerheder og tidsrammen for prædiktationens udfald (Brandt et al., 2011, s.43).

*Probability and Prediction of outcome* tilgangen er, som navnet røber, en ramme for at prædiktere et *outcome* eller sandsynligheden for *outcome*. Rammen opererer på et

generisk plan, da den ikke inkorporer spilteoretiske komponenter såsom specifikke aktører og disses handlingsrum (Brandt et al., 2011, s.44). Dermed er det muligt at lave prædiktioner af "borgerkrig" for alle lande, som er formålet med dette speciale. Traditionelt har tilgangen været baseret på simple regressionsmodeller, hvor at sandsynligheden for borgerkrig er vurderet på en længere årrække. Faktorerne i modellerne har ofte bygget på det teoretiske fundament som f.eks. Collier & Hoeffler (2004) & Fearon & Laitin (2003) har bidraget med til litteraturen. Igen er modellerne blevet kritiseret for, at operere med for lange tidsperioder og en manglende afrapportering af prædiktionernes usikkerheder (Brandt et al., 2011, s. 45). Jeg har valgt at forfølge denne ramme i udarbejdelsen af mine prædiktionsmodeller, da mit mål er at lave en generisk global prædiktionsmodel, der kan implementeres uanset kontekst. For ikke at reproducere de fejl og udfordringer som studier indenfor denne tilgang har stået overfor, vil jeg i følgende afsnit se nærmere på eksisterende studier.

### 2.2. Prædiktion af borgerkrige

Litteraturen der behandler borgerkrig som et prædiktionsproblem er komplekst, både hvad angår genstanden for prædiktion og valget af algoritmer og data. Fremfor at lave en forfejlet kronologisk gennemgang af litteraturen, har jeg i dette afsnit valgt at fremhæve studier, der synes særligt fremtræden, når man følger en *snowballing* afsøgning af litteraturen<sup>4</sup>.

Som en start på dette afsnit kan det omdiskuterede studie "Predicting armed conflict, 2010-2050" af Hegre et al., 2013 fremhæves. Forfatterne forsøger gennem brug af *machine learning* at prædiktere borgerkrig op til 40 år ud i fremtiden. Denne store tidshorisont har netop været omdrejningspunktet for diskussion omkring studiets anvendelighed (Cederman and Weidmann, 2017). Studiet forventer, at konflikt opstår, eskalerer og afsluttes på baggrund af eksogene forhold, der lader sig påvirke af endogene variable. Metodisk gør forfatterne brug af en spilteoretisk simulation, hvor ændringer i forhold et givent år, betyder ændrede sandsynligheder i det efterfølgende år. F.eks vil tilstedeværelsen af politiske uroligheder og små oprør, øge sandsynligheden for større oprør og borgerkrig (Hegre et al., 2013, s. 2-7). Forfatterne finder, at der er en stiafhængighed i transitionen mellem konflikters intensitet og udbredelse og at lande, der har oplevet konflikt derfor er tilbøjelige til at ende i en *conflict trap*, hvor at volden spiralerer (Hegre et al., 2013, s. 31-32). Denne indsigt er særlig relevant i forhold til det datasæt over politiske begivenheder, herunder uroligheder, jeg vil introducere.

Med en kortere tidshorisont for prædiktionen finder man blandt andet studiet "A global risk assessment model for civil wars" 2009. Her udarbejder forfatterne en række prædiktionsmodeller for den sandsynlighed, et land har for at opleve borgerkrig de kommende 5 år. Meget lig den tilgang jeg anlægger udspringer deres studie af *greed*

---

<sup>4</sup>De mest centrale *keywords* har været: Civil war, predicting og machine learning

litteraturen og forsøger at lave konkrete risikovurderinger for alle lande i deres datasæt. At målet er at give en risikovurdering for en kommende 5-årig periode, skal ses i lyset af, at studiet udelukkende bygger op samfundsstrukturelle data (Rost et al., 2009, s. 922), hvilket igen understreger vigtigheden i at inddrage data om et lands politiske landskab, for at kunne lave tidsligt nære prædiktioner. Studiet er i vidt omfang i stand til at sandsynliggøre borgerkrig, hvor der faktisk har været borgerkrig (Rost et al., 2009, s. 931-932).

Foruden de generiske og langsigtede prædiktionsmodeller findes der en række af casestudier, der har et enkelt land eller region som case. Disse studier har vist sig at være i stand til at prædiktere udbrud af konflikt meget præcist, på baggrund af færre variable, der til gengæld er mere kontekst specifikke, se blandt andre Blair et al., 2016; Brandt et al., 2011. Den klare fordel ved disse studier er, at det er muligt at inddrage flere aktørspecifikke variable og gøre brug af ekspertviden. Skræddersyningen til en case resulterer i gode prædiktioner, mens den klare ulempe ligger i praksisanvendelsen, da det internationale samfund ville være nødsaget til at lave unikke modeller for hvert enkelt land (Brandt et al., 2011)

I artiklen *Reading Between the Lines: Prediction of Political Violence Using Newspaper* Mueller and Rauh, 2018 tiltaler forfatterne det faktum, at der i de empiriske undersøgelser er meget variation mellem lande, men sjældent variation inden for hvert enkelt land i dets strukturelle variable. Gennem en emneordssøgning i 700.000 nyhedsartikler, skaber forfatterne en *within*-variation i det enkelte land. Det gør de ud fra en tese om, at de emner der kommer til udtryk i nyhedsmedierne, vil være forskellige i perioder med fred og perioder, der leder op til konflikt (Mueller and Rauh, 2018, s. 2-4). Forfatterne er på baggrund af deres model i stand til at prædiktere konflikt og flygtningestrømme 1 år før borgerkrigene bryder ud i deres *out-of-sample* datasæt (Mueller and Rauh, 2018, s. 31-32).

Værdien i nuanceringer i de algoritmer som *machine learning* bygger på og i mængden af data, der inddrages, er blandt andet blevet afsøgt af Perry, 2013. I sit studium afsøger Perry, hvordan at præcisionen i prædiktionerne varierer alt efter valg af algoritme. Samtidig finder forfatteren at antallet af uafhængige variable har en stor betydning for algoritmernes *out-of-sample performance*. Dette forklarer han gennem *Bayes theorem*, der antager at forekomsten af en variabel er uafhængig af forekomsten af en anden variable, og hver bidrager med en sandsynlighed for et bestemt udfald (Perry, 2013, s. 4-9). Perry, 2013 finder i sin afdækning at *ensemble* algoritmen Random Forest performer marginalt bedre, når det kommer til prædiktation af udbrud af borgerkrig.

Et andet eksempel på et ambitiøst prædiktionsstudium, der undersøger på metodikken i *machine learning* og samtidig evaluere teoriforgreningerne indenfor borgerkrigslittera-

turen er Muchlinski et al., 2016. Med afsæt i et lande-år datasæt for perioden 1945-2000, der rummer 88 variable, søger forfatterne at forudsige udbrud af borgerkrig. Her tester forfatterne variationer af de modeller, der hver er anvendt i flere af de paradigmesættende inferens-orienterede kausalestimationsstudier (f.eks. Fearon and Laitin, 2003; Collier and Hoeffler, 2004) op mod Random Forest algoritmen. Forfatterne finder, at Random Forest algoritmen udkonkurrerer de andre modeller og prædiktere borgerkrig i alle de tilfælde, hvor der faktisk har været borgerkrig (Muchlinski et al., 2016, s. 94).

### 2.3. Kritikken og centrale indsigter

Kritikken af prædiktionsstudierne har ikke været usagt. Således sætter Cederman and Weidmann, 2017 fokus op de faldgrupper, der er i prædiktionsstudierne og forsøger at dæmpe den *hype*, der synes at være omkring *machine learnings* potentiale.

Den første kritik som Cederman and Weidmann leverer er, at de eksisterende prædiktionsstudier opererer med en for stor tidshorisont eller opløsning. Hvis målretningen af policy-tiltag skal være effektive, nytter det ikke at man ser 1-10 år ud i fremtiden (Cederman and Weidmann, 2017, s. 1). For at imødegå denne kritik vil jeg strukturere mit data således, at det er muligt at prædiktere sandsynligheder og *outcomes* måned for måned. En markant forbedring af den temporale opløsning fra de eksisterende generiske studier.

Den anden kritik der leveres er, at eksisterende studier i for høj grad beror på strukturel data, der ikke kan indfange pludselige ændringer i det politiske landskab. Et fokus på strukturelt data er ikke nok. Som forfatterne selv opsummerer kritikken:

In the absence of full knowledge of how all theoretical components interact and sufficient data to measure the relevant variables, all that can be hoped for is risk assessment on the basis of structural features that increase the probability of conflict (Cederman and Weidmann, 2017, s. 2)

Citatet ligger op til, at prædiktion af borgerkrig ikke kan udvikle sig til meget mere end baseline vurderinger af sandsynligheden for borgerkrig, grundet kompleksiteten i feltet og den overraskende adfærd, der kan være blandt aktørerne. Kritikken imødegås ved at introducere en ny anvendelse af GDELT datasættet (Leetaru and Schrodtt, 2013), der med sine 250.000.000 politiske nyhedsartikler sammen med den høje tidslige opløsning, forhåbentligt kan skabe den dynamiske og specifikke prædiktion som forfatterne efterspørger.



---

### 3. TEORETISKE PERSPEKTIVER PÅ BORGERKRIG

Når man beskæftiger sig med prædiktion gennem *machine learning* er det ikke muligt at udlede de korrelationer, der måtte være mellem de uafhængige variable og den afhængige responsvariabel. Selvom *machine learning* dermed gør prædiktionsmodeller til en teoretisk *blackbox*, bør teorien alligevel have en fremtrædende karakter, da teorien netop derfor er den vigtigste kilde til information om udvælgelse af relevante variable (Hastie et al., 2017, s. 459). De teoretiske perspektiver på borgerkrig er derfor fokuset i dette kapitel.

Før man kan gøre sig forhåbninger om at prædiktere borgerkrig, må man først spørge sig selv, *hvorfor opstår borgerkrig?* Netop dette spørgsmål har været omdrejningspunktet for store dele af litteraturen omhandlende borgerkrige f.eks. (Collier and Hoeffler, 2004; Fearon and Laitin, 2003; Wig, 2016; Collier et al., 2008; Esteban and Ray, 1994) (Murshed, 2009 for et overblik). Borgerkrige er en kompleks størrelse, der er svær at forklare med simple kausalhistorier og forudsætninger. Samfundsstrukturer, økonomiske faktorer, gruppedynamikker og enkeltstående begivenheder spiller alle en central rolle, når det kommer til borgerkrig. Samtidig opstår borgerkrig typisk ikke fra den ene dag til den anden, men er kulminationen på eskalerende lokale konflikter. Netop denne kompleksitet har resulteret i en række forskellige teoretiske bud på, hvordan at borgerkrig skal forstås. Her står særligt 3 teoretiske grene centralt.

Den første teoretiske retning, der også kan anskues som fællesmængden for de tre, er det populært benævnte *vækst og konflikt nexus*. Grundpillen er her, at oprør og borgerkrig bliver sandsynliggjort, når befolkningsgrupper oplever fattigdom og manglende økonomisk- og menneskelig udvikling. Selvom teoriretningen indeholder en del nuancer, kan fattigdom og underudvikling ikke alene forklare borgerkrig. Derfor bliver denne forklaringsmodel ofte suppleret med teoretiske perspektiver, der forsøger at forklare den ulogiske handling som borgerkrig kan anses for at være. Her spiller *rational choice* en særlig rolle som ramme for de teoretiske tilgange *greed & grievance*. Hvor at *greed* vægter de økonomiske motiver bag borgerkrig, fokuserer *grievance* tilgangen på den politiske utilfredshed, der kan være i befolkningsgrupper og som kan danne rammen for organiseringen af borgerkrig. I de følgende afsnit vil jeg zoome ind på disse teoretiske perspektiver på borgerkrig

For at dette teoriredegørende kapitel ikke kommer til at hænge som en løs ramme uden om specialet, har jeg valgt løbende at introducere de variable og datakilder, som jeg på baggrund af teorierne, inddrager i mine prædiktionsmodeller. Det samlede datasæt vil blive opsummeret i kapitel 5, hvor jeg vil lægge vægt på de variable, som jeg vil prædiktere, samt det datasæt, der er mit selvstændige bidrag til litteraturen.

#### 3.1. Vækst, velfærd & borgerkrig

Økonomisk velstand synes åbenlyst forbundet med risikoen for borgerkrig gennem dets effekt på borgernes velfærd og tilfredshed. Det er med andre ord svært at forestille sig, at en befolkningsgruppe iværksætter en borgerkrig, hvis borgerne er har et højt velstandsniveau og som oplever en menneskelig udvikling. Hvis man acceptere præmissen om, at økonomisk velstand fører til velfærd og tilfredshed i befolkningen, må økonomisk vækst ses som borgerkrigs antitese. Vækst og velstand forebygger kun borgerkrig, hvis disse ikke er biased mod bestemte befolkningsgrupper (Murshed, 2009, s. 15). Under theomet er det altså både relevant at se på, hvilke faktorer der fører til økonomisk vækst og øget velfærd, samt hvilke mekanismer der sikrer at den økonomiske vækst er fattigdomsreducerende for hele samfundet og skaber en samlet menneskelig udvikling. Dette afsnit tjener til formål netop at beskrive hvilke faktorer, der kan være med til at påvirke et lands langsigtede vækst og på den måde bidrage enten positivt eller negativt til den økonomiske velstand i befolkningen.

I afsnittet vil jeg indledningsvist redegøre for den overordnede sammenhæng mellem vækst, velstand og borgerkrig, for på den måde at konsolidere præmissen. Herefter vil jeg se nærmere på den samfundsinstitutionelle ramme, der ofte fremhæves som garanten for at vækst er fattigdomsreducerende. Afslutningsvist vil jeg se nærmere på et lands økonomiske afhængighed af eksport af uforarbejdede naturressourcer og øvrige geografiske faktorer, der i litteraturen har vist sig at påvirke den langsigtede økonomiske vækst og velstandsniveau.

##### 3.1.1. Vækst og konflikt nexuset

Selv om man kan have sine reservationer, og vejen ikke altid er lige, er økonomisk vækst hovedfaktoren når fattigdom skal bekæmpes. Uden økonomisk vækst vil omfordeling af samfundets goder gøre fattige mere lige, men ikke løfte det samlede velstandsniveau. For at løfte hele befolkninger eller befolkningsgrupper ud af fattigdom, er man nødt til at sikre, at samfundskagen bliver større. Selvom vækst ofte vil komme de i forvejen rige til gode, kan økonomisk vækst også være fattigdomsreducerende, hvis dele af den økonomiske vækst finder vej ned til samfundets fattige gennem *tickle down* mekanismer (Murshed, 2009, s. 15). Hvis den økonomiske vækst er vellykket, vil den reducere fattigdom, sikre menneskelig udvikling og reducere uligheden i et samfund og dermed havde fjernet forudsætningen for borgerkrig.

Hvis man opdeler alle lande i mit datasæt i to grupper; dem der har oplevet borgerkrig og dem der ikke har, ses der en tydelig forskel på det gennemsnitlige bruttonationalprodukt (BNP) og væksten af denne. Forskellen er beskrevet i tabel 1. De økonomiske variable stammer fra *World Development Indicators* (The World Bank, 2018), mens opdelingen af lande er med udgangspunkt i den borgerkrigs definition jeg intro-

ducere i afsnit 5 på baggrund af data fra Uppsala Conflict Data Program (Croicu and Sundberg, 2017; Sundberg and Melander, 2013).

**Tabel 1:** Gennemsnitlige BNP pr capita og økonomisk vækst

	BNP pr capita (\$)	Vækst (%)
Lande med borgerkrig i perioden	8708	4.13
Lande uden borgerkrig i perioden	11579	3.53

Gennemsnittene er udregnet for perioden 1989-2017, der er styret i den begrænsning, der er i dataet fra Uppsala Conflict Data Program. Som det ses af tabellen, har gruppen af borgerkrigsramte lande i gennemsnit 2871 \$ pr capita mindre end gruppen af lande, der ikke har oplevet borgerkrig. Selvom det for denne fordeling ikke er udført statistiske test på dette resultats signifikans, er det bemærkelsesværdigt og teoriunderstøttende, at der er så stor forskel på de to grupper. Som det også ses af tabel 1, har gruppen af borgerkrigsplagede lande har en højere gennemsnitlige økonomisk vækst. I følge teorien kunne dette ses som kontra intuitivt, da der forventes en lavere økonomisk vækst i lande med borgerkrig, samtidig med at økonomisk udvikling burde reducere lysten til at indgå i en borgerkrig. Resultatet er bemærkelsesværdigt, men modbeviser ikke nødvendigvis tesen af særligt tre grunde:

Den første grund er i sin essens regneteknisk. Teorien ligger basalt set op til, at det kun er relevant at se på den økonomiske vækst, der ligger forud for en borgerkrig og beskæftiger sig som så ikke med den efterfølgende vækst. Derfor kunne man argumentere for at kun væksten for borgerkrig skulle indgå i opgørelsen. Et muligt resultat af en borgerkrig er nemlig regimeskifte, hvilket kan medføre ændrede økonomiske forhold, herunder andre vækstbetingelser. Med andre ord kunne man have overvejet, kun at opsummere væksten forud for borgerkrigene i stedet for hele perioden som jeg har gjort.

Det andet argument er at finde i den neoklassiske vækstteori. En af forskrifterne er her, at fattige lande bør have en højere vækstrate end rigere lande, da investeringer og afkastet på kapital er større, hvor der er mindre af det (Murshed, 2009, s. 16). Det er med andre ord grundessensen af konceptet i *faldende marginalnytte*, der her er anvendt til at betegne gruppen af fattigere lande som potentielle *catching-up* lande. En diskussion af, om det er denne *catching-up* effekt, der er udtrykt i forskellen mellem de to grupper, ligger uden for dette speciales formål. I stedet skal noten ses som en af flere mulige forklaringer, der kan være på forskellen.

Den sidste forklaring er at finde i samfundsinstitutionernes karakter og deres betydning for fordelingen af ressourcer. Her spiller forhold som kontrol med den udøvende magt, *patrimonialisme* og favorisering af bestemte befolkningsgrupper en central rolle. Da dette er kernen i de følgende afsnit, vil jeg ikke gå dybere ind i det her.

For at indfange et lands økonomiske performance, har jeg valgt at inddrage data på BNP og vækst som proxier for landenes økonomiske *performance*. Dataet er hentet fra *World Development Indicators* (The World Bank, 2018). Da den langsigtede økonomiske performance er central, har jeg på baggrund af dataet skabt en række variable, der indeholder information om den nære historiske performance.

#### 3.1.2. Betydning af institutioner for vækst og fordeling

De politiske institutioner i et land har, som vi skal se, en dobbelt betydning i forhold til sammenhængen mellem økonomisk velstand og borgerkrig. For det første kan institutionernes gensidige indretning ses som en ramme for den økonomiske politik, der føres i et land. For det andet har de politiske institutioner betydning for de beføjelser, der gives til udøvende magt og den kontrol der er af denne. Forhold som har vist sig at stå centralt, når det kommer til fordelingen af de økonomiske ressourcer i et land.

Når man ser på samfundsinstitutionerne i et land, er den første pointe man støder på, at de økonomiske institutioner bør være centraliserede samt uafhængige af de politiske institutioner. Fordelen ved at have centraliserede økonomiske institutioner skal ses i lyset af den historiske udvikling, der er sket i gennem forskellige finansielle regimer, der ligger uden for specialets ramme. For et overblik henvises der til Goodman and Pauly, 1993. At de økonomiske institutioner skal være uafhængige af de politiske institutioner, kan anskues som et mere politisk spørgsmål, hvor svaret ikke er klart. Fordelen ved uafhængigheden er at politiske interesser, som ikke er forenelige økonomiske interesser for landets som helhed, ikke kommer til at stå i vejen for hinanden (Murshed, 2009, s. 22). Grundtanken er, at de økonomiske institutioner, såsom en nationalbank, er politisk uafhængige er de bedre i stand til at stille de institutionelle rammer for økonomien, der optimale for vækst. f.eks. ved at være at styre renten med det formål at sikre en prisstabilitet eller tiltrække investeringer (McNamara, 2002, s. 47-48).

Foruden den politiske uafhængighed af de økonomiske institutioner i et land, er de politiske institutioner af stor betydning for velstanden i befolkningen, særligt når det kommer til fordelingen af ressourcer. For at sikre, at politiske institutioner kan skabe de bedste vilkår for vækst og velstand, argumenterer blandt andre Acemoglu, Johnson og Robinson 2005 ligeledes for, at samfundets økonomiske- og politiske institutioner skal være uafhængige af hinanden. Forfatterne argumenterer for, at politiske institutioner skal være indrettet således, at der er kontrol med den udøvende magt i et land, så dennes interesse ikke leder til økonomisk ineffektivitet og fattigdom (Acemoglu et al., 2005, s. 2-4). Kontrollen er central, da særligt sub-sahariske afrikanske lande har været plaget af depotisme, *patrimonialisme*, *patronage* og korrupsion, der resulterer i et ineffektivt bureaukrati, hvor den enkeltes velstand er afledt af forholdet til den udøvende

magt (Taylor, 2017, s.74). Den manglende demokratiske *governance*, der kendetegner disse systemer har desuden vist sig at have en sammenhæng med politisk ustabilitet og *coup d'états* (Reno, 2017, s. 137-138).

For at undgå et sådan politisk system, med utøjlet økonomisk politik, der kun tilgodeser magteliter og smalle samfundsgrupperinger, taler blandt andre Acemoglu et al., 2005 for, at der må være en balancering af den udøvende magt samt en række rettigheder i befolkningen. Her har eksistensen af ejendomsretten og en håndhævelse af selvsamme en afgørende rolle. Hvis ejendomsretten ikke er beskyttet, forsvinder dele af incitamentet økonomisk aktivitet da der ingen sikkerhed er for afkastet af investering og ens arbejde, da magteliter vil være i stand til at drive rov på samfundet og allokere de økonomiske ressourcer på egne hænder (Murshed, 2009, s.23).

For at indfange i hvilken grad samfundsinstitutionerne i et land er gearret mod forhold, der fordrer samfundsvækst og velfærd for befolkningen frem for kun at tilgodese magtelitens egne økonomiske interesser, må man altså se på i hvilken grad at borgerne kan balancere magtelisten gennem de politiske institutioner. Et centralt element i denne balancering er altså borgernes mulighed for at sige fra overfor politikere og magtelisten gennem samfundsinstitutionerne. Her er graden af demokrati altså af afgørende karakter. Uanset om man ser på et minimums demokrati eller lader begrebet omfatte frihedsrettigheder, ligger der i demokratiet altid en mulighed for balancere magthaverne gennem valg (Dahl, 1991).

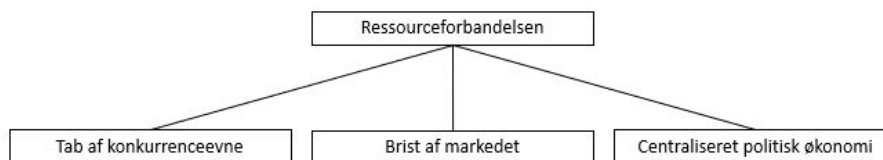
På baggrund af ovenstående indsigter, har jeg valgt at inddrage data på et lands samlede grad af demokratisering, samt selvstændige mål for reguleringen, konkurrencedygtigheden og åbenheden omkring den proces, hvormed den udøvende magt bliver valgt med udgangspunkt i *POLITY IV* datasættet (Marshall et al., 2016). Demokratiseringsvariablen fra *POLITY IV* datasættet udtrykker netop graden hvormed borgere kan udtrykke præferencer for den førte politik, valg af den politiske ledelse, institutionelle begrænsninger på den udøvende magt samt borgerrettigheder (Marshall et al., 2016, s. 14). Alle elementer, der må vurderes til at være relevante i forhold til institutionernes rolle for fordelingen af samfundets goder.

#### 3.1.3. Ressourceforbandelsen

Målet med dette afsnit er at afsøge den populært benævnte ressourceforbandelses indvirkning på en lands langsigtede vækst og derigennem dennes direkte og indirekte påvirkning af risikoen for borgerkrig. Ressourceforbandelsen er en tese, der i alt sin enkelthed går ud på, at hvis et lands økonomi baserer sig på eksport af naturressourcer, så vil det potentielt være forbundet med en række eksternaliteter, der kan have en negativ indvirkning på et lands langsigtede økonomiske performance. Tesen går på flere ben, herunder det potentielle tab af konkurrenceevne, risikoen for at markedet brister,

og en tendens til centraliseret grådighed i det politiske system. Dette er illustreret i figur 2

**Figur 2:** Oversigt over ressourceforbandelsen dimensioner



#### **Tab af konkurrenceevne**

Theomet første ben *tab af konkurrenceevne* har sine rødder i afhængighedsteorien som udviklingsparadigme, introduceret af blandt andre Samir Amin (1976) og Raúl Prebisch (1950). Hvis man lægger kritikken af det internationale system til side inden for udviklingsteorien, er udgangspunktet at økonomier, der baserer sig på eksport af uforarbejdede primærvarer på langsiget vil klare sig dårligt. Det skyldes at prisen på uforarbejdede primærvarer såsom fødevarer er indkomst u-elastisk. Konkret betyder dette at efterspørgslen på varerne ikke vil stige, når indkomsten stiger og der dermed er en øvre grænse på, hvor stor eksporten kan være. Samtidig vil eksporten af uforarbejdede primærvarer, såsom naturressourcer, have en lavere værdi end forarbejdede varer og dermed vil der ske en udbytning mellem lande, der eksporterer uforarbejdede naturressourcer og lande der importerer disse for at bruge dem som input i deres produktion (Murshed, 2009, s.29). Hvis produktionen i et land kun beror på udvinding og eksport af uforarbejdede primærvarer, vil der dermed være et mindre vækst potentiale end lande, med en mere divers produktion.

#### **Volatiliteten af prisen på naturressourcer**

Det andet element i ressourceforbandelsen relaterer sig også til de uforarbejdede primærvarers relation til markedet. Prisen på ubehandlede primærvarer såsom olie og fødevarer, er på det internationale marked meget flygtige og fra tid til anden vil specifikke markeder opleve bobler og brist (Murshed, 2009, s.29). Hvorvidt at prisen går op eller falder dramatisk er afhængigt af dette specifikke produkt, men årsager såsom udtømningen eller opdagelsen af en given naturressource, fejlslagne høster eller et øget internationalt udbud, står som nogle af de mest udslagsgivende faktorer for markederne. Hvis et lands økonomi er orienteret mod eksport af en specifik uforarbejdet naturressource med få eller ingen produktionskæder til resten af økonomien, vil et boblebrist have stor negativ betydning for landets økonomi og den langsigtede vækst. Som et tilnærmet mål for brist med kollapsende har jeg udarbejdet en dummyvariable, der indikerer om et land et givent år har haft en negativ vækst i BNP.

#### **Den politiske økonomi og indtægter fra ressourcer**

Som vist i forrige afsnit omhandlende sammenhængen mellem institutioner og vækst,

er institutioner vigtige når det kommer til at sikre den langsigtede vækst. Den rette og gode politik vil være stort set ineffektiv i et dårligt institutionelt miljø. Indenfor ressourceforbandelse theomet er det også blevet foreslået, at et lands indtægter fra naturressourcer, kan have en negativ effekt på institutionernes kvalitet. Skåret ud i pap har indtægter fra naturressourcer en negativ effekt på samfundsinstitutionerne og derigennem vækst, når indtægterne fører til en *rent seeking* adfærd hos den politiske ledelse, der ofte assisteres af korruption og klientalisme (Murshed, 2009, s. 38). Dette er i hvert fald tilfældet, når man anskuer sagen politisk økonomi ud fra et *rational choice* perspektiv.

Hvorvidt denne kleptokratiske gevinstsøgende og korruptionsprægede adfærd gør sig gældende, eller om den politiske økonomi er mere godgørende og altruistisk er dermed afhængig af de politiske ledes adfærd. Begge tilgange kan ses som det rationelle valg for lederne. Selvom det som altid må ses som irrationelt at skade landets langsigtede økonomiske performance, kan den kortsigtede egennytte maksimerende overskygge dette mål, særligt hvis indtægterne fra naturressource er lettilgængelige for den politiske ledelse.

I et land, hvor at det er let for magteliten at monopolisere indtægterne fra naturressourcer og disse er omsættelige, vil det givetvis være mere tiltalende for grupperinger at søge disse indtægter på en grådig vis (Murshed, 2009, s. 40). Alt i alt afhænger ressourceforbandelsens negative effekt på den politiske økonomi ikke kun af tilgængeligheden af de naturressourcer, men også af elitens egennytte maksimerende interesser. Det er svært at opstille en generel model for, hvordan at den politiske elite vælger at forvalte indtægterne fra naturressourcer, men flere studier tyder på (se blandt andre Caselli and Cunningham, 2017 & Andersen and Aslaksen, 2008), at det afhænger af centraliseringen af magt og fraktionisering i et lands politiske land.

Mere konkret er tesen at præsidentielle systemer koncentrerer mere magt i den politiske ledelse og dermed skaber flere fraktioner og rent seeking behaviour (Andersen and Aslaksen, 2008). Præsidentielle systemer giver dermed anledning til at politikere udtrækker flere indtægter fra naturressourcer og at det offentlige forbrug er mere orienteret mod eliten og grupper med et bånd til den politiske ledelse. Dette kan ses som et særligt problem når nu at store dele af udviklingslandene, der er i en særlig farezone for borgerkrig, er præsidentielle systemer og hvor at der kun er en svag deling af magten, og der i et nævneværdigt omfang praktiseres klientalisme (Murshed, 2009, s. 41, Pitcher, 2017, s. 46).

Hvorvidt naturressourcer skal ses som en velsignelse eller en forbandelse er ikke givet på forhånd, da dette rette institutionelle *setup* og en forskelligartet økonomi kan være bestemmende for udfaldet.

#### 3.1.4. Geografiens betydning for vækst

Et lands velsignelse eller forbandelse med tilgængeligheden af naturressourcer er ikke den eneste eksogent givne geografisk betonedede faktor, som spiller en rolle for et lands vækstpotentiale og velstanden i befolkningen. Her spiller forhold som landets placering og jordens dyrkbarhed også en afgørende rolle.

Landbrugets udvikling og den økonomiske vækst afledt heraf, er i praksis og teoretisk en af de væsentligste faktorer for, at udviklingslande kan opleve et økonomisk *take-off* og løftes ud af fattigdom (Dicken, 2011, s. 271, Rostow, 1991, s. 36-39). Hvis man først ser nærmere på jordens bonitet og klima, er dette direkte givne for landbrugsmulighederne i et land. Fødevarereproduktion er på trods af et globalt marked og en stigning i internationale handelskæder, fortsat primært en lokal proces i verdens fattigste lande. Derfor er jordens bonitet og klimaet af afgørende betydning for fødevarereproduktionen i land og det eksistensgrundlag som befolkningen har. Hvis klimaet er barsk og boniteten er dårlig, er det svært at igangsætte en udvikling og industrialisering af landet eller blot opretholde et eksistensgrundlag. I udviklingslande udgør dyrkningsmulighederne derfor en endnu større risikozone, da befolkningen ofte driver subsistenslandbrug, der er direkte bestemmende for deres velfærd. Her kan fejlslagne høste og naturkatastrofer skabe hungersnød og reducere den menneskelige udvikling i sådan en grad at vækstpotentialet hæmmes og risikoen for konflikt øges. For at indkapsle landbrugets potentiale og afkast har jeg i mit datasæt inddraget et mål for andelen af dyrkbar jord i et land fra Verdensbankens udviklingsindikatorer (The World Bank, 2018).

Den geografiske placering af et land er ikke uden betydning for landets vækst. Et lands forbindelse til hav eller mangel på samme har en afgørende betydning for landets handelsmuligheder. Som Poul Collier udtrykker det i bogen *The Bottom Billion* (2007):

if you are coastal, you serve the world; if you are landlocked you serve your neighbors' and are limited by their economies. (Collier, 2007, s. 56-57)

Fundamentet for tesen er de forhøjede transportomkostninger, der er i forbindelse med handel, når et land ikke har sin egen kystadgang, der særligt skyldes at handlede varer kan være underlagt afgifter til de nabolande, der transiterer varerne (Lahiri and Masjidi, 2012, s. 506). Som mål har jeg taget udgangspunkt i Harvard Universitys liste over *landlocked countries* (Harvard University, 2018)

#### 3.2. Muligheden for oprør og det økonomisk motiverede oprør

Selvom overstående afsnit har tjent til formål at skitsere nogle af de grundbetingelser, der er for borgerkrig - nemlig at borgerkrig er mere forventelige i lavindkomst lande med udbredt fattigdom, må man ikke glemme, at oprør og borgerkrig er resultatet af et menneskeligt valg. Hvis man anskuer dette valg ud fra et *rational choice* perspektiv,



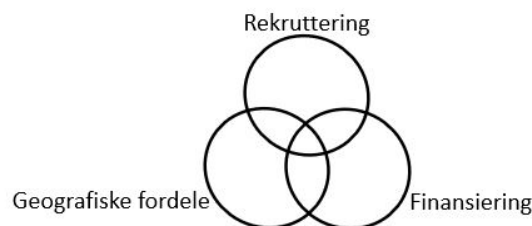
er valget produktet af en nyttemaksimerende kalkule, hvor at den potentielle gevinst ved borgerkrig, holdes op mod omkostningerne forbundet med handlingen. I en sådan nyttemaksimerende kalkule kan det være svært at se, hvordan et aktivt valg om at starte eller indgå i et oprør, der udvikler sig til en borgerkrig, kan være pareto-optimalt for oprørerne. Dette gælder både for magtelisten som beslutningstager for den gruppe de repræsenterer og for det enkelte individ, der sætter livet på spil.

Denne kontraituitive problemstilling for aktørerne, har de fremtonende teoretikere Paul Collier & Anke Hoeffler 2004 undersøgt<sup>5</sup>, ud fra individets og magteliternes økonomiske motiver for at indgå i borgerkrig. I den økonomisk betonedede del af deres studier, anvender forfatterne den simple antagelse, at oprør kan være økonomisk motiveret både for magtelitens side, mens også for det enkelte individ (Collier and Hoeffler, 2004, s. 565). Her findes nytten både i den økonomiske gevinst, der kan være for aktørerne under en borgerkrig i form af plyndringer, men i særlig grad også den økonomiske og politiske magt, der er forbundet med en succesfuld borgerkrig.

Oversat til *rational choice frameworked* kunne man sige, at aktørerne i et oprør optræder som en form for *economic man*, der ønsker at maksimere sin egen "profit" (Murshed, 2009, s.63). Her skal profitten forstås ud fra den magt og økonomiske gevinst, der er forbundet med en succesfuld borgerkrig. Den kalkule, der ligger til grund for om handlingen, resulterer i oprør eller ej, er ikke kun afhængig af incitamenterne bag borgerkrig, men også omkostningerne, der er forbundet med handlingen.

I (Collier and Hoeffler, 2004) studie oversættes omkostningerne i kalkulen til *mulighederne* for oprør. For forfatterne er der en række faktorer, der påvirker disse mulighederne for oprør og borgerkrig, og som derfor må indtænkes i kalkulen (Collier and Hoeffler, 2004, s. 565). Når man anskuer disse incitamenter og omkostninger, kan man med fordel inddele dem i tre hovedgrupper: finansiering af oprøret, rekruttering af soldater & geografiske fordele. Som illustreret i figur 3 er der overlap mellem de tre delelementer, der tilsammen udgør potentialet for det oprør - uanset om det er økonomisk motiveret eller ej.

**Figur 3:** Hovedkomponenterne i greed-tilgangen der tilsammen udgør mulighederne for oprør



<sup>5</sup>blandt andre konkurrerende motivationer

### 3. Teoretiske perspektiver på borgerkrig

---

I de følgende tre afsnit vil jeg gennemgå de faktorer, der i empirien har vist sig have betydning inden for disse hovedgrupper.

#### 3.2.1. Finansiering af oprør

Hvordan man skaffer finansiering til sit oprør er ikke det typiske spørgsmål man stiller sig selv. Ikke desto mindre er det netop det spørgsmål som dette afsnit vil forsøge at besvare. I Collier & Hoefflers paradigmesættende studium, lægger forfatterne vægt på særligt to kilder til finansiering: Indtægter fra naturressourcer og donationer fra diaspora (Collier and Hoeffler, 2004, s. 565).

##### Indtægter fra naturressourcer

Som jeg belyste i forrige afsnit 3.1 kan indtægter fra naturressourcer, være med til at forme og skabe et patrimonialistisk politisk system, hvor magthaveren favoriserer bestemte grupperinger i bytte for deres loyalitet. Ud fra dette perspektiv kan tilstedeværelsen af naturressourcer ses som et af de økonomiske motiver, der er for oppositionens magteliter da en magttransition ville betyde at disse ressourcer skifter hænder. Tilstedeværelsen af letomsættelige naturressourcer spiller også en central rolle, når det kommer til selve finansieringen af oprør og kan ses som en ramme, der er med til at sænke omkostningerne for oprøret, da de giver *muligheder* som kilde til indtægter.

Når det kommer til indtægter fra naturressourcer er den generelle fremstilling, at oprørsgrupper kan erobre områder og virksomheder, hvorfra eksporten af naturressourcer foregår. Derigennem får de adgang til de indtægter, som er forbundet med disse. Dette er, hvad enten der er tale om tømmer, olie, diamanter eller narkotikum (Collier and Hoeffler, 2004, p. 565). I litteraturen har man forsøgt at måle de potentielle indtægter fra naturressourcer som oprørsgrupper kan drive rov på. Både på aggregeret niveau, som den andel af den eksport af primærprodukter udgør af det samlede bruttonationalprodukt (Collier and Hoeffler, 2004), eksporten af olies andel af den samlede eksport (Fearon and Laitin, 2003), Hvorvidt der er en olieeksport i landet (Rost et al., 2009), om der udvindes diamanter og om udvinding af diamanter er overjordisk eller foregår i mineskafter (Lujala et al., 2005a).

Formålet med disse nuanceringer af tilstedeværelsen af naturressourcer som indtægtskilde for oprør, er at åbne op for de forskellige dynamikker, der kan gøre sig gældende i det pågældende land. Eksempelvis finder Lujala et al., 2005a at tilstedeværelsen af sekundære diamanter<sup>6</sup> øger risikoen for at oprør og forlænger varigheden af borgerkrig, mens dette ikke er tilfældet for primære diamanter udvundet i miner.

I et forsøg på at begrænse antallet af variable der måler naturressourcer som en in-

---

<sup>6</sup>Sekundære diamanter referer til diamanter, der findes i geologiske aflejringsmiljøer og som til modsætning til primære diamanter ikke skal udvindes af klippemateriale

dtægtskilde i finansiering af oprør, har jeg valgt at fokusere på olie og diamanter. Det har jeg gjort ud fra en tanke om, at eksempelvis inddragelse af den samlede eksport af naturressourcer, siger meget lidt om dynamikken bag finansieringen og da det ikke er forventeligt, at eksporten af eksempelvis korn giver samme muligheder for finansiering af borgerkrig som plyndringer af diamanter gør. Samtidig er netop diamanter lettere at transportere på tværs af grænser og sælge på illegale markeder (Lujala et al., 2005a, s. 542-543). Olie på den anden side er medtaget, da denne kan ansues som en stor post-konflikt indtægtskilde for den aktør, der vinder konflikten. Dette betyder dog også, at mit data bliver biased mod lande, der netop har olie og diamanter som en naturressource og forsømmer andre naturressourcer, der kan bruges til at finansiere oprør eksempelvis kokain og heroin. Som mål for tilstedeværelsen af olie i et land, har jeg oprettet en dummyvariable med baggrund i Fearon and Laitin, 2003 datasæt. Til at måle tilstedeværelsen af diamanter, har jeg benyttet data fra (Lujala et al., 2005b), med fokus på tilstedeværelsen af diamanter, primære diamanter og sekundære diamanter.

#### **Donationer fra diaspora**

En anden indtægtskilde til finansieringen af oprør er donationer fra migranter og diaspora, der sympatiserer med oprørerne. Som jeg beskriver i afsnit 3.3 er oprør og borgerkrige ofte centreret omkring et gruppetilhørsforhold, såsom etnicitet og religion. Diaspora der deler et sådant tilhørsforhold med oprørsgruppen kan forventes at sympatisere med oprørsgruppen i kraft af deres status som diaspora, der ofte er ufrivilligt. Derfor kan det forventes at en andel af diaspora vil støtte finansieringen af oprøret økonomisk. I litteraturen findes der ikke noget direkte mål for denne økonomiske støtte. Diasporas medfinansiering er tidligere blevet proxiet ud fra antallet af diaspora som er bosiddende i USA, og som deler etnicitet med en af aktørerne i en borgerkrig (Collier and Hoeffler, 2004, s. 568). Selvom denne tilgang har den ulempe at man udelader diaspora, der lever i andre lande, har det den fordel at alle diasporaerne lever under den samme politiske, økonomiske og juridiske enhed. Der er altså en lighedsdannelse mellem deres forudsætninger for at støtte økonomisk. Et andet mål man i fremtiden kunne bruge til at måle diasporaernes finansiering er de indadgående pengeoverførsler der er fra diaspora. Desværre er datagrundlaget for disse pengeoverførsler fortsat mangelfuldt<sup>7</sup>, men bør i fremtiden overvejes som en nærliggende proxy for diasporaernes finansiering af oprør.

På baggrund af disse indsigter har jeg valgt at inddrage data om det totale antal flygtninge fra et land. Dataet tager udgangspunkt i *Ethnicity of Refugees* datasættet (Rüegger and Bohnet, 2018) og vurderes som en bredere proxy for den potentielle finansiering fra diaspora, end den Collier and Hoeffler, 2004 anvender.

---

<sup>7</sup>Verdensbanken er i de seneste år begyndt at opgøre dette data, men dataet lider fortsat af store huller, der gør dataet uanvendeligt i dette speciale (The World Bank, 2018)

#### 3.2.2. Rekrutteringspotentialiet

Foruden muligheder for at skaffe finansiering til våben og "lønninger for oprøret" er rekrutteringspotentialiet for borgerkrigens aktører ligeledes afgørende for om det er mulighed for at gennemføre et oprør. Rekrutteringspotentialiet for en oprørsgruppe kan basalt set koges ned til de individer, der har et økonomisk eller andet politisk motiv for at indgå. Her kan man lave den grundlæggende antagelse, at det rekrutteringsgrundlag vil være større, jo større selve befolkningen i landet er. Volumen af befolkningen kan i midlertidig ikke indfange alle aspekter af rekrutteringsgrundlaget. Her spiller faktorer såsom uddannelsesniveaue, lønniveauet og arbejdsløsheden i et land ligeledes ind, særligt for unge mænd, der kan ses som den befolkningsgruppe, der udgør hovedgruppen i rekrutteringen (Murshed, 2009, s. 65).

For arbejdsløse og uuddannede mænd, der ikke ser et potentiale i at tage et arbejde til lave lønninger, kan rekrutteringen til et oprør, hvor at lønnen er højere ses som det økonomiske motiv og det optimale valg, på trods af den risiko, der er forbundet ved at lade sig rekruttere. For magteliten kan samme faktorer ses en atypisk omstændighed, der er med til at sænke omkostningerne forbundet med rekrutteringen og derigennem give *mulighed* for oprøret. Indkomsten forud for en borgerkrig er derfor af betydning for det økonomiske motiv for at indgå i en borgerkrig. Dette kan illustreres med det omvendte historiske eksempel, hvor at deltage i den russiske borgerkrig disreterere i større omfang om sommeren, hvor at lønningerne som bønder var op mod ti gange højere (Collier and Hoeffler, 2004, s. 569).

Som proxy for det økonomiske motiv, der er for deltagerne i et oprør og som kan ses som omkostningen forbundet med rekrutteringen for oprørets ledelse, har jeg valgt at inddrage 5 variable. De første 2 variable relaterer sig til uddannelsesniveaue i befolkningen, der kan ses som omkostningsforøgende i rekrutteringen. Her er der tale om andelen af befolkningen, der har været indskrevet på et sekundært skoleniveau, og andelen af mænd som har været indskrevet på sekundært skoleniveau. De sidste tre variable relaterer sig til landets økonomi. Her har jeg inddraget data for bruttonationalproduktet, bruttonationalproduktet pr capita samt den økonomiske vækst år for år. Alle 5 variable er hentet fra *world development indicators* (The World Bank, 2018)

#### 3.2.3. Omkostningsreducerende geografiske fordele

Som sagt består incitamentet for at indgå i borgerkrig blandt andet af de økonomiske gevinster modregnet de omkostninger, der er forbundet med oprøret. Hvis man ser nærmere på de omkostninger, der er forbundet med den risiko der er ved oprør i form af at blive overrumplet af et overlegent militær. Her kunne variable såsom regeringens militære kapacitet og budget være oplagte variable at inddrage (Collier and Hoeffler, 2004, p. 569), men som desværre ikke har været mulige at inddrage. I stedet har jeg

valgt at fokusere på to variable, der kan mindske risikoen for at blive overrumplet af regeringens militær, der relaterer sig oprørernes mulighed for at have sikre baser, hvorfra militæroperationer og plyndringer kan udspringe.

Her vil jeg lig litteraturen lægge vægt på oprørernes mulighed for at gemme sig i terrænet. Den grundlæggende tanke er her, at hvis et terræn er bjergrigt og kuperet, vil det være nemmere for et oprør at skjule sig i terrænet og svært fremkommeligt for regeringsstyrker, hvilket giver oprøret mulighed for taktiske geografisk reorganisering (Collier and Hoeffler, 2004, s. 569). Den anden faktor som jeg selv vil introducere her, er arealet af landet. Dette har jeg valgt at gøre ud fra en tankegang om, at jo større et land er, jo mere omkostningsfuldt vil det være for regeringen at projicere sin magt over hele territoriet. Dette kan ses som en fordel for oprørsgrupper, da dette giver mulighed for at operere fra områder, hvor at den militære tilstedeværelse af regeringen, er svagere eller ikke tilstedeværende. Dataet for terræn er hentet fra Fearon and Laitin, 2003, mens arealet er udregnet på baggrund af World Borders Datasættet fra Sandvik, 2018

En anden omkostningsreducerende faktor er den sociale sammenhængskraft i et oprør. Her er forventningen, at oprør vil have en større sammenhængskraft, hvis den er bundet op på gruppeidentiteter relateret etnicitet eller religion. Samtidig vil det være lettere for en oprørsgruppe at opnå sympati i en lokalbefolkning, der deler samme gruppeidentitet som oprørerne (Collier and Hoeffler, 2004, s. 570). Netop gruppeidentitet og tilhørsforhold spiller en central rolle i *grievance* tilgangen til borgerkrig, der er fokuspunktet for følgende afsnit.

### 3.3. Gruppeidentitet: Politisk og økonomisk utilfredshed

En ensartet gruppeidentitet kan i omkostnings-frameworket for borgerkrig, ses som omkostningsreducerende for et oprør, da det skaber en social sammenhængskraft for oprøret. Foruden denne indsigt, spiller identitet og gruppeidentitet en særlig rolle, når det kommer til motivet for at indlede en borgerkrig. Som belyst i afsnit 3.1.1, kan fattigdom og økonomisk ulighed i et samfund ses som fundamentet for borgerkrig. Dette gør sig i særlig grad gældende, når samfundsinstitutionerne skaber et *patrimonialistisk* politisk system, hvor at magthaveren favoriserer bestemte samfundsgrupperinger på bekostning af andre.

Her kan den teoretiske tilgang *grievance* ses et væsentlig bidrag. *Grievance*, der ikke har en dækkende oversættelse til dansk, udtrykker den utilfredshed, der udspringer af politisk undertrykkelse og eksklusion, samt økonomisk ulighed, der er forbundet med tilhørsforholdet til en bestemt samfundsgruppe (Collier and Hoeffler, 2004, s. 570, Murshed, 2009, s. 76). Her kan had mellem etniske og religiøse grupperinger, politisk repression og eksklusion - akkompagneret med økonomisk ulighed ses som væsentlige motiver for oprør og borgerkrig (Collier and Hoeffler, 2004, s. 571)

Som center for den *grievance* der er i et samfund, står individet og gruppeidentitet som individet deler med en samfundsgruppe. Hvis individet og samfundsgruppen som helhed oplever en *grievance*, der kan knyttes til gruppeidentiteten, vil gruppeidentiteten kunne bruges som rekrutteringsgrundlag for et oprør, med omvæltningen af den oplevede *grievance* som motiv. Ved at identificere sig med én bestemt gruppe, bliver der automatisk skabt *in- og out-groups* (Sidnik and Condor, 2014), der er let omsættelige til fjendtlige binære billeder. For oprørslederne kan dette binære billede igen bruges til at dyrke hadet til andre samfundsgrupperinger og som samlingspunkt for oprøret. Et vigtige element i skabelsen af denne gruppeidentitet er de narrativer som oprørsledelse kan bruge som redskab til at udpensle og begrunde det binære forhold, der er mellem samfundsgrupperingerne (Andrews, 2014; Hammack and Pilecki, 2014; Selimovic, 2014).

Når man ser på, hvordan den politiske uretfærdighed manifesterer sig i et oprør er det værd både at anskue uretfærdigheden ud fra et mikro- og et makroperspektiv. Her skal mikroperspektivet forstås som individets oplevede politiske uretfærdighed, der manifesterer sig i samfundsgruppen og gruppeidentiteten, der dermed udgør makroperspektivet. I de følgende to afsnit, vil jeg udpensle henholdsvis disse mikro- og makro-perspektiver på politisk utilfredshed.

#### 3.3.1. Individuel politisk- og økonomisk utilfredshed

Det er vigtigt at forstå, at *grievance* som motiv for borgerkrig, hvad enten den er økonomisk betonet eller ej, udspringer af en individuel utilfredshed. Hvis man løfter denne individuelle utilfredshed ud af dens kontekst, kan den koges ned til diskrepansen mellem *dét* individet mener det fortjener og hvad de mener de modtager, hvad enten der er tale om økonomi eller politisk magt (Murshed, 2009, s. 77).

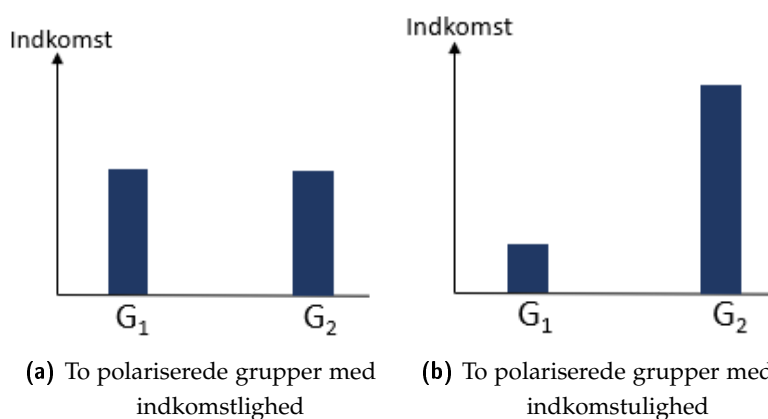
Her er det vigtigt at holde sig for øje, at denne forventning til velfærd og politisk magt er relativ til andre mennesker. Hvis individet oplever en diskrepans mellem *dét forventede* og *det oplevede* samtidig med individet vurderer, at andre individer har mere end de burde, vil der opstå en utilfredshed, der er fundamentet for den oplevede *grievance* (Murshed, 2009, s. 77). Nøgleordet her er altså *det oplevede afsavn*, der findes i diskrepansen mellem individets forestilling om, hvordan at ting bør være og hvordan de faktisk er. Dette relative afsavn bør ses som mikro-fundamentet for *grievance*-tilgangen. Her kan man forvente, at potentialet for kollektiv vold i form af oprør og borgerkrig er afhængig af intensiteten og bredden på det relative afsavn, der er blandt individer, der har det samme gruppetilhørsforhold.

### 3.3.2. Polarisering og Horisontale ulighed

Hvis man antager at oprør er afhængigt af det relative afsavn, der er blandt individer som har samme gruppetilhørsforhold, gennem en forstillet fælles gruppe identitet, er det værd at se nærmere på polarisering og horisontal ulighed mellem grupper. Polarisering som fænomen kan forstås som en tilstand i et samfund, hvor der er mindst to samfundsgrupper, der indadtil er homogene, men som har en stor inter-gruppe heterogenitet. Estaban & Ray 1994, der har beskæftiget sig med konceptualiseringen og målbarheden af begrebet har defineret det som:

Suppose that the population is grouped into significantly-sized “clusters,” such that each cluster is very “similar” in terms of the attributes of its members, but different clusters have members very “dissimilar” attributes. In that case we would say that the society is “polarized” (Esteban and Ray, 1994, s.819)

I studiet af borgerkrig, er disse grupper ofte blevet defineret ud fra etnicitet eller religiøse tilhørsforhold, der dermed har dannet rammen for polarisering (Fearon and Laitin, 2003; Collier and Hoeffler, 2004). Her er det nemt at se den interne homogenitet i form af traditioner, sprog og kultur, der adskiller sig fra andre grupperinger. Som udgangspunkt bør denne polarisering, mellem distinkte grupper ikke give anledning til kollektiv vold, men kan i lyset af den individuelle *oplevede afsavn* af f.eks. indkomst, helbred, magt eller uddannelsesniveau netop danner rammen for konflikt. Som illustreret i figur 4 behøver dette *oplevede afsavn* ikke være begrænset til en forestilling, men være produktet af et politisk system, der enten af demokratiske eller autokratiske veje favoriserer bestemte grupper i samfundet.



Figur 4: Eksempel på to-dimensionel polarisering

Figur 4a skildrer et samfund med to grupper, der er homogene indadtil, men som er heterogene i forhold til hinanden. Selvom indkomsten mellem de to grupper er lige, kan der i grupperne godt være et *oplevet individuelt afsavn*, mens det ikke er forventeligt

### 3. Teoretiske perspektiver på borgerkrig

---

at grupperne oplever afsavnet som *grievance*. Figur 4b skildrer derimod et samfund, hvor der er en reel forskel i indkomstniveauet for de to grupper. På baggrund af polariseringen og den horisontale ulighed, må man forvente at det *individuellet oplevede afsavn* vil manifestere sig i gruppe  $G_1$ , hvilket vil øge fremmedgjorthed over for samfundet, forstærke gruppeidentiteten og antagonismen overfor gruppe  $G_2$ . Dermed kan polarisering føre til en antagonisme, der er motivet i borgerkrig (Murshed, 2009, s. 78-80).

Indenfor politisk psykologi illustreres denne selvforstærkede gruppeidentitet ofte med en telt-analogi. Analogien er ret fundamental og beskriver hvordan at individer vil samles under et teltsejl, der udtrykker gruppeidentiteten. Når der er trusler mod gruppen eller gruppen gennemgår hårde tider, vil menneskerne samles omkring midterstolpen i teltet, der udtrykker den politiske ledelse som er styrende for gruppeidentiteten. Jo mere der bliver rusket i teltsejlet, jo større truslen er, jo mere af den individuelle identitet vil blive tilsidesat for gruppeidentiteten, hvilket netop kan overkomme det kollektive handlingsproblem et oprør kan anskues som (Volkan, 2005).

Begrebet polarisering og de problematikker som jeg har belyst ovenfor relateres i litteraturen ofte til etnisk og religiøs fraktionisering. En skelen mellem polarisering og fraktionisering ligger i antallet af grupper i et land, hvor polaritet udtrykker få grupper, mens fraktionisering udtrykker mange grupper. Skelnen er vigtig, da studier peger på, at der er et kurvelinært forhold mellem konflikt og antallet af grupper, med flest konflikter ved få grupperinger. *Rational choice* forklaringen på dette ligger i, at omkostningerne ved oprør øges jo flere grupperinger der er i landet (Murshed, 2009, s. 79). Denne forventning bliver i litteraturen bekræftet. Fearon & Laitin finder af etnisk fraktionisering ikke er en robust estimator for udbruddet af borgerkrig (Fearon and Laitin, 2003). Montalvo & Reynal-Querol 2005 finder ligeledes ikke en effekt af etnisk fraktionisering, men derimod at etnisk polarisering er en robust estimator for udbruddet af borgerkrig. Sagt med deres egne ord (Montalvo and Reynal-Querol, 2005, s. 812).

Som mål for det individuelt *oplevede afsavn* og horisontale skillelinjer i samfundet har jeg inddraget en række variable. Den første række variable udtrykker den etniske og religiøse fraktionisering i et land og hentet fra Fearon and Laitin, 2003. Et vigtigt element i det *oplevede afsavn* er mulighederne for at ændre på det politiske system. Her står den politiske magt, gruppen har, som det stærkeste virkemiddel (Wig, 2016, s. 513-515).

På baggrund heraf, har jeg med udgangspunkt i *Ethnic Power Relations* datasættet inddraget en række variable, der udtrykker variationen i etniske gruppers adgang til politisk magt både som gruppernes andel af befolkningen, dummyvariable og befolkn-



ingstal. Her har jeg lagt vægt på følgende 4 kategorier: Grupper der enerådigt har magten, grupper der deler magten, grupper der er ekskluderet fra magten og grupper der i datasættet vurderes som irrelevante. Alle disse variable der indfanger *grievance* på et strukturelt niveau. For at komme tættere på den faktiske politiske utilfredshed, der måtte være indenfor en samfundsgruppe, har jeg, som jeg vil vende tilbage til i afsnit 5 på side 47, inddraget data over politiske begivenheder, der udtrykker blandt andet politisk utilfredshed.

### 3.4. Samspillet mellem de teoretiske tilgange

I dette kapitel har jeg set nærmere på den eksisterende teori for borgerkrig, der tager udgangspunkt i spørgsmålet *hvorfor opstår borgerkrig?* I kapitlets første afsnit tog jeg fat i den ramme for borgerkrig, der er gennemgående, nemlig at borgerkrig udspringer af fattigdom og økonomisk ulighed. Som jeg belyste, spiller samfundets institutioner her en særlig rolle i forhold til fordelingen af ressourcer og for de vækstbetingelser der i et land. Ud fra et vækstperspektiv belyste jeg desuden nogle af de geografiske faktorer, der kan påvirke den langsigtede vækst herunder et lands afhængighed af naturressourcer, der kan ses som særligt relevant i relation til borgerkrig. I kapitlets andet afsnit, så jeg nærmere på nogle af de omstændigheder, der kan gøre det mere eller mindre omkostningsfuldt at gennemfører et oprør, for desuden at introducere det økonomiske motiv, der kan være for oprørernes side. I kapitlets sidste afsnit var fokus ligeledes på motiverne bag oprør og borgerkrig med en vægtning på den *grievance*, der kan opstå i forskellige samfundsgrupperinger.

### 4. MACHINE LEARNING OG PRÆDIKTIONSMODELLER

I dette kapitel vil jeg zoome ind på specialets metodiske tilgang. Her vil jeg først introducere begrebet *machine learning* med det formål, at åbne begrebet op. Herefter vil jeg introducere *opbygningen af en prædiktionsmodel* i praksis og som funktion. Disse indledende afsnit har til formål at skildre det potentiale som *machine learning* rummer i prædiktionen af borgerkrig, samt lægge retningslinjerne for evalueringen af prædiktionsmodellerne i analysen. Kapitlet afsluttes med en introduktion til de *machine learning* algoritmer, der anvendes i analysen.

#### 4.1. Machine learning og algoritmer

Hvad er *machine learning* egentlig? For uindviede fremstår begrebet som et *buzzword*, der dækker alt fra kunstig intelligens, selvkørende biler til spam-filteret i ens e-mail indbakke. *Machine learning* kan, på trods af dets mange facetter, bedst beskrives som en form for statistisk læring, hvor en computer gennem en række algoritmer træner en model til at genkende mønstre i data, der igen bruges til at forudsige et udfald af en *responsvariabel* (James et al., 2013, s.15, Knox, 2018, s. 1-2). Der hvor *machine learning* algoritmer adskiller sig fra de, i samfundsvidenskaben, velkendte statistiske læringsmodeller, er ved at modellens algoritmer ikke eksplicit er programmeret til at anvende en bestemt form - eksempelvis lineær, men i stedet selv at finde frem til den form, der giver den optimale *performance* i forudsigelsen af responsvariablen.

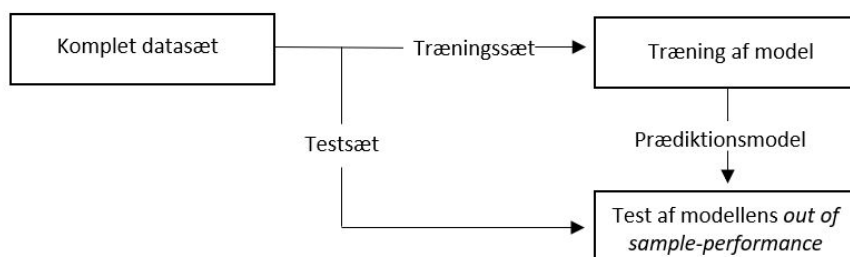
Algoritmerne står altså centralt i *machine learning*, da de er afgørende for mønstergenkendelsen og dermed udregningen af forudsigelsen af responsvariablen. I denne sammenhæng skal algoritmen forstås som det sæt af regler og den fremgangsmåde *machine learning* modellen opererer under. Begrebet algoritme kan i vid udstrækning oversættes til de mere samfundsvidenskabelige begreber "model" eller "metode", men har typisk en mere matematisk karakter. Inden jeg introducerer de algoritmer, der er centrale i dette speciale, vil jeg først introducere en generel note om opbygningen af prædiktionsmodeller, der er det grundlag som algoritmerne arbejder ud fra. Dette er temaet for følgende afsnit.

##### 4.1.1. Prædiktionsmodeller i praksis

En prædiktionsmodel er grundlæggende statistiskmodel der er *fittet* på baggrund af et kendt datasæt, som anvendes til at prædiktere udfaldet af *responsvariablen* i et ukendt datasæt (James et al., 2013, s.15). Dette giver en grundlæggende implikation for evalueringen af en prædiktionsmodel, da evalueringsgrundlaget, det ukendte data, ligger ude i fremtiden. Det er med andre ord ikke relevant at evaluere en prædiktionsmodel på baggrund af, hvor godt modellen *fitter* det kendte data (Varian, 2014, s. 6). Da det ikke kan ses som attraktivt at vente på at ny data generes før man kan evaluere en

prædiktionsmodels forventede performance, opdeler man i praksis sit kendte datasæt i to distinkte dele, som illustreret i figur 5

**Figur 5:** Opdeling af data i træningssæt og testsæt



Grundtanken ved opdelingen af datasættet er, at man således har en tilfældig distinkt delmængde af sit data, træningsdatasættet, hvorpå man kan træne prædiktionsmodellen og en tilfældig distinkt delmængde data, testdatasættet, hvorpå man kan teste og simulere modellens *out-of-sample performance*.

Denne tankegang har jeg videreført til mit datasæt, hvor jeg i stedet for at opdele dataet gennem tilfældigudvælgelse, har opdelt data på baggrund af observationernes tidslige placering. Det har jeg gjort, da der f.eks. ikke er meget værdi i at kunne prædiktere borgerkrig i et land i 1995, hvis prædiktionsmodellen er bygget på data for landet i henholdsvis 1994 og 1996. Jeg har lagt mit split i 2012, så min træningsdatasæt udgør observationer i perioden 1989-2011 (20 år) mens min testdatasæt består af observationer fra perioden 2012-2017 (7 år). Denne såkaldte blok-opdeling skal altså sikre, at træningsdataet ligger temporalt tidligere end træningsdatasættet og dermed sikre at jeg ikke performer bedre end muligt (Varian, 2014, s. 6).

#### 4.1.2. Prædiktionsmodeller som funktion

Som sagt kan *machine learning* i sin essens ses som statisk læring, hvor at en computer på baggrund af en algoritme, finder mønstre i et datasæt, som den bruger til at forudsige responsvariablen (James et al., 2013, s.15-16). Der er med andre ord tale om at *fitte* observationer til en model, der anvendes på et nyt datasæt. I nærliggende speciale er formålet at lære algoritmen at forudsige omstændigheder omkring borgerkrige, med så stor præcision som mulig. Det data der skal gøre algoritmen i stand til at prædiktere benævnes *prædiktorer*<sup>8</sup>, mens det der skal prædiktes benævnes *responsvariablen* (James et al., 2013, s. 15)

I sin mest generelle forstand er et prædiktionsproblem, når man er interesseret i den betingede fordelingen af responsvariablen  $y$  givet prædikatorene  $x$  (Varian, 2014, s.

<sup>8</sup>uafhængige variable, features og blot variablene er alle gængse ord for samme fænomen

6). Når man vil estimere responsvariablen, foregår det på baggrund af en funktion, der beskriver forholdet mellem prædiktorerne og responsvariablen. Dette kan matematisk illustreres som i ligning 4.1 nedenfor:

$$\hat{Y} = \hat{f}(X), \quad (4.1)$$

hvor at  $\hat{f}$  udtrykker estimatet for den funktion, der forudsiger  $\hat{Y}$ . Som et nærtliggende eksempel kan *udbruddet af borgerkrig* ses som en funktion af den politiske utilfredshed i et land og den økonomiske vækst i et land, der i tilfældet udgør prædiktorerne. Indenfor prædiktionsmodeller behandles  $\hat{f}$  oftest som en teoretisk *black box*, i den forstand at man ikke er interesseret i kausalfortællingerne og korrelationsforholdet mellem  $\hat{Y}$  og  $\hat{f}$ , forudsat at modellen giver præcise forudsigelser af  $Y$  (James et al., 2013, s. 17).

Præcision af  $\hat{Y}$  som en prædiktion af  $Y$  er afhængig af den usikkerhed, der er forbundet med  $\hat{f}$ , da denne ikke vil være et perfekt estimat for  $f$ . En del af denne usikkerhed kan reduceres ved at vælge den med mest optimale statistiske læringsmodel til at estimere  $f$  (James et al., 2013, s. 18). Dette er grundlaget for, at jeg har valgt at introducere to forskellige *machine learning* algoritmer. Uanset hvilken algoritme, der anvendes til at prædiktere  $\hat{Y}$  vil  $\hat{f}$  aldrig blive et perfekt estimat for funktionen  $f$  der, som vist i ligning 4.2, er et udtryk for den sande sammenhæng.

$$Y = f(X) + \epsilon, \quad (4.2)$$

Hvor godt et estimat  $\hat{f}$  er af  $f$  afhænger af henholdsvis det reducerbare og det ikke-reducerbare fejllid.  $\hat{f}$  er som udgangspunkt nemlig ikke et perfekt estimat af  $f$ , og resulterer derfor i en række fejl. Disse fejl kan delvist afhjælpes ved at sikre at al relevant data inddrages i lighed med tankegangen om *omitted variable bias* (Stubager and Sønderskov, 2011, s. 10) og ved at afsøge den mest optimale algoritme og statistiske læringsmetode til at estimere  $f$  (James et al., 2013, s.18). Selvom  $\hat{f}$  var et perfekt estimat for den sande sammenhæng  $f$ , ville prædiktionen fortsat være forbundet med en række fejllid, udtrykt gennem det ikke reducerbare fejllid  $\epsilon$ . Spillerummet for at optimere en prædiktionsmodel ligger altså i inddragelse af data, valg af algoritme og justering af denne. Dette er næste afsnits genstandsfelt, der vil ligge særlig fokus på det *trade-off*, der ligger mellem en models fortolkningsmuligheder og fleksibilitet.

Eftersom at  $\hat{f}$  altid er en approksimation af  $f$  vil denne være forbundet med en ikke-reducerbar fejl. Her kan det argumenteres at denne fejl vil være større i lineære modeller, da virkeligheden sjældent er lineær (James et al., 2013, s. 82, 90)

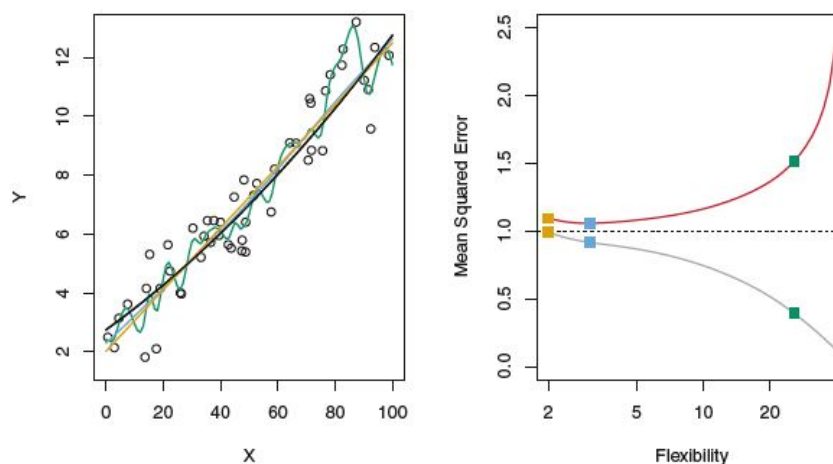
### 4.1.3. Bias-varians trade-off og krydsvalidering

Når man vælger hvilken statistisk læringsmetode eller algoritme, der skal danne  $\hat{f}$  er det ofte et *trade-off* mellem restriktive modeller og fleksible modeller. Restriktive mod-

eller er kendetegnet ved at være sat på en let fortolkelig formel, der dikterer hvordan at forholdet mellem prædiktorerne og responsvariablen kan være, som det kendes fra eksempelvis lineære modeller. Dette har sin klare fordel, når der skal skrives kausal-historier, men er ikke nødvendigvis en force ved prædiktioner. De fleksible modeller er kendetegnet ved at de tage højde for flere dimensioner i estimationen af  $\hat{f}$ , der kan give et mere nuanceret og bedre fit, for til gengæld at være svært fortolkelige i forhold til variablenes koefficienter (James et al., 2013, s. 23)

Hvor at en lineær model med få frihedsgrader ofte vil være for fast i sin form, og dermed skabe fejl grundet en lav grad af fleksibilitet, vil en fleksibel model med for mange frihedsgrader kunne risikere at *overfitte* observationerne og dermed skabe store fejl, når modellen anvendes på ny data, som illustreret i figur 6 (James et al., 2013, s. 31)

**Figur 6:** illustration af fleksibilitetens betydning for fejl



Figuren er baseret på (James et al., 2013, s. 33)

Her ses det, at den lineære model og den fleksible model med få frihedsgrader *performer* godt, mens at den fleksible model med mange frihedsgrader *overfitter*, hvilket resulterer i en stor RMSE når modellen møder nyt data. Når modellen skal fittes, er det en afvejning mellem bias og varians, hvor at varians refererer til hvor meget  $\hat{f}$  ville ændre sig, hvis denne blev dannet på baggrund af et andet datasæt, mens bias referer til den fejl, der opstår når en model forsimples komplekse forhold i virkeligheden (James et al., 2013, s. 34-35)

*Machine learning* algoritmerne bliver nødt til at foretage denne afvejning for ikke at *overfitte*. Dette gøres ved at lave en intern krydsvalidering i bygningen af modellen. Her trænes modellerne på en delmængde af træningsdatasættet, hvilket resulterer i, at modellerne tilegner sig information omkring, hvornår de begynder at overfitte datapunkterne. Der findes flere forskellige måder at lave denne krydsvalidering på, men typisk vil metoden bygge op omkring en metodik, hvor at træningssættet splittes i 5-10

*folds*, hvorfra der bygges individuelle modeller. Den endelige model bliver så resultatet af en afvejning mellem disse modellers prædiktioner, der sikrer at der tages højde for fejl og tendensen til *overfitting*, modellerne i mellem (Knox, 2018, s. 107).

Kombinationen af *machine learning* algoritmernes høje grad af fleksibilitet, og muligheden for at imødekomme tendensen til *overfitting* gennem krydsvalidering, er det metodiske argument for, at algoritmerne forventes at *performe* bedre end simple ufleksible statistiske læringsmodeller. I min analyse har jeg anvendt krydsvalidering med 5 *folds*

### 4.2. Evaluering af prædiktionsevner

Flere gange undervejs i dette speciale har jeg gjort det klart, at estimationsstudier og prædiktionsstudier adskiller sig så fundamentalt fra hinanden, og at værktøjerne i den samfundsvidenskabelige normalvidenskab med fokus på kausalitet og inferens, ikke kan bruges til at evaluere en prædiktionsmodels *performance* (King et al., 1995a; King et al., 1995b). I stedet bør en prædiktionsmodels *performance* vurderes ud fra dens evne til præcist at prædiktere responsvariablen. I klassiske regressionsanalyser fokuseres der her ofte på *root mean squared error* (RMSE) og  $R^2$ . RMSE vil tage en lille værdi, hvis vore prædikterede responsvariable er tæt forbundne med de faktiske værdier af responsvariablen.  $R^2$  værdien vil være høj, hvis modellen er i stand til at forklare en variationen i responsvariablen (James et al., 2013, s. 30-31).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2} \quad (4.3)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{f}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (4.4)$$

I prædiktionsstudier er både RMSE og  $R^2$  værdierne relevante evalueringskriterier for kontinuerte responsvariable. Det er værd her at huske på, at evalueringen af prædiktionsmodellen skal ske på baggrund af *out-of sample* dataet fra testdatasættet. Det betyder også at en prædiktionsmodel godt kan have en høj  $R^2$  værdi i det trænedes datasæt, men ikke nødvendigvis har det i testdatasættet (James et al., 2013, s. 31).

Mens *our-of sample*  $R^2$  og RMSE værdier er gode evalueringskriterier for kontinuerte responsvariable, er de ikke anvendelige for binære responsvariable. I dette speciale er et af formålene, at prædiktere *udbrudet af borgerkrig* og *eksistensen af borgerkrigslignende tilstande*, der begge er binære responsvariable.

Når man prædikterer et binært udfald, vil prædiktionsmodellerne returnere en sandsynlighed for det positive udfald ( $t = 1$ ). Det betyder at prædiktionerne ikke vil være

kategoriske i deres prædiktions af  $\hat{f}$ , men i stedet give en sandsynlighed for eksempelvis *udbruddet af borgerkrig*. For at omdanne denne sandsynlighed tilbage til et binære *outcome*, må der sættes en grænseværdi  $\delta$ . På baggrund af grænseværdien  $\delta$  klassificeres prædiktionerne således i de to grupper, f.eks. om der på baggrund af sandsynligheden prædikteres borgerkrig eller ikke-borgerkrig (Knox, 2018, s. 220, James et al., 2013, s.364). Grænseværdien kan udtrykkes som:

$$\hat{y}_i > \delta \Rightarrow \hat{y}_i = 1 \Rightarrow \text{udbrud af borgerkrig}$$

$$\hat{y}_i < \delta \Rightarrow \hat{y}_i = 0 \Rightarrow \text{ikke udbrud af borgerkrig}$$

Uanset hvordan grænseværdien  $\delta$  sættes, findes der kun 4 udfald for prædiktionen (James et al., 2013, s. 145). De to korrekte prædiktioner betegnes *sande positive* (SP), når modellen korrekt prædikterer borgerkrig, hvor der faktisk er borgerkrig og *sande negative* (SN), når modellen korrekt prædikterer "ikke-borgerkrig", hvor der ikke er borgerkrig. Omvendt er der tale om *falske negative*, når modellen prædikterer "ikke-borgerkrig", hvor der er borgerkrig, og *falske positive* når den prædikterer borgerkrig, hvor der er ikke borgerkrig. Udfaldsrummet kan skitseret i en confusion-matrix som i tabel 2

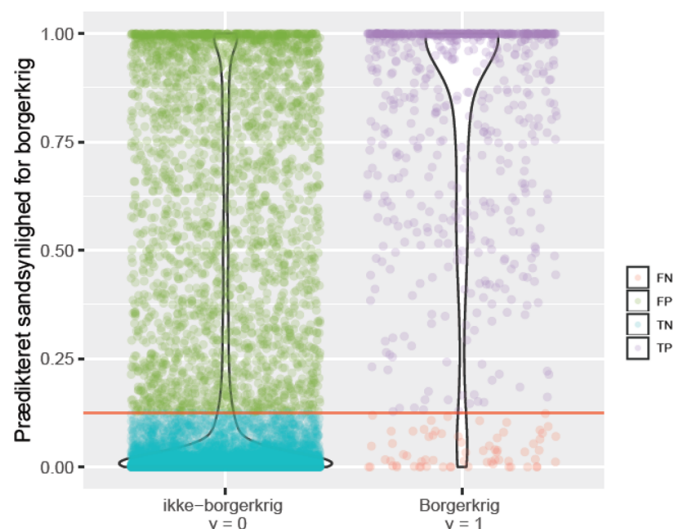
		Prædikteret respons	
		$\hat{y} = 1$	$\hat{y} = 0$
Faktiske respons	$\hat{y} = 1$	Sande positive	Falske negative (type-I fejl)
	$\hat{y} = 0$	Falske positive (type-II fejl)	Sande negative

**Tabel 2:** Confusion-matrix.

Hvordan at grænseværdien  $\delta$  skal sættes, er ikke uproblematisk for beslutningstagere i en politisk kontekst. Hvis grænseværdien  $\delta$  sættes for lavt, vil der være en større sandsynlighed for at prædiktere FP, hvorimod at en høj grænsesætning af  $\delta$  vil øge risikoen for at prædiktere FN med store konsekvenser for befolkningen, staten og stabiliteten i det internationale samfund. I figur 7 ses et eksempel på et densitetshistogram for prædiktionen af eksistensen af borgerkrig, kontra det faktiske udfald<sup>9</sup>. I eksemplet er  $\delta$  sat til 0.125, hvilket giver fordelingen mellem FP, FN, SP og SN.

<sup>9</sup>Eksemplet tager udgangspunkt i en tidlig prædiktion af eksistensen af borgerkrig, på baggrund af en logisk regression

Figur 7: Densitetshistogram over prædiktions typer - opstillet eksempel



Som det ses af figuren, er der i eksemplet en forholdsvis fornuftig fordeling mellem de prædikterede sandsynligheder og de faktiske udfald. Overordnet set prædikteres der lave sandsynligheder for eksistensen af borgerkrige i de tilfælde, hvor der ikke har været borgerkrige og højere sandsynligheder i de tilfælde, hvor der faktisk var borgerkrig. For begge grupper gælder det dog, at en ikke væsentlig del af prædiktionerne er henholdsvis FN og FP. Ved at ændre på grænseværdien  $\delta$  er det muligt at justere forholdet mellem sande og falske prædiktioner og dermed optimere prædiktionen efter formålet.

Hvis det internationale samfund skal kunne bruge prædiktionerne som beslutningsgrundlag, gælder det først og fremmest om at reducere mængden af FN, når  $\delta$  skal sættes. I eksemplet i figur 7, vil en lavere sætning af  $\delta$  betyde en voldsom forøgelse af mængden af FP. Hvis det internationale samfund skulle igangsætte policy tiltag på baggrund af alle disse fejlagtige prædiktioner, ville det blive meget omkostningsfuldt og formentlig på bekostning af indsatser målrettet de *sande positive*. Der er med andre ord et centralt *trade-off*, når grænseværdien skal sættes og det er blandt andet derfor det er vigtigt at finde frem til den mest optimale prædiktionsmodel, for at reducere de omkostninger, der er forbundet med dette *trade-off*.

Værdisætningen af  $\delta$  og balancegangen mellem de fire prædiktions typer er i høj grad et politisk og normativt spørgsmål, der er svært at give et objektivi korrekt svar på. En måde hvorpå, at man kan forsøge at tilnærme sig en optimal sætning af grænseværdien, er ved at se på forholdet mellem sand-positiv-raten (SPR) og falsk-positiv-raten (FPR). Terminologien er indenfor den statiske læring ikke skarp og SPR benævnes derfor også modellens *sensivitet* eller *power*, mens FPR også benævnes 1-modellens specificitet



(James et al., 2013, s. 149). Uagtet den upræcise termologi kan de to forholdstal matematisk udtrykkes således:

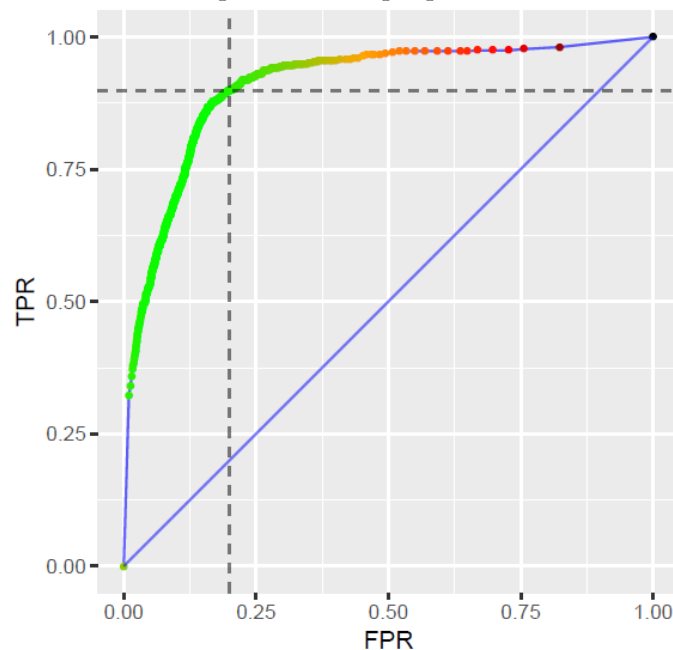
$$\text{Sand positiv rate} = \frac{\text{Sande positive}}{\text{Sande positive} + \text{Falske negative}} = \frac{\text{Sande positive}}{\sum y_i = 1} = \text{SPR} \quad (4.5)$$

$$\text{Falsk positiv rate} = \frac{\text{Falske positive}}{\text{Falske positive} + \text{Sande negative}} = \frac{\text{Sande positive}}{\sum y_i = 0} = \text{FPR} \quad (4.6)$$

Statistisk set er det muligt at optimere forholdet mellem SPR og FPR ud fra en funktion, der udtrykker optimale mellem de to forhold. At udarbejde en sådan funktion på en meningsfuld måde, ville kræve et dybt kendskab til økonomien i internationale interventioner - bløde såvel som hårdere - opstillet mod de menneskelige og samfundsmæssige omkostninger, der er forbundet med borgerkrig.

Frem for at forsøge at lave en sådan funktion, har jeg valgt at tage det normative standpunkt, at verden ikke er tjent med borgerkrig og at prædiktionen af FN skal straffes hårdere end FP. Gennemgående i min analyse har jeg sat omkostningen ved at prædiktere FN til den arbitrære værdi 10 mod en omkostning på 1 ved at prædiktere FP. Tallene udtrykker dermed at det i sætningen af  $\delta$  skal være 10 gange så "dyrt" at lave en type-I fejl som at lave en type-II fejl. Forholdet mellem SPR og FPR kan grafisk illustreres med en *Receiver Operating Characteristic* kurve (James et al., 2013, s. 365-366), der plotter forholdet mellem FPR og SPR ved alle  $\delta$ . Dette eksemplificeret i figur 8.

**Figur 8:** Opstillet eksempel på ROC-kurve



ROC-kurven afspejler den afvejning der er mellem at maksimere SPR, og dermed i eksemplet indfange alle borgerkrige og mængden af type-II fejl. Som sagt er målet i min undersøgelse at afsøge potentialet for *machine learning* prædiktionsmodeller og dermed have en SPR så tæt op 1, uden at FPR stiger markant. I eksemplet ses det at  $\delta = 0,125$  fører til en SPR på omkring 0,9 og en FPR på omtrent 0,2. Den ideelle position for en ROC-kurve er en visuel tilnærmelse af det øvre venstre hjørne. For at måle denne tilnærmelse matematisk, kan man udregne arealet under ROC-kurven, det såkaldte *area under curve* (AUC-værdien) (James et al., 2013, s. 147). I eksemplet er AUC-værdien 0,94, hvilket er ganske tæt på maksiværdien 1.

AUC-værdien kan ses som et godt sammenligningsgrundlag for forskellige prædiktionsmodeller, da den for det første udtrykker *performancen* for hvert enkelt model ved alle mulige grænseværdier af  $\delta$ , samtidig med at der gives informationer om forholdet mellem SP og FP, der må ses som de vigtigste udfald for selve *performancen* (James et al., 2013, s. 147). En AUC-værdi på 0.5 er et udtryk for en klassifikation af udfaldende, der ikke er bedre, end hvis klassifikationen var fuldstændig tilfældig. AUC-værdien giver samtidig mulighed for sammenligning mellem modeller, da værdien udtrykker alle mulige grænsesætninger af  $\delta$  og sammenligningen af modeller er derfor ikke afhængig af, hvor netop  $\delta$  er sat for den enkelte model.

I dette afsnit har jeg argumenteret for, at *out-of sample* RMSE og  $R^2$  værdier tjener som gode evalueringskriterier for kontinuerte responsvariable. For binære responsvariable har jeg introduceret confusion-matrix, densitets-histogrammer, ROC-kurver og AUC-værdier som evalueringsgrundlag for prædiktionsmodellernes performance. Alle evalueringsmetoder, der ser nærmere på den eksterne validitet af prædiktionsmodellerne. I det følgende afsnit vil jeg belyse de algoritmer, jeg anvender i analysen.

### 4.3. Valg af algoritmer

I ovenstående afsnit, har jeg gennemgået de brede koncepter for både *machine learning* og prædiktionsstudier. I dette afsnit vil jeg beskrive de 2 *machine learning* algoritmer, som jeg vil teste op mod baselinemodellerne, der udgøres af henholdsvis en logistisk og en lineær regressionsmodel. Som det blev skitseret i afsnit 2, findes der ikke én model, der definitivt kan kåres til den bedste og derfor er det nødvendigt at teste flere algoritmer (James et al., 2013, s. 29)

De to algoritmer jeg vil introducere, er *Random Forest* og *Extreme Gradient Boosted Tree* (EGBtree). Begge algoritmer er såkaldte ensemble metoder, der bygger deres forudsigelser på gennemsnittet af en lang række klassifikationstræer (James et al., 2013, s. 303). Da klassifikationstræerne er byggestenene for begge algoritmer, vil jeg i følgende afsnit først introducere disse, for herefter at belyse de to algoritmernes særpræg.

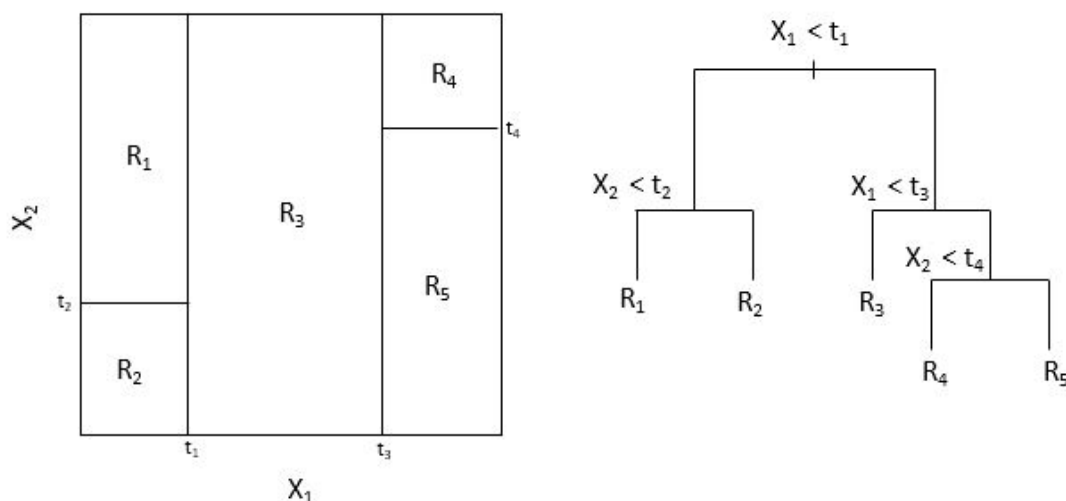
### 4.3.1. Klassifikationstræer

Klassifikationstræer og de såkaldte træbaserede algoritmer har vundet terræn i den samfundsvidenskabelige forskning i de senere år, grundet algoritmes intuitive fortolkning og deres iboende fleksibilitet i modelleringen af data (James et al., 2013, s. 303). Helt generelt fungerer klassifikationstræs algoritmen, ved at segmentere observationerne i et datasæt, på baggrund af en række sand/falsk spørgsmål om de uafhængige variable.

I sin essens forsøger klassifikationstræet at stille de spørgsmål til de uafhængige variable, der segmenterer dataet således, at hver gruppe, afledt af spørgsmålet, danner den mest ensartede fordeling af responsvariablen. F.eks. kunne et spørgsmål til forudsigelsen af *udbrud af borgerkrig* være, om der havde været over 25 kamprelaterede døde i den foregående måned. Hvis udsagnet er sandt, placeres observationen i den ene gruppe, mens den placeres i den anden gruppe, hvis udsagnet er falsk. Når dette spørgsmål er stillet, vil det næste spørgsmål stilles, der igen vil forsøge at skabe den optimale inddeling af responsvariablen (James et al., 2013, s. 304).

Den første segmentering, altså den første opdeling af data, betinger de følgende segmenteringer og er altså styrende for, hvordan senere segmenteringer i datasættet skal foregå. Dette kaldes en *grådig* segmentering. Segmenteringen, der lag efter lag forsøger at optimere opdelingen, resulterer i, at modellen ikke bygger på én konkret antagelse om modellens form, f.eks. en lineær sammenhæng, men i stedet giver rum til modelkompleksitet, som belyst i afsnit 4.1.3. Den segmentering af data, som klassifikationstræerne bygger på, er illustreret i figur 9

**Figur 9:** Logikken bag segmenteringen i klassifikationstræer



(a) Segmenteringen af datasættet  
Inspireret af (James et al., 2013, s. 22)

(b) Klassifikationstræet

I eksemplet kunne det første split på  $X_1$  eksempelvis være om der var over eller under 25 døde i foregående måned. Hvis denne segmentering ikke opdeler responsvariablen optimalt, vil det give anledning til endnu en segmentering. F.eks. om antallet af kamphandlinger i den foregående måned er større end 0 ( $X_1 > 0$ ). hvilket ville resultere i de to regioner  $R_1$  &  $R_2$ , på baggrund af klassifikationstræets venstreben.

I klassifikationstræets højre ben, kunne man segmentere videre på  $X_1$ , for en yderligere nuancering inden at modellen segmenterede på  $X_2$ . I klassifikationstræet terminologien resulterer hver segmentering i yderligere 2 grene, mens at segmenter, der ikke opdeles yderligere benævnes blade. Klassifikationen sker efter en *top-down*, binær grådig tilgang. Klassifikationen er *top-down*, da den første segmentering er afgørende for de senere split; binær da der kun groes 2 grene af gangen, og grådig da modellen vælger den segmentering, der er bedst på det pågældende niveau, uden omtanke for at dette kan føre til dårligere segmenteringer senere hen (James et al., 2013, s. 306)

Når man skal *fitte* et klassifikationstræ til et datasæt og dermed danne en model, foregår det på baggrund at *minimere* forskellen mellem de faktiske *outcomes*  $y_i$  og de prædikterede *outcomes*  $\hat{y}_i$  for hver segmentering af variabel  $j$  til alle værdier variabelen ( $s$ ) (James et al., 2013, s. 307). For hver segmentering kan denne *loss-funktion*, der ønskes at minimeres udtrykkes som:

$$Tree_{loss} = \sum_{x_i \in R_1(j,s)} \ell(y_i, \hat{y}_{R_1}) + \sum_{x_i \in R_2(j,s)} \ell(y_i, \hat{y}_{R_2}) \quad (4.7)$$

Når man gror klassifikationstræer, er det med fare for at *overfitte* data, da et klassifikationstræ kan blive ved med at segmentere observationerne, indtil det klassificerer samtlige observationer korrekt. Som diskuteret vil dette give gode resultater *in-sample*, men formentlig en dårlig *out-of sample performance*. For at undgå dette er det nødvendigt at regularisere sit klassifikationstræ, det vil sige styre kompleksiteten og tendensen til *overfitting*. Det kan gøres ved at beskære klassifikationstræet i forhold til hvor dybt det må gro, et mindste antal observationer i hvert blad og samt en indregnet omkostning ved høj kompleksitet (James et al., 2013, s. 307-311)

Med denne præsentation af klassifikationstræer, der er byggestenen i de udvalgte *machine learning* algoritmer, er vi nu klar til at få præsenteret algoritmerne.

### 4.3.2. Random Forest

Random Forest er på mange måder den naturlige forlængelse af klassifikationstræet som algoritme. Fremfor at gro ét enkelt klassifikationstræ, groes der i stedet et ensemble af træer, der hver især fittes til data. I al sin simpelhed tager Random Forest algoritmen gennemsnittet af hvert klassifikationstræs prædiktioner for den enkelte observation og skaber dermed sin egen algoritme (James et al., 2013, s. 319). Denne tilgang bør sikre,

at hvis et klassifikationstræ laver en dårlig segmentering på et parameter, vil det blive opvejet af de andre klassifikationstræer.

Foruden at være en forlængelse af klassifikationstræet som algoritme, understøttes Random Forest af en resampling teknikkene *bootstrap aggregating* (Bagging), der fungerer som modellens interne krydsvalidering (Bühlmann and Yu, 2000, s. 2). Bagging-metodikken resampler datasættet på to parametre.

For det første benyttes der kun en delmængde af træningsdatasættet, hver gang at der groes et klassifikationstræ. Som belyst i afsnit 4.1.3, er dette en ekstra instans mod *overfitting* på træningsdatasættet. For det andet udvælges kun en tilfældig delmængde af de uafhængige variable til at gro klassifikationstræerne. Dette har den værdi, at variablenes effekt på prædiktionen bliver mere stabile og mindre afhængig af, præcis hvilket data der bruges i træningsdatasættet. Samtidig åbner *bagging* op for muligheden for at vurdere hvilke uafhængige variable, der har den største effekt på prædiktionen og dermed også usikkerheden, der er forbundet med  $\hat{f}$  (James et al., 2013, s. 187)

Den klare ulempe ved Random Forest er, at når man gror flere hundrede klassifikationstræer, er det på samme måde som i det illustrerede eksempel i figur 9 ikke let at fortolke vejen til prædiktionen, da det er svært at gennemskue gennemsnittet af flere 100 klassifikationer. Som Hal. R Varian forklarer det:

One defect of random forests is that they are a bit of a black box—they don't offer simple summaries of relationships in the data. As we have seen earlier, a single tree can offer some insight about how predictors interact. But a forest of a thousand trees cannot be easily interpreted. However, random forests can determine which variables are "important" in predictions in the sense of contributing the biggest improvements in prediction accuracy. (Varian, 2014, s. 15)

Forventningen til Random Forest algoritmen er, at dens opbygning gennem klassifikationstræer sikrer en fleksibilitet til modellens form, samtidig med at *bagging* reducerer risikoen for overfitting til træningsdatasættet.

#### 4.4. Extreme Gradient Boosted Trees

Den anden algoritme jeg vil introducere, er Extreme Gradient Boosted Trees (EGBtree), der ligesom Random Forest algoritmen er en ensemble algoritme, der baserer sig på klassifikationstræer. Jeg har valgt denne algoritme over andre, da det i litteraturen har vist sig at *tuning* af algoritmerne ofte er frugtbart (Perry, 2013) samtidig med at modellen har vundet flere *case-competitions* på Kaggle (Gorman, 2017).

Til forskel fra Random Forest algoritmen, der benyttede *bagging* i den interne

krydsvalidering, benytter EGBtree sig af *boosting*. Begrebet *boosting*, der dækker over en proces, hvor at klassifikationstræerne groes sekventielt. For hvert træ der groes tillægges de observationer, der klassificeres forkert, en større betydning i det følgende klassifikationstræ (James et al., 2013, s. 321). Ulig Random Forest bruges der altså ikke nødvendigvis et subsample af observationerne, når klassifikationstræet udregnes. I stedet bruges information fra det foregående klassifikationstræ til at gro det efterfølgende klassifikationstræ. Det betyder også, at EGBtree modellen har en tendens til at overfitte data, hvis at antallet af klassifikationstræer bliver for stort (James et al., 2013, s. 322)

Ved boosting er der grundlæggende tre parametre, der kan justeres eller *turnes*. Det første af antallet af træer, der skal groes. Som udgangspunkt bør antallet sættes lavere end ved Random Forest, da mange træer kan føre til overfitting. Det andet parameter er *shrinkage* parameteret  $\lambda$ .  $\lambda$  kan tolkes som det parameter, der styrer hvor stor en vægt fejl klassificerede observationer skal tillægges i det følgende klassifikationstræ.

Når  $\lambda$  skal sættes er det en afvejning med antallet af klassifikationstræer, da en meget lav  $\lambda$  vil betyde en lille flad læringskurve, der kræver flere træer. Det sidste tunings-parameter er antallet af segmenteringer i hvert klassifikationstræer. Her er dybe træer ikke nødvendigvis en fordel, da algoritmen netop lærer på baggrund af de fejl klassifikation den løbende foretager (James et al., 2013, s. 322)

For kort at opsummere, adskiller EBGtree algoritmen sig ved at fejlklassifikationer i klassifikationstræerne gives en øget betydning i en iterativ proces, der til sidst ender ud i en samlet model, der har forsøgt at lære af sine egne fejl og på baggrund af dette give disse give den bedste prædiktionsmodel (Varian, 2014, s. 14)

---

## 5. DATA

I dette kapitel, vil jeg introducere mit samlede datasæt med fokus på de variable, jeg ikke har berørt gennem teori-redegørelsen i kapitel 3. Her vil jeg først beskrive de fire forhold omkring borgerkrig, der udgør mine responsvariable. Herefter vil jeg vende opmærksomheden mod det datasæt over politiske begivenheder, som jeg på baggrund af GDELT-databasen introducerer til litteraturen. I forlængelse af denne introduktion vil jeg diskutere nyhedsdataets validitet som datakilde. Afslutningsvist vil der være en oversigt over det samlede datasæt, samt en kort note om behandlingen af manglende observationer i datasættet.

Datasættet er konstrueret således, at hver observation udgør en måned i et land. Datasættet er tidsmæssigt afgrænset til perioden 1989-2017. Denne grænse er trukket på baggrund af datasættet fra Uppsala Conflict Data Program, der kun indeholder information om konflikter fra 1989 og frem (Sundberg and Melander, 2013).

### 5.1. Responsvariablene

Som kernen i analysens prædiktionsmodeller står *responsvariablene*, der udgør de fire forhold for borgerkrig, som jeg vil prædiktere. Alle responsvariablene tager deres udgangspunkt i Uppsala Conflict Data Programs datasæt *Georeferenced Event Dataset*. Datasættet består af 142902 enkeltstående tilfælde af organiseret vold (Sundberg and Melander, 2013; Croicu and Sundberg, 2017). I datasættet udgør hver observation én kamphandling med beskrivelse af kamphandlingens aktører, antallet af døde<sup>10</sup>, samt et geografisk og et tidsligt stempel, så at observationen kan placeres rumligt og temporalt.

For at fokusere dette datasæt på intrastatslige uroligheder og borgerkrige, har jeg frasorteret interstatslige konflikttilfælde og konflikttilfælde, hvor ingen af parterne har en direkte tilknytning til en stat. Det reducerede datasæt er efterfølgende grupperet og summeret på lande-måned niveau. Hver observationer udtrykker således det samlede antal af døde<sup>11</sup> og antallet af kamphandlinger i et land en given måned.

Dette grupperede konfliktdatasæt udgør rammen for de fire responsvariable. De første to responsvariable er binære variable, hvor den ene udtrykker om der udbryder borgerkrig, mens den anden udtrykker om der eksisterer borgerkrigslignende tilstande i et land. De to øvrige responsvariable er kontinuerte og udtrykker konflikternes intensitet gennem henholdsvis summen af antallet af døde og antallet af kamphandlinger en given måned.

---

<sup>10</sup>Informationen er opdelt, således at det kan udledes, hvor mange dødsfald der tilfalder hver part samt antallet af civile ofre

<sup>11</sup>igen underopdelt således at der er et dødstal for begge parter og civile

Mens de to kontinuerte variable ikke behøver en yderligere operationalisering, forholder det sig anderledes med de to binære responsvariable. Her er der behov for en målbar definition af borgerkrig i forhold til intensiteten af den interne konflikt. Dette er nødvendigt for at skelne enkeltstående voldelige konflikter og terrorist-angreb fra borgerkrig. I litteraturen findes der en håndfuld *data definitioner* på borgerkrig, f. eks.:

1. 1000 døde i en intern konflikt et givent år (Collier and Hoeffler, 2004)
2. 1000 døde i en intern konflikt et givent år, hvor at der er mindst 100 døde på hver side (Fearon and Laitin, 2003)
3. 25 kamprelaterede døde et givent år (Soysa and Neumayer, 2007)

Med udgangspunkt i disse definitioner har jeg valgt at definere borgerkrig som en et år, hvor at der er 1000 kamprelaterede døde i interne konflikter, og hvor mindst 200 af de døde kan relateres til en af de kæmpende aktører. Argumentet for at definere borgerkrige ud fra 1000 døde ligger i, at der skal være en meget alvorlig tilstand af intern konflikt, før at de indledende overvejelser omkring borgerkrig gør sig gældende. Samtidig er der et behov for at tilskrive en andel af de døde til konfliktens parter. Hvis ikke man gør dette, vil folkemord og voldsomme terroristangreb optræde på listen over borgerkrig. Begge fænomener er relevante prædiktionsproblemer, men ligger uden for dette speciales genstandsfelt.

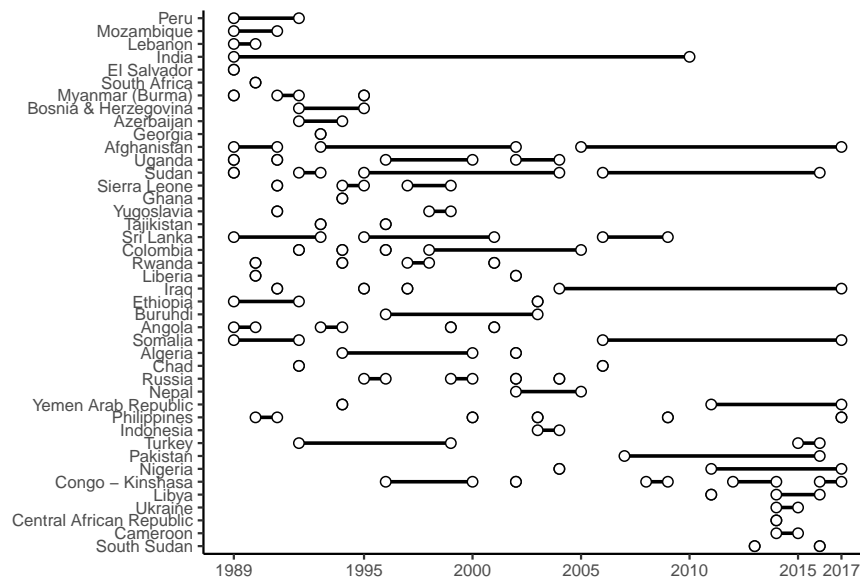
For at applicere denne definition af borgerkrig på mit datasæt, har jeg transformeret definitionen til måneds-niveau, der også udgør definitionen for responsvariablen "borgerkrigslignende tilstand". Her vil der være tale om *borgerkrigslignende tilstande*, hvis der er 83 døde, hvoraf 16 kan relateres til en af konfliktens aktører en given måned.

Responsvariablen "Udbrud af borgerkrig" bliver på baggrund heraf, den første måned med *borgerkrigslignende tilstande* et givent år, hvor at kriteriet for borgerkrig på årsniveau bliver opnået. Hvis kriterierne er opfyldt i det efterfølgende år, er der ikke tale om et nyt tilfælde af borgerkrig men i stedet for en sammenhængende borgerkrig. På figur 10 ses en oversigt over de identificerede borgerkrige og deres udstrækning.

I oversigten identificerer jeg 91 borgerkrige i perioden. Ud fra datadefinitionen og operationaliseringen af borgerkrig er der således ganske få tilfælde og udbrud af borgerkrig af prædiktere. Det måske mest i øjenfaldende ved figur 10 er, at konflikten i Syrien ikke optræder. Dette skyldes ikke, at konflikten falder uden for datadefinitionen, men derimod at Syrien er det eneste land, der ikke indgår i *Georeferenced Event Dataset* (Croicu and Sundberg, 2017, s. 2).



**Figur 10:** Oversigt over borgerkrige og borgerkrigs-lignende uroligheder



Dette meget lave antal *udbrud af borgerkrig*, vil formentlig umuliggøre en prædiktionsmodel af responsvariablen, da der er meget få tilfælde som modellerne kan lære at genkende mønstre ud fra. Dette er et af de centrale argumenter for, at at nuancere prædiktionsmodellen med responsvariablen *eksistensen af borgerkrigs-lignende tilstande*. Denne responsvariable har 4697 "positive" tilfælde, og indfanger både måneder med sporadiske alvorlige interne konflikter, udbrud af borgerkrig samt den samlede borgerkrigsperiode.

For kort at vende tilbage til de to kontinuerte responsvariable *antal døde* og *antal kamphandlinger*, tjener de til formål at måle den nærmere intensitet af konflikterne, der ellers går tabt i de to binære responsvariable. Begge variable kan direkte udledes af det grupperede og summerede *Georeferenced Event Dataset*. Formålet med variablene er, at viden om intensiteten af borgerkrig kan guide allokering af ressourcer i et indgreb og dermed gøre policy-indgreb mere målrettet og dimensioneret til den faktiske konflikt.

Variationer af responsvariablene indgår også som uafhængige variable i prædiktionsmodellerne. F.eks. bruges det gennemsnitlige antal døde i de tre forgående måneder i prædiktionsmodellen af responsvariablene.

## 5.2. Det politiske landskab: Politisk tilfredshed og utilfredshed

Som introduceret i kapitel 3, står politisk utilfredshed som et centralt motiv for borgerkrig i teorien og særligt i *grievance* tilgangen. I litteraturen har politisk utilfredshed været operationaliseret på et meget strukturelt niveau. Her har antagelsen været, at politisk utilfredshed vil være mere sandsynligt i lande, der er etnisk og religiøst fraktioniseret. Dette er en markant antagelse, som jeg har forsøgt at gøre op med,

ved at fokusere på faktiske mål for politisk utilfredshed og tilfredshed. Her har jeg konstrueret et datasæt, der gennem en analyse af nyhedsartikler udtrykker forskellige politiske forhold og som samlet datasæt udgør en proxy for det *aktuelle politiske landskab*.

At konstruere et datasæt, der måned for måned og fra land til land, beskriver det politiske landskab ved at indkapsle den politiske tilfredshed og utilfredshed, er en enorm proces. Jeg har valgt at tilgå dannelsen af et sådan datasæt med udgangspunkt i det verdensomspændende projekt Global Database of Events, Language and Tone (GDELТ) (Leetaru and Schrodt, 2013). GDELТ-databasen består af over 250.000.000 millioner observationer, der hver indeholder relevant metadata for én enkelt politisk nyhedsartikel i perioden 1979 og frem<sup>12</sup>. I konstruktionen af datasættet over det politiske landskab, har jeg med udgangspunkt i dette data, fokuseret på typen af de politiske begivenheder, hvem aktørerne er i de politiske begivenheder og hvordan at tonen er i den nyhedsartikler, der beskriver de politiske begivenheder. For kort at introducere GDELТ beskriver projektet sig selv som:

”The GDELТ Project is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what’s happening around the world, what its context is and who’s involved, and how the world is feeling about it, every single day” (GDELТ, 2018).

Datasættet er altså skabt ud af en ambition om at kortlægge begivenheder af høj samfundsmæssig betydning, hvem der er involveret og hvordan disse begivenheder opleves. Inden jeg beskriver de konstruerede variable, er det oplagt at spørge sig selv, *hvordan at nyhedsartikler kan udtrykke det politiske landskab?*

Her er det korte svar, at nyhedsartikler over politiske begivenheder netop er et udtryk for det politiske landskab herunder den politiske tilfredshed og utilfredshed i et land, da det er en afrapportering af faktiske politiske begivenheder. F.eks. er en nyhedsartikel omhandlende en demonstration et udtryk for den politiske utilfredshed, der er i den deltagende gruppe. En anden nyhedsartikel, der beskriver politiets voldelige opløsning af en demonstration ses som en manglende politisk anerkendelse af demonstranternes sag. Som modsætning er en nyhedsartikel, der beskriver en aftale mellem opposition og regering, et udtryk for politisk samarbejde.

Hvis man fokuserer på den politiske utilfredshed, der teoretisk står centralt i dette speciale, er der en fare ved at læne sig blindt på af nyhedsdata. Det skyldes at man udelukkende vil indfange den aktørbaserede politiske utilfredshed, der eksplicit kommer til udtryk. Dermed indfanges den latente politiske utilfredshed ikke. Omvendt

---

<sup>12</sup>Da datasættet har en vis teknisk karakter i dets opbygning og har central rolle i dette speciale, har jeg beskrevet processerne og de kilder, der har dannet grundlaget for datasættet i Appendix B

kan man argumentere for, at politisk utilfredshed eller tilfredshed, der ikke kommer eksplicit til udtryk, ikke har en karakter, der er værd at inddrage som mål, grundet dennes udtalte karakter. Nyhedsdatatet kan altså en indgang til en informations rigdom om udtalte politiske begivenheder.

### 5.2.1. De konstruerede variable

Som sagt har jeg med udgangspunkt i GDELT-databasen 250.000.000 observationer af politiske begivenheder valgt at fokusere på *typen af politiske begivenheder og aktørerne i de politiske begivenheder*.

I GDELT-databasen optræder over 300 forskellige typer af politiske begivenheder, der strækker sig fra alt mellem udtalelser, fredsmægling, appel til det internationale samfund, demonstrationer og vold (GDELT, 2013). En oversigt over de forskellige typer findes i appendiks D. Da inddragelsen af alle disse 300 forskellige typer af politiske begivenheder, vil resultere i et *information overload* i mine prædiktionsmodeller, har jeg foretaget en aggregering af typerne. Med udgangspunkt i GDELT-databasen datastruktur, har jeg således inddelt de 300 begivenhedstyper i 4 kategorier. Hver kategori er således en samling af de begivenhedstyper der udtrykker henholdsvis *verbalt samarbejde, verbalt modarbejde, materielt samarbejde og materielt modarbejde*.

Det næste skridt har herefter været at se, hvor mange begivenheder inden for hver af disse grupperinger, der kan tilskrives en bestemt samfundsgruppe. Som skildret i *grievance* litteraturen tillægges etnicitet og religiøse forhold en stor betydning. På baggrund af denne teoretiske forventning om, at etnicitet og religion kan være udgangspunktet for politik, har jeg isoleret de begivenheder, hvor at en af aktørerne i handlingen har et udtalt religiøst eller etnisk tilhørsforhold. Dette har skabt tre aktørgrupper: Etniske aktører, religiøse aktører og "resten" Min grundlæggende tanke med at danne denne række af variable har været, at det dermed i højere grad er muligt at spore politiske begivenheder, der favoriserer eller underkender bestemte etniske eller religiøse grupperinger. Politiske begivenheder der kan forventes at skabe skillelinjer i samfundet, og understøtte en følelse af uretfærdighed i den u-favoriserede gruppe.

For hver af de tre aktørgrupper, findes der således 4 begivenhedstyper og det sidste element har således været at danne meningsfulde indikatorer inden for disse. Her har jeg dannet 3 variable. Den er første variable tæller antallet af politiske begivenheder for hver gruppe/kategori på lande-måned niveau og må ses som et simpelt mål for den politiske aktivitet.

Den anden variable er et mål for den gennemsnitlige tone, i nyhedsartiklernes beskrivelse af begivenhederne<sup>13</sup>. Indikatoren er dannet ved at fratrække mængden af

<sup>13</sup>variablen er udregnet i GDELT projektet og som beskrevet i appendiks B

positive konnotationer fra mængden af negative konnotationer. Formålet med denne variable er, at der formentlig vil være forskel i tonen af politiske begivenheder alt efter hvor sympatiske eller usympatiske disse fremstår, hvilket kan hjælpe med at afdække opbakningen til en politisk handling.

Det sidste mål er den gennemsnitlige *goldstein-scale* for begivenhederne inden for gruppen. *Goldstein-scalen* dækker over den forventede betydning en begivenhed har for den politiske stabilitet i et land (GDELT, 2013, s. 5). Variablen er altså en proxy for den stabiliserende eller destabiliserende betydning en begivenhed har for et land.

Som en *feature engineering* af variablene, er der for hvert variabel udregnet det rullende gennemsnit for de forgående 6 måneder. Resultatet er således 18 variable for hver af de 4 kategorier af politiske begivenheder. I appendix C ses det samlede SQL-statement der, på baggrund af det indhentede data, danner det samlede datasæt over det politiske landskab<sup>14</sup>

### 5.3. Validiteten af nyhedsdata som datakilde

Anvendelsen af nyhedsdata i akademisk arbejde rejser naturligvis en række spørgsmål om dataets validitet og anvendelighed. I dette afsnit vil jeg først genbesøge spørgsmålet om nyhedsdataets validitet i lyset af analysens resultater. Som en afledt diskussion her vil jeg perspektivere nyhedsdataets potentielle værdi i forhold til andre aspekter af borgerkrig og samfundsvidenskabelige spørgsmål.

Når man arbejder med nyhedsdata som datakilde det med en naturlig skepsis. *For hvordan kan vi være sikker på, at nyhedsartiklerne formidler begivenhederne objektivt og ikke er biased? Og kan vi overhovedet være sikker på, at alle relevante begivenheder bliver rapporteret som nyhedsartikler og ikke er underlagt censur?* Det korte svar til disse spørgsmål er, at *det kan vi ikke*, men at man ved at kortlægge præmisserne for dataet kan være opmærksomme på de fejlkilder, der kan være forbundne med dataet.

Når man arbejder med data af enhver art, er dataets validitet genstand for megen opmærksomhed. Når det kommer til data, der bygger på nyhedsdata kan man spørge sig selv om hvilken gyldighed dataet har og om nyhedsartiklerne er sandfærdige. GDELT-databasen forsøger, som beskrevet i appendix B i så vidt muligt omfang at undgå biased nyhedsdata, ved at basere sig på nyhedsartikler fra store internationale nyheds beureauer<sup>15</sup> og dermed undgå partipolitiske og "farvede" nyhedsmedier. Selvom dette er et skridt på vejen kan der fortsat stilles spørgsmålstejn ved nyhedsartiklernes

---

<sup>14</sup>Forbehandlingen og indhentningen af data ses i R-filen "Dataload.R" og "DataGather.R"

<sup>15</sup>Africa News, Agence France Presse, Associated Press, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, The New York Times, United Press International and The Washington Post (GDELT, 2018)

objektivitet. Frem for at anskue dette som en problematik kan man anerkende denne potentielle kilde til fejl og i stedet se det som et vindue, hvor i gennem man kan få en indblik i den informations rigdom, der ligger i kulturen og kontekstens påvirkning den nationale bevidsthed.

Dette syn på nyhedsartiklernes værdi som data bliver hjulpet på vej af den analyse af artiklerne som ligger til grund for GDELT-databasen. Med undtagelse af målet for den tone, der er i nyhedsartiklen, bygger GDELT-databasen på de objektive oplysninger i nyhedsartiklerne frem for den vinkel nyhedsartiklen som nyhedsmedieret lægger ned over historien. I GDELT-databasen lægges der med andre ord vægt på, hvem aktørerne er og hvad handlingen er, fremfor hvordan at disse skildres i artiklen.

#### 5.4. Behandling af manglende data

Når man arbejder med data fra forskellige kilder løber man næsten altid ind i problemer med ufuldstændige observationer, hvor der mangler observationer på enkelte variable. Konsekvensen af én manglende værdi i en observation er, at hele observationen må udgå af analysen (W Beck et al., 2018). Dataet i dette speciale har ikke været en undtagelse.

I samfundsvidenskaben afværges dette problem til tider med *data-imputation*, hvor man gennem forskellige metoder giver et bud på, hvad værdien af den manglende værdi er og dermed beholder observationen i analysen. Eksklusionen af relevante variable grundet manglende variable er ikke at foretrække, da det vil føre til *omitted variable bias*, der i sidste ende påvirker de estimerede parametre og prædiktionernes *performance* (Stubager and Sønderkov, 2011, s.10). Ulempen ved at *imputere* værdier er, at det ikke er sande observerede værdier, men i stedet forventede værdier eller det bedste bud på værdien. Fejlen der vil være forbundet med *imputationen* vil resonere ned gennem analysen, og i sidste ende behæfte analysen med flere fejl end den afhjælper. Ikke desto mindre kan det argumenteres for at *data-imputation* er et nødvendigt onde, der kan give gode estimater for de manglende værdier på (W Beck et al., 2018, s.223).

Jeg har valgt at begrænse min brug af *imputation* til få strukturelle variable<sup>16</sup> og benyttet henholdsvis lineær-imputation for de variable hvor der forventes en lineær udvikling og *last-observation-carried-forward* metode, der hvor den sidste observation er det bedste bud. For resten af variablene har jeg accepteret det tab i antal i observationer, der er forbundet med manglende værdier. I tabel 3 ses en oversigt over variablene i datasættet.

<sup>16</sup>Befolkningstal, terrænet i et land, indskrivning i skole og dyrkbarjord

**Tabel 3:** Oversigt over data

Variabel	Type	Kategorier	Antal
<b>Konfliktdata</b>			
Døde i konflikt	Kontinuert		5
Døde i konflikt (side a)	Kontinuert		1
Døde i konflikt (side b)	Kontinuert		1
Antal kamphandlinger	Kontinuert		4
Udbrud af borgerkrig	Kategorisk	Ja/nej	1
borgerkrigslignende tilstand	Kategorisk	Ja/nej	1
<b>Politiske begivenheder</b>			
Materielt samarbejde	Kontinuert		18
Materielt modarbejde	Kontinuert		18
Verbalt samarbejde	Kontinuert		18
Verbalt modarbejde	Kontinuert		18
<b>Styreform</b>			
Demokrati-score	Kontinuert		1
Autokrati-score	Kontinuert		1
Polity 2 score	Kontinuert		1
Regulering af valg	Kategorisk	Ureguleret, transition, reguleret	1
Konkurrence ved valg	Kategorisk	Udpegning, udpegning og valg, valg	1
Åbenhed ved valg	Kategorisk	lukket, lukket med udpegning, lukket med valg, åben	1
Begrænsninger på udøvende magt	Kategorisk	7 kategorier af begrænsninger	1
Historisk koloniseret	Kategorisk	Nej, britisk, fransk, andet kolonimagt	1
<b>Vækst og handel</b>			
BNP pr capita	Kontinuert		1
log(BNP)	Kontinuert		1
Vækst	Kontinuert		2
Handels andel af BNP	Kontinuert		1
import af <i>good and services</i>	Kontinuert		1
<b>Gruppe data</b>			
# Politisk relevante grupper	Kontinuert		5
Politiske gruppers befolkningsandel	Kontinuert		4
# Etniske flygtninge	Kontinuert		1
Etnisk fraktionisering	Kontinuert		1
Relgiøs fraktionisering	Kontinuert		1
<b>Geografiske faktorer</b>			
Areal	Kontinuert		1
Adgang til kyst	Kategorisk	Ja/nej	1
Olieressourcer	Kategorisk	Ja/nej	1
Diamantressourcer	Kategorisk	Ja/nej	3
Terræn	Kontinuert		1
<b>Befolkningsdata</b>			
Befolkningstal	Kontinuert		3
Adgang til uddannelse	Kontinuert		3

---

## 6. ANALYSE

I dette kapitel vil jeg gennemgå resultaterne af min analyse. I kapitlets første afsnit (6.1) vil jeg sammenligne resultaterne af Random Forest og EGBtree algoritmerne, med baseline modellerne for de 4 responsvariable introduceret ovenfor. I forlængelse af dette afsnit, vil jeg i afsnit 6.3 knytte en kommentar om validiteten af analysens resultater. Herefter vil jeg i afsnit 6.4 dykke nærmere ned i EGBtree-prædiktionsmodellerne, der alt i alt vurderes til at *performe* bedst på tværs af responsvariablene. I denne subanalyse, vil jeg først se nærmere på de *vigtigste* variable for prædiktionsmodellerne. Afslutningsvist vil jeg afsøge *machine learning* algoritmens yderligere potentiale i prædiktionen af borgerkrig, gennem en *tuning* af EGBtree algoritmen.

Alt dataindsamling, lagring, forarbejdning, visualisering og analyse af data er foretaget i open source-programmeringssproget R (Foundation, 2018). Kodningen i R er vedlagt specialet som en projektmappe, der kan eksekvere specialet fra ende til anden og sikre således muligheden for total replikation og viderebygning af projektet<sup>17</sup>. En kommenteret udgave af den kode, der udgør selve analysen kan derudover findes i appendiks E.

Til at træne baseline modellerne har jeg brugt de indbyggede funktioner i R. For *machine learning* algoritmen Random Forest har jeg benyttet tillægspakken *randomForest* (Liaw and Wiener, 2002). Dertil har jeg benyttet mig af pakken *caret* (Wing et al., 2018), der sikrer en let tilgang til træning af modeller, herunder justering af krydsvalideringen og andre parametre. *Caret*-pakken trækker blandt andet på indholdet af *xgboost*-pakken, der er en let anvendelig implementering af EGBtree algoritmen (Chen et al., 2018).

### 6.1. Modellernes performance

Som en indledende note til analysen skal det påpeges, at ingen af observationerne i prædiktionsmodellerne bruger data, der tidsmæssigt ligger samtidig med eller efter den prædikterede responsvariabel. Data struktureret på lande-år er således *lagget* med et år, mens data struktureret på lande-måned niveau er *lagget* med mindst 1 måned. Alle modellerne er trænet på et datasplit, hvor at data fra før 2012 udgør træningsdatasættet, mens data fra 2012 og frem udgør testdatasættet.

Samtidig må det igen gøres klart, at rammen for evaluering af prædiktionsmodellerne i mellem. For responsvariablene *udbrud af borgerkrig* og *borgerkrigslignende tilstande* anvendes densitetshistogrammer, confusion-matricer, ROC-kurver og AUC-værdier. For de to kontinuerte responsvariable, *antal døde* og *antal kamphandlinger* anvendes prædiktionsmodellernes *out-of sample RMSE* og  $R^2$ -værdier.

---

<sup>17</sup>Projektet vil efter bedømmelse desuden blive offentliggjort på fil-delings platformen GitHub - <https://github.com/FrederikTomPetersen>

### 6.1.1. Udbrud af borgerkrig

I dette afsnit vil jeg se nærmere på prædiktionsmodellernes evne til at prædiktere *udbruddet af borgerkrig*<sup>18</sup>. Denne responsvariable må nok erkendes som den mest interessante af de fire mål for borgerkrig. Det skyldes at netop prædiktionen af fremtidige udbrud af borgerkrig, giver spillerum for meningsfulde præventive indsatser.

I prædiktionen af *udbruddet af borgerkrig*, har jeg i modelleringen af prædiktionsmodellerne stået overfor den udfordring, at der heldigvis har været relativt få udbrud, som jeg har kunnet træne modellerne på. I mit træningsdatasæt har der været 50 observationer af udbrud af borgerkrig som modellerne kunne lære ud fra, mens der i testdatasættet kun har været 8 udbrud af borgerkrig<sup>19</sup>. Det giver en udfordring for modellerne, da det statistisk bliver sværere at finde mønstre, der kan adskille de to distinkte *outcomes* fra hinanden. Med dette in mente må det alligevel siges, at modellerne performer ganske udmærket.

I vurderingen af modellerne er første skridt at finde den optimale værdi for  $\delta$ . Her har jeg, som beskrevet i afsnit 4.2, valgt at sætte en omkostning for at forudsige *falske positive* til 1 og en omkostning for at forudsige *falske negative* til 10. Tallene er som sagt arbitrære og virkelighedsfjerne, men afspejler et ønske om at forudsige flere *sande positive*, selvom dette vil betyde flere *falske positive*. Den samlede omkostningen for forudsigelsen kan udtrykkes som ligningen:

$$\text{Omkostning} = \sum(FP \text{ ved } \delta) * 1 + \sum(FN \text{ ved } \delta) * 10 \quad (6.1)$$

Normalt ville den optimale tærskelværdi for  $\delta$  være ved den laveste omkostning. Dette har jeg forsøgt at nuancere ud fra rationale omkring, at for at forudsigelserne er brugbare, skal de forudsige faktiske udbrud af borgerkrig. Derfor har jeg valgt at acceptere omkostninger op til 400, selvom dette ikke er det sande optimale. For tærskelværdien  $\delta$  betyder det, at forudsigelserne maksimalt resultere i 320 FP og 8 FN eller en kombination herimellem. Dette resulterer i følgende optimale tærskelværdier for  $\delta$ :

**Tabel 4:** Udbrud af borgerkrig: Optimale værdier for  $\delta$  og omkostning

	Baseline	RF	EGBT
$\delta$	0,056	0,044	0,025
Omkostning	118	253	94
SPR	0,125	0,25	0,375
FPR	0,004	0,018	0,004

Af tabellen ses det, at EGBtree algoritmen klarer sig bedre, end både Random Forest algoritmen og baseline modellen, når man tager omkostningsbegrænsningen i betragt-

<sup>18</sup>Kodningen for denne del af analysen findes i appendiks E.1

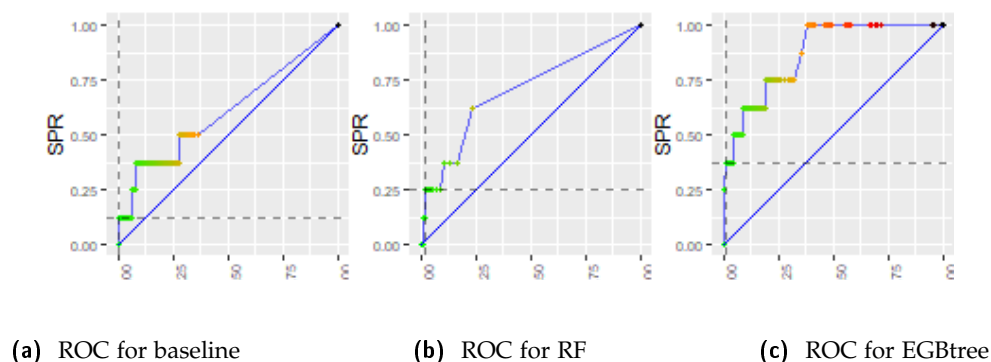
<sup>19</sup>Det samlede tal er i det fulde datasæt 91, men er grundet manglende værdier for enkelte variable reduceret til 58



ning. Dette er både hvad angår andelen af udbrud der prædikteres korrekt, men også i forhold til den samlede omkostning, der er ved at prædiktere disse. EGBtree prædiktionsmodellen prædikterer således 37,5 % af udbruddene af konflikt, mens andelen er 25 % for Random Forest algoritmen og 12,5% for baseline modellen. Man kunne undre sig over, at alle tre omkostninger er lavere end loftet på 400. Det skyldes at ingen af modellerne kunne forbedre deres *sand positiv rate* (SPR) indenfor omkostningsloftet og det dermed ikke har været meningsfuldt at maksimere omkostningen på tværs af modellerne. Tabellen kårer EGBtree modellen som den foretrukne, når man isoleret ser på den optimale  $\delta$  givet forbeholdet for den maksimale omkostning på 400.

En lidt anden fortælling finder man, når man udtrykker modellerne som ROC-kurver, der som beskrevet i afsnit 4.2, er et godt udgangspunkt for sammenligning af modellerne på tværs af alle værdier af  $\delta$ . I figur 11 ses modellerne udtrykt som ROC-kurver med de optimale  $\delta$ -værdier markeret med stiplede linje. Hvor at Random Forest algoritmen ikke synes særlig værdifuld sammenholdt med baseline modellen, under hensyntagen til den maksimale omkostning, ser modellen ud til generelt at performe bedre på tværs af  $\delta$ .

Figur 11: ROC-kurver: Udbrud af borgerkrig



Ingen af modellerne er optimale i deres visuelle fremtoning, eftersom ingen af modellerne tilnærmer sig øvre venstre hjørne. Alle modellerne klarer sig dog bedre end en tilfældig klassificering af responsvariablen, der er illustreret ved linjen med formlen  $1X + 0$ . Af figurerne kunne man få det indtryk, at EGBtree prædiktionsmodellen performer bedst på tværs af alle værdier af  $\delta$ . Dette er ikke tilfældet hvis man sammenligner AUC-værdierne for modellerne. Her har Random Forest modellen en AUC-værdi på 0,8; EGBtree modellen en AUC-værdi på 0,79 og baseline modellen en værdi på 0,57. At ROC-kurven for Random Forest modellen ikke udtrykker dette skyldes, at R-funktionen, der plotter dataet, ikke er i stand til at udregne værdierne, når grænseværdien ændres langt ude i decimalerne.

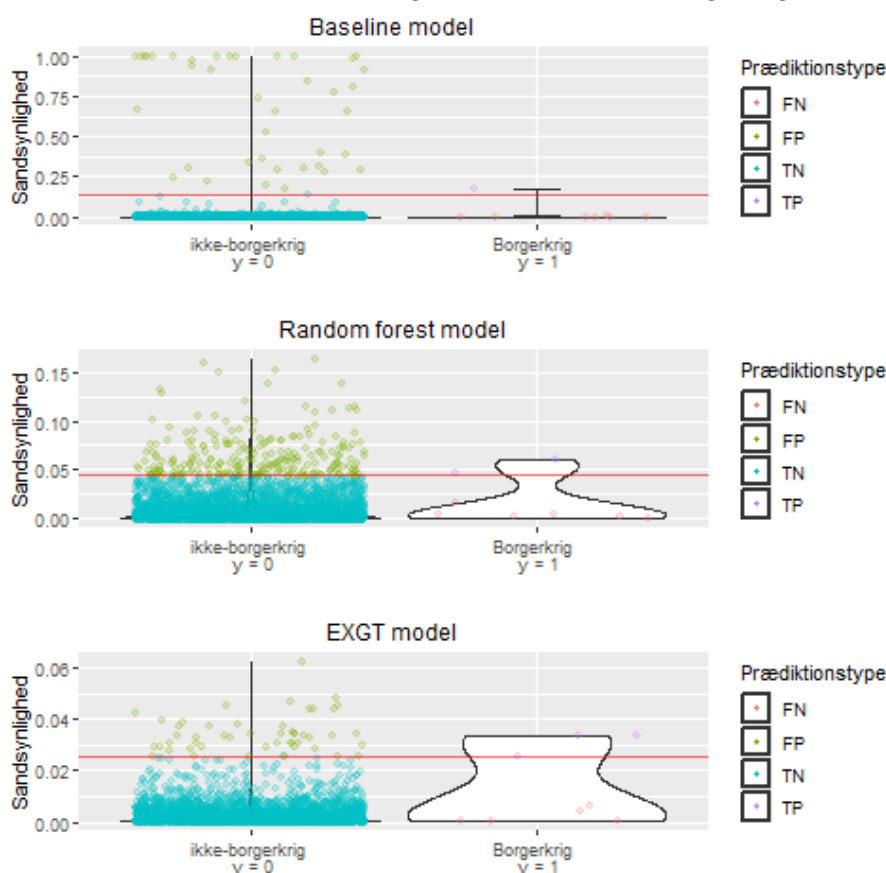
Som det ses af figur 11, er det kun muligt at opnå en sand positiv rate på 1, hvis

## 6. Analyse

man samtidig vil acceptere en falsk positiv-rate, der nærmer sig 1. Som afrunding på denne delanalyse, må det konkluderes at Random Forest modellen *performer* bedst på tværs af alle grænseværdier, mens XGBtree modellen *performer* bedst indenfor den omkostningsramme, der er opsat.

Som en sidste visuel kontrol af resultaterne har jeg i figur 12 illustreret forholdet mellem forudsagte SP, SN, FP og FN i et densitetshistogram. Her ses det tydeligt hvilke typer af fejlprædiktioner, der bidrager mest til den samlede omkostning forbundet med den optimale tærskelværdi af  $\delta$ . For baseline modellen er det særligt de 7 FN, der vejer tungt. For Random Forest algoritmen er det henholdsvis de 7 FN samt en stor mængde FP, der giver den høje omkostning. EGBtree algoritmen formår at ramme en gylden middelvej ved at forudsige færre FN, uden at mængden af FP øges markant. Figuren gør det samtidig klart, at andre tærskelværdier for  $\delta$  ville kunne have optimeret mulighederne for at forudsige faktiske tilfælde af udbrud, hvis man er villig til at acceptere en højere omkostning.

**Figur 12:** Densitets histogrammer: Udbrud af borgerkrig



Noget af det, der er interessant at se nærmere på, er de tilfælde, hvor at modellerne på tværs har vurderet en høj sandsynlig for borgerkrig, men hvor at datadefinition af

borgerkrig, ikke har defineret et udbrud af borgerkrig. I tabel 5 nedenfor ses en oversigt over de 10 lande, der i EGBtree algoritmen optrådte som falske positive med en høj sandsynlighed for *udbrud af borgerkrig*, men hvor at datadefinitionen på en borgerkrig har vurderet, at der ikke er borgerkrig.

**Tabel 5:** Falske positive i EGBT-algoritmen med høj sandsynlighed af borgerkrig

Land	År	Måned	forudsagte sandsynlighed
Congo	2012	1	0,058
Burma	2013	3	0,051
Indien	2016	1	0,042
Filippinerne	2014	1	0,042
Centralafrikanske Republik	2017	2	0,042
Irak	2013	1	0,031
Chad	2016	1	0,029
Niger	2016	1	0,028
Egypt	2017	3	0,028
Pakistan	2017	1	0,028

Hvis man sammenligner denne tabel med de faktiske udbrud af borgerkrig<sup>20</sup>, ses det, at der er overlap mellem de to lister, hvor Congo, Filippinerne og den Centralafrikanske Republik udgør fællesmængden. At vurdere at sandsynligheden for borgerkrig i disse lande har været stor, har alt i alt altså kun været tidsmæssigt ved siden af. Den resterende gruppe er lande, som man i et vist omfang kan nikke genkendende til som lande, hvor at borgerkrig har været en reel risiko. I Burma er dette udtrykt gennem fordrivelsen af rohingyaerne, I Chad og Niger i form af de politiske uroligheder mellem etniske grupper og manglende økonomisk udvikling, Egypten i form af den magttransition, der har været i kølvandet på det arabiske forår og Pakistan og Indien i forhold til Taleban og konflikten i Kashmir-regionen. Disse falske positive, hvad enten de optræder i listen over faktiske borgerkrige eller ej, må trods deres bidrag til upræcise forudsigelser, tænkes at have en praktisk anvendelse, da målretningen af præventive tiltag i disse lande, formentlig vil kunne gyde olie på vandende.

EGBtree modellen, der vurderes til at performe bedst i prædiktionen af *udbrud af borgerkrig* synes delvist anvendelig. Prædiktionsmodellen fanger 37,5 % af de faktiske udbrud af borgerkrig i test-datasættet. Samme sand positiv-rate må derfor forventes at eksistere i den virkelige verden, hvor prædiktionsmodellen appliceres på ukendt data. De 37,5 % korrekte prædiktioner er akkompagneret af 49 FP fordelt på de 10 lande listet i tabel 5. 49 FP synes at være en forsvindende lille andel af de 10252 negative og i lyset af landenes politiske tilstand, bør det ikke være disse FP, der fraholder en fra at anvende modellen i praksis.

<sup>20</sup>Cameroun(2014), Centralafrikanske Republik(2014), Congo (2012 & 2016), Libyen(2014), Filippinerne(2017), Tyrkiet (2015), Ukraine (2014)

### 6.1.2. Eksistensen af borgerkrigslignede tilstande

I dette afsnit vil jeg se nærmere på forudsigelsen af responsvariablen *eksistensen af borgerkrigslignede tilstande*. Kodningen i R, der ligger til grund for modelopbygning og visualiseringerne kan ses i appendiks E.2

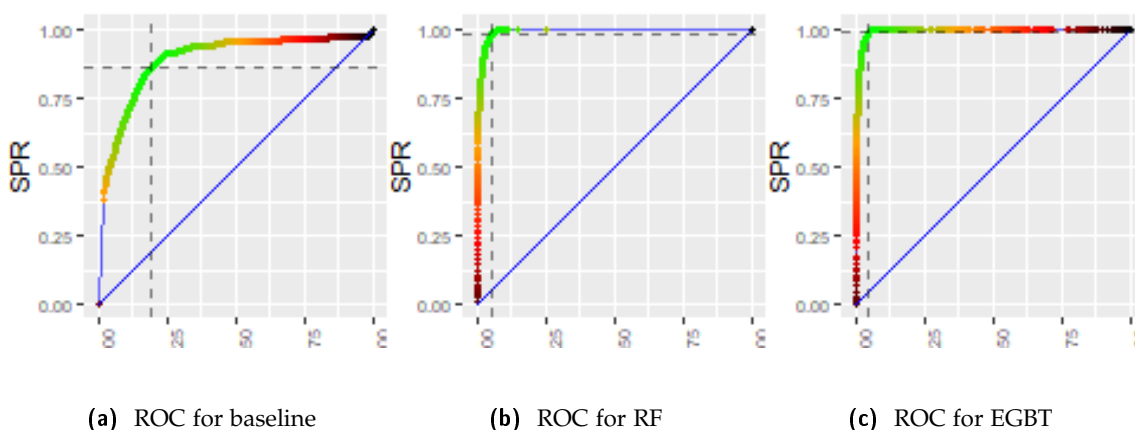
For igen at kunne se nærmere på algoritmernes performance, skal den optimale tærskelværdi af  $\delta$  sættes for hver af modellerne. Da jeg satte  $\delta$  for responsvariablen *udbrud af borgerkrig*, var det en vægtning mellem at maksimere antallet af *sande positive*, samtidig med at begrænse den samlede omkostning. I dette tilfælde har jeg udelukkende fokuseret på at minimere den samlede omkostning ved sætningen af tærskelværdien  $\delta$ . Det har jeg gjort ud fra en overvejelse omkring, at det er mere værdifuldt at prædiktere *udbruddet af borgerkrig*, end det er at kende eksistensen af borgerkrig. Samtidig er prædiktionen af *borgerkrigslignende tilstande* et meget videre begreb, hvor antallet af *positive* langt overstiger de otte udbrud af borgerkrig i testdatasættet. I tabel 6 ses oversigten over værdisætningen af  $\delta$  og de forbundne omkostninger.

**Tabel 6:** Eksistensen af borgerkrig: Optimale værdier for  $\delta$  og omkostning

	Baseline	RF	EGBT
$\delta$	0,337	0,125	0,042
Omkostning	2926	595	494
SPR	0,932	0,989	0,993
FPR	0,189	0,053	0,05

Her er det benævnelsesværdigt, at baseline modellen har en markant højere minimumsomkostning end Random Forest og XGBtree algoritmerne. Samtidig ses det, at den mindst omkostningsfulde værdisætning af  $\delta$  er væsentligt højere, og det må derfor forventes, at *machine learning* modellerne generelt performer bedre. Det er også det billede der går igen, når man ser nærmere på ROC-kurverne for modellerne. ROC-kurverne er illustreret i figur 13, hvor at  $\delta$ -optimalen er markeret med stiplet linje.

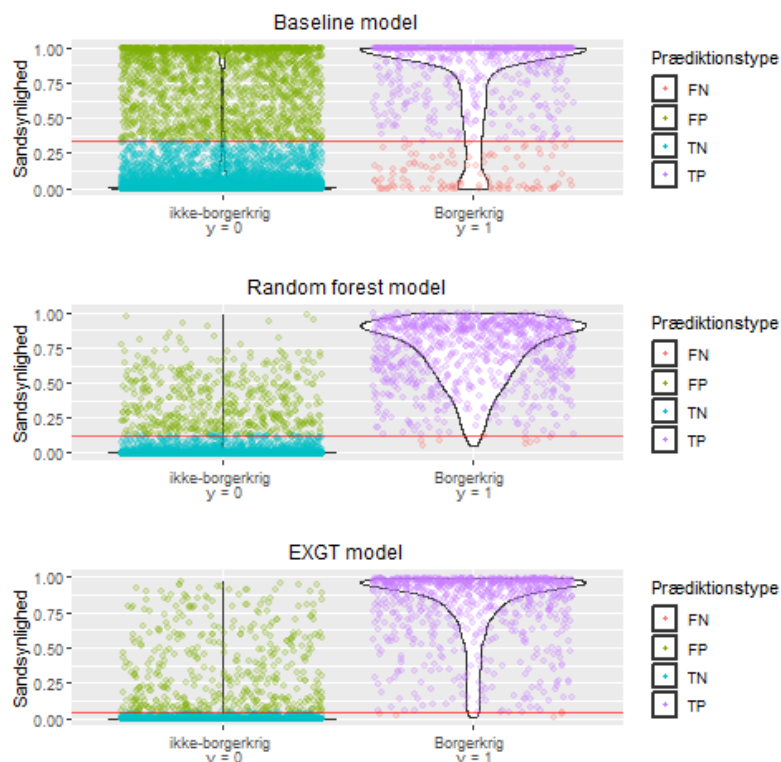
**Figur 13:** ROC-kurver: Eksistensen af borgerkrig



Som det ses af figuren, er ROC-kurverne for Random Forest og EGBtree algoritmerne næsten perfekte i deres tilnærmelse af øvre venstre hjørne. Baseline modellen er mindre perfekt, men er stadig markant bedre end en tilfældig prædiktion, udtrykt ved den perfekte korrelation mellem falsk positiv-raten og sand positiv-raten. For alle tre modeller gælder det, at sand positiv-raten er tilnærmelsesvis 1. For EGBtree algoritmen optræder der kun 5 FN ud af de i alt 830 positive. Reelt betyder dette at tærskelværdien  $\delta$  kan sættes så lavt at næsten alle sande positive prædikteres uden at det går ud over falsk positiv-raten. For Baseline modellen må der accepteres en falsk positiv-rate på 18,5 %, hvis tilnærmelsesvis alle de sande positive skal prædikteres af modellen.

Når man ser på algoritmernes performance i forhold til at prædiktere borgerkrigslignede tilstande er det iøjnefaldende, at begge *machine learning* algoritmer igen klarer sig bedre end baseline modellen. Hvis man ser på densitetshistogrammerne i figur 14 er det tydeligt at der er en forskel i fordelingen af forudsigelsestyperne. "Paddehatte fordelingen" af observationerne for gruppen med borgerkrig, vidner om en model, der i store træk er i stand til at skelne de grupper for hinanden. Omvendt ses en *anker-formet* fordeling af observationerne, hvor der ikke er borgerkrig. Igen betyder dette at modellerne er i stand til at danne særskilte mønstre i dataet. Undtagelsen er baseline modellen, der for *ikke-borgerkrigs* observationerne ikke har en entydig fordeling af observationerne i gruppen, der ikke har oplevet borgerkrig.

**Figur 14:** Densitets histogrammer: Eksistensen af borgerkrig



Resultatet af denne delanalyse er klokkeklart. *Machine learning* algoritmerne *performer* markant bedre end baseline modellen, og vidner dermed om et stærkt potentiale for *machine learning*. Samtidig kan særligt EGBTree algoritmen igen fremhæves, da modellen korrekt prædikterer 99,3 af de *borgerkrigslignende tilstande*. EGBtree prædiktionsmodellen fejl-prædikterer kun 5 falske negative og 444 falske positive fordelt på 29 lande.

Af de 5 falske negative er det værd at undersøge, om netop disse observationer er de samme falske negative som EGBtree modellen prædikterede for *udbruddet af borgerkrige*. Dette er i midlertidig ikke tilfældet<sup>21</sup>. Dette har en markant betydning for modellens performance, da modellen reelt set prædiktere alle udbrud af borgerkrig korrekt, samtidig med at den er i stand til præcist at prædiktere måneder med borgerkrigslignende tilstande med en måneds varsel.

### 6.2. Intensiteten af konflikter

I dette afsnit vil jeg se nærmere på forudsigelsen af responsvariablene *antal døde & antal kamphandlinger*. Prædiktionen af *antallet døde & kamphandlinger* har ikke deres eksistensberettigelse i prædiktionen af, hvor og hvornår der udbryder borgerkrig eller borgerkrigslignende tilstande. Værdien i prædiktionen af disse to responsvariable ligger i den information om omfanget af konflikterne, der muliggør en skalering og tilretning af internationale interventioner. Prædiktionen af intensiteten kan derfor ses som et supplement til de to øvrige, prædiktioner, der giver et indblik i konfliktens karakter. Kodningen i R, der ligger til grund for denne del af analysen, kan ses i appendiks E.3.

Som nævnt adskiller intensitets-responsvariablene sig fra de to øvrige ved at være kontinuerte. Dette betyder, at evalueringen af modellernes performance i stedet må læne sig op ad RMSE og  $R^2$  værdierne for modellerne i test-datasættet, som beskrevet i afsnit 4.2. Disse værdier er opsummeret i tabel 7 og igen i tabel 9. Foruden de numeriske værdier, vil jeg i evalueringen se nærmere på fordelingen af fejllædende.

#### 6.2.1. Antallet af døde

I tabel 7 nedenfor ses RMSE og  $R^2$  for modellerne i prædiktionen af *antallet af døde*. Som det ses af tabellen, har baseline modellen den absolut højeste RMSE. Modellen prædikterer i gennemsnit 2082 døde forkert. Resultatet vidner om, at antallet af døde ikke let lader sig beskrive med en lineær model. Derudover ses det, at XGBtree modellen *performer* bedst af de to *machine learning* modeller.

I figur 15 er de tre modellers prædiktioner opsummeret visuelt. På den øverste figur ses det prædikterede antal døde, plottet mod det faktiske antal døde<sup>22</sup>. Her udtrykker den sorte linje den perfekte prædiktion. Som det kan ses af grafen, prædikterer særligt

---

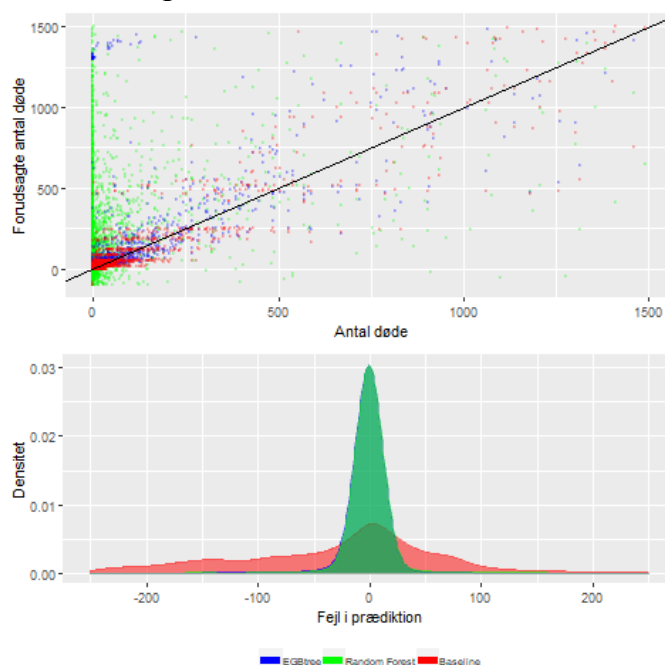
<sup>21</sup>De 5 falske negative er: Algeriet april 2014, Brasilien februar 2017, Colombia januar, 2014, Niger januar 2017 og Saudi Arabien oktober 2015

<sup>22</sup>Grafen er skåret til, således at værdier der ligger udenfor 0,0 og 1500,1500 ikke er medtaget

**Tabel 7:** Sammenligning af modellernes performance: Antal døde

	RMSE	R <sup>2</sup>
Baseline	2082	0,002
Random Forest	486	0,148
EGBtree	65	0,82

baseline modellen meget forkert. Ligeledes ses det at Random Forest modellen prædikerer meget forkert, og det er muligt at identificere to klynger ved omkring 600 og 1300 prædikterede døde. EGBtree modellen prædiktioner ser ud til at placere sig nærmere den perfekte prædiktation, hvilket også understøttes af RMSE-værdierne for de to modeller.

**Figur 15:** Prædiktation af antal døde

(a) Øverst: Prædikeret vs faktiske antal døde

(b) Nederst: Kernel densitetsdiagram for residualerne

I den nederste figur i figur 15 har jeg visualiseret modellernes residualer. Det vil sige forskellen mellem det prædikterede antal døde og det faktiske antal døde. I densitets-histogrammet er det muligt at definere en *smoothing* parameter, der udtrykker hvor mange standardafvigelser modellen skal gå ud for den enkelte observation, når figuren skal udregne densiteten (Wickham et al., 2018). Parameteret styrer således hvor opmærksom visualiseringen skal være på de nærmeste observationer i forhold til observationer, der værdimæssigt ligger længere væk.

Her har jeg haft sat værdien 12, for at skabe en glat visualisering frem for en mere

partikulær visning. Som det ses af 16b ligger residualerne for Random Forest og EGBtree algoritmen tilnærmelsesvist oven på hinanden indenfor det fremvise udsnit af residualerne<sup>23</sup>. Værdierne for de to algoritmer er centreret omkring 0, hvor langt størstedelen af observationerne ligger indenfor  $\pm 50$ . Den markant højere RMSE-værdi for Random Forest algoritmen er således et udtryk for en række *outliers*, der meget vel kunne være de to klynger, der kan identificeres af figur 16a. Af densitetshistogrammet ses den tydeligt at baseline modellen performer markant dårligere. Residualerne er ikke på samme måde centreret omkring 0 og grafen fremstår fladere.

Alt i alt performer EGBtree algoritmen bedst, med en centrering af prædiktionernes residualer omkring 0, med færre *outliers* end både baselinemodellen og Random Forest algoritmen. XGBtree modellens prædiktioner er overraskende præcis. På trods af den gennemsnitlige fejl-prædiktion på 65 døde vurderes det til, at prædiktionsmodellen kan tjene som et kvalificeret bud på intensiteten i en beslutningsproces.

### 6.2.2. Antallet af kamphandlinger

Som supplement til antallet af døde, som proxy for intensiteten af interne konflikter, vil jeg i dette afsnit se på prædiktionen af antallet af kamphandlinger. Antallet af kamphandlinger er ikke et entydigt mål intensiteten, da flere forhold gør sig gældende, hvis man f.eks. skal vurdere om 5 kamphandlinger, er mere intense end 1 kamphandling. Ikke desto mindre er det en oplysning, der er relevant i interventioner, da det kan bidrage med information om, hvor mange steder der skal iværksættes indsatser. I tabel 8 ses en opsummering af modellernes performance.

**Tabel 8:** Sammenligning af modellernes performance: Kamphandlinger

	RMSE	R <sup>2</sup>
Baseline	3,84	0,954
Random Forest	4,93	0,930
EGBtree	4,27	0,944

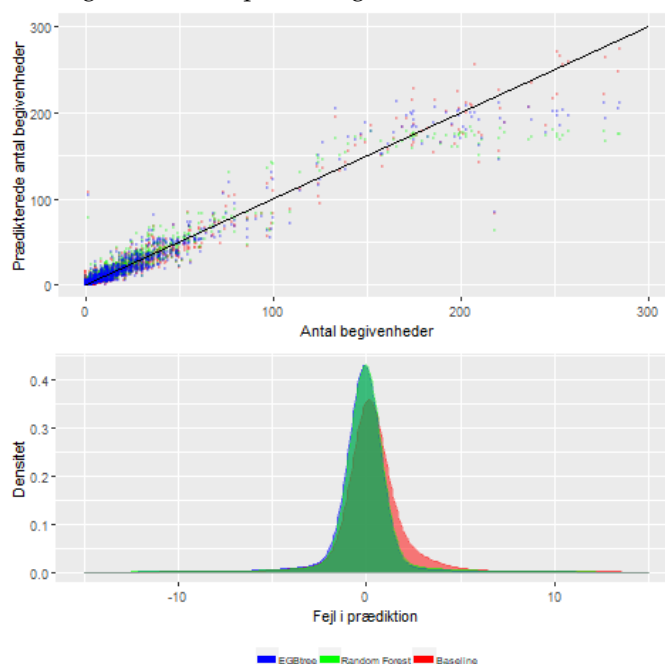
Her er det overraskende at baseline modellen *performer bedst* med en RMSE på 3,84 og en R<sup>2</sup> på 0,954. Af de to *machine learning algoritmer* performer XGBtree modellen bedst. Hvis man ser nærmere på visualiseringerne af modellernes performance i figur 16, er det tydeligt at modellernes prædiktioner er væsentlig mere konsistente end for *antallet af døde*. I visualiseringen af de prædikterede værdier mod de faktiske værdier ses en pæn tilnærmelse af de faktiske værdier, med få prædiktioner, der skyder markant forbi. Særligt ved de ekstreme værdier prædikterer baseline modellen tættere på de faktiske værdier, end både Random Forest og EGBtree algoritmen. På tværs af spektret af antal kamphandlinger performer baseline modellen bedre, hvilket understøttes af RMSE og R<sup>2</sup> værdierne.

<sup>23</sup>graffen er skåret således at residualer større end  $\pm 250$  ikke vises



Hvis man ser nærmere på densitetsplottet i figur 16, ser det mærkværdigt nok ud til, at Random Forest og EGBtree algoritmernes residuaer er *mere* centreret omkring 0, hvis man isoleret ser på de prædiktioner, der ligger indenfor  $\pm 15$  af de faktiske værdier. Dette vidner om, at det netop er de ekstreme værdier for *antallet af kamphandlinger* som modellerne har svært ved at prædiktere. En forklaring på dette kan være, at der er relativt få ekstreme værdier i datasættet og algoritmerne derfor har svært ved at udnytte sin fleksibilitet til at tage højde for disse tilfælde.

**Figur 16:** Forudsagte antal kamphandlinger mod faktiske antal kamphandlinger



(a) Øverst: Prædikteret vs faktiske antal kamphandlinger

(b) Nederst: Kernel densitetsdiagram for residualerne

Gennemgående må det erkendes at modellerne ikke *performer* bemærkelsesværdigt godt, når det kommer til prædiktionen af antallet af kamphandlinger, som mål for intensiteten. Baseline modellen, der viste sig at være den bedste model, prædikterede i gennemsnit 3,8 kamphandlinger forkert. Prædiktionen kan dog bruges som en løs indikator på, hvor mange kamphandlinger, der kan forventes i en intervention og understøtter dermed specialets problemstilling.

### 6.2.3. Sammenhængen mellem resultaterne

I tabel 9 ses en opsummering af resultaterne for de tre prædiktionsmodeller performance på de fire responsvariable: *udbruddet af borgerkrig*, *eksistensen af borgerkrig*, *antallet af døde* & *antallet af kamphandlinger*.

Det bredeste sammenligningsgrundlag for de to binære responsvariable er AUC-værdien, der som illustreret i afsnit 4.2, er et godt mål for modellens samlede performance på tværs af tærskelværdien  $\delta$ . Her udkonkurrerer de to *machine learning* algoritmer baselinemodellen, og har begge AUC-værdier mellem 0,79 og 0,99 for de to responsvariable. Når man sammenholder disse ganske fornuftige AUC-værdier med SPR, bliver *machine learning* algoritmernes potentiale i prædiktionen af borgerkrig konsolideret. Særligt EGBtree algoritmen *performer* her godt. Dette understreges af, at 37,5 af de ukendte *udbrud af borgerkrig* prædikteres korrekt en måned før de bryder ud. Samtidig prædikterer algoritmen 99,3 % af de borgerkrigslignede tilstande med én måneds varsel. Et resultat, der er marginalt bedre end Random Forest algoritmens prædiktioner.

**Tabel 9:** Sammenligning af modellernes performance

	<i>Area under curve</i>		<i>RMSE</i>	
	Udbrudet	Tilstand	Døde	Kamphandlinger
Baseline model	0,57	0,894	2084	3,84
Random Forest	0,80	0,993	486	4,93
EGBtree	0,79	0,994	65	4,27

	<i>Andel af SP</i>		<i>R<sup>2</sup></i>	
Baseline model	12,5 %	93,2%	0,002	0,954
Random Forest	25 %	98,9%	0,14	0,930
EGBtree	37,5 %	99,3%	0,62	0,944

	<i>Andel af SN</i>			
Baseline model	99,6 %	81,1%		
Random Forest	98,2%	99,4%		
EGBtree	99,6%	99,5%		

Samtidig er det interessant at se, at AUC-værdien for alle tre modeller stiger mellem *udbrud af borgerkrig* og *borgerkrigslignende tilstande*. Dette tyder på, at modellerne generelt *performer* bedre, når modellen får flere “positive” observationer at træne på, hvilket stemmer overens med den indledende forventning til prædiktionernes *performance*.

For de to mål for konflikt intensitet, er den gennemsnitlige fejl-forudsigelse udtrykt ved RMSE-værdien og  $R^2$  for modellens forklaringskraft det bedste sammenligningsgrundlag. For *antallet af døde* står det klart at baseline- og Random Forest modellen ikke har nogen værdi i kraft af deres høje RMSE-værdier. EGBtree modellen er derimod anvendelig som en proxy for konfliktens forstående intensitet. EGBtree modellen prædiktere således  $\pm 65$  døde forkert, der i forhold til intense konflikter må anses som en acceptabel fejlmargen.

For *antallet af kamphandlinger performede* baselinemodellen overraskende nok marginalt bedre end de to *machine learning* algoritmer. Baselinemodellen prædiktede her gennemsnitligt 3,84 kamphandlinger forkert. Dette resultat kan alt i alt også anvendes

som en proxy for intensiteten af forstående konflikter, der kan være med til at guide udformning af en militær intervention og nødhjælpsarbejde.

For at vende tilbage til anskuelsen af borgerkrig som et prædiktionsproblem, kan man på baggrund af ovenstående resultat konkludere, at prædiktion af borgerkrig har et potentiale i tilretningen af præventive og reaktionære tiltag mod borgerkrig. Modellerne kan i et vist omfang både prædiktere udbrud og eksistensen af borgerkrig, samtidig med at prædiktionerne kan bidrage med information om intensiteten af konflikterne. Information, der er essentiel for det internationale samfund, hvis de skal gøre sig håb om at forhindre eller nedskalere konflikter.

### 6.3. Modellernes validitet og resultaternes pålidelighed

Et centralt led i enhver forskning er at afrapportere de usikkerheder, der er forbundet med undersøgelsens resultater. I forrige afsnit afrapporterede jeg således usikkerhederne for prædiktionerne af intensiteten af konflikterne gennem *RMSE*-værdierne. *AUC*-værdierne for de binære variable er på samme måde forbundet med en statistisk usikkerhed, der særligt stammer fra opsplitningen af dataet i et trænings- og test-datasæt, samt den *resampling*, der foregår i algoritmerne.

Når man laver sit datasplit for at skabe henholdsvis trænings- og testdatasættet, skabes der en usikkerhed for modellen, da alternative opdelinger vil føre til andre modeller. Samtidig har re-sampling teknikken og de opdelinger af observationer som denne foretager, betydning for modellens usikkerhed. Samplingen af data i prædiktionsmodellerne har altså betydning for *AUC*-værdierne, men også  $\beta$ -koefficienterne for de kontinuerte variable. Her kan *RMSE*-værdierne dog ses som en proxy for usikkerheden.

I min afrapportering af resultaterne har jeg fravalgt at beskrive usikkerheden for *AUC*-værdierne. Det skyldes først og fremmest, at ikke kan antages at være normalfordelte omkring et gennemsnit, hvilket er den grundlæggende forudsætning for, at man kan udregne og standardfejl (James et al., 2013, s. 65). Samtidig er *AUC*-værdien et overordnet performance mål, der på samme måde som  $R^2$  beskriver en models samlede performance, der ikke er meningsfuldt at udregne standardfejl for. Disse årsager er den almindeligste grund til, at der i litteraturen ikke afrapporteres standardfejl for *AUC*-værdier (James et al., 2013, s. 187)

Denne overvejelse omkring den manglende afrapportering af standardfejlene for modellerne resulterer i, at den observerede forskel mellem modellernes performance i tabel 9 ikke statistisk signifikant. Denne problemstilling står ikke centralt i min problemstilling og jeg vælger derfor at fokusere på EGBtree algoritmen, der er den *machine learning* algoritme, der gennemgående har leveret de bedste resultater.

### 6.4. Den bedste model

På tværs af de fire responsvariable har EGBtree algoritmen gennemgående *performet* bedre end både baseline modellen og Random Forest algoritmen. Som jeg har belyst er algoritmen i vid udstrækning i stand til at prædiktere centrale mål for borgerkrig. Resultaterne rummer dog et potentiale for forbedringer, der indenfor den ramme mit dataet giver, må afsøges en yderligere modelspecifikation. I det følgende afsnit vil jeg derfor dykke ned i det rum for forbedringer som algoritmen rummer gennem en finere *tuning* af algoritmens parametre.

Som en opfølgning på de resultater som *tuning* processen giver, vil jeg afslutte analysen ved at se nærmere på variable, der har været særligt givende for præcisionen af prædiktionerne. Foruden at være et selvstændigt bidrag til analysen i form af den information om disse vigtige variable giver for fremtidige modelspecifikationer, tjener afsnittet også et andet formål. I min analyse har jeg som et selvstændigt bidrag introduceret data over politiske begivenheder udledt fra GDELT-databasen (Leetaru and Schrodt, 2013). Ved at se nærmere på, om netop disse variable er vigtige for prædiktionsmodellernes performance, kan det sandsynliggøres hvorvidt at disse informationer i fremtidige studier skal inddrages i prædiktioner og hvorvidt samfundsvidenskaben i højere grad skulle indskrive disse kontekstuelle variable i den teoretiske litteratur.

At gennemgå den yderligere tuning proces for EGBtree modellen for alle 4 responsvariable, vil være en omfattende gennemgang, der ligger uden for afsnittets formål. I gennemgangen vil jeg derfor fokusere på prædiktionen af udbrud af borgerkrig.

#### 6.4.1. De vigtigste variable i prædiktionsmodellerne

I studier, hvor man foretager kausal-estimationer, er måske det mest centrale analytiske greb at se på korrelationen mellem de uafhængige variable og den afhængige variabel og analysere de usikkerheder, der er forbundet med estimationerne. I prædiktionsstudier, der bygger på *machine learning* algoritmer, er det ligeledes interessant at se på variabelenes betydning for prædiktionerne. Som tidligere belyst må algoritmerne anses som en teoretisk *black box*, hvor det ikke er meningsfuldt at udlede samvariationen med responsvariablene. Som alternativ til analysen af påvirkningen på responsvariablene, kan man i stedet se på, hvor vigtige variablene er for prædiktionernes præcision.

I litteraturen benævnes denne vigtighed for *variables importance* (Hastie et al., 2017, s. 367 - 368). At identificere de uafhængige variable, der har stor betydning for prædiktionen af responsvariablene, er ikke kun teoretisk interessant, det er også vigtigt for at optimere de udregninger, der ligger til grund for prædiktionerne. F.eks. har jeg i mine modeller haft inddraget 124 uafhængige variable. Der skal ikke meget tankevirksomhed til at komme frem til, at nok ikke alle variablene er lige relevante for

prædiktionsmodellernes *performance*. Gennem en analyse af variablenes *importance* for prædiktionernes præcision, er det potentielt muligt at reducere antallet af variable og dermed reducerer tidsforbruget for udregningerne og den nødvendige computerkraft. Samtidig kan analysen bidrage til den teoretiske forståelse af feltet.

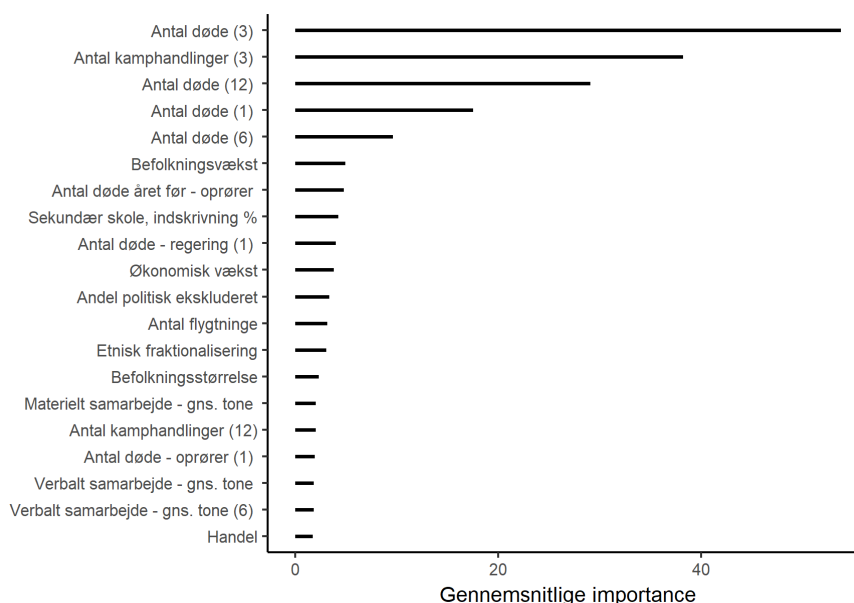
En måde hvorpå man kan tilnærme sig information om i hvilket omfang en variabel bidrager til prædiktionen, er at se på hvor ofte variablene bliver brugt til at foretage den segmentering af data, der foregår når algoritmen udregner klassifikationstræerne. For et enkelt klassifikationstræ vil et sådan relativt mål for variabelens *importance* kunne udtrykkes således:

$$I_x^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = x), \quad (6.2)$$

hvor at  $t$  er et udtryk for hvert blad i klassifikationstræet, der er dannet på baggrund af segmentering af variable  $X_{v(t)}$ , der giver den største estimerede forbedring af prædiktionen. Ved hver af segmentering tillægges en responsvariablen en konstant alt efter hvilket blad observationen ender i. Dette betyder at den relative *importance* af en variabel, kan udregnes som summen af den kvadrede værdi, der tillægges i hvert blad for en given variabel (Hastie et al., 2017, s. 368). For ensemble metoder som Random Forest og EGBtree algoritmen kan denne *importance* værdi udregnes ved at tage gennemsnittet af værdierne for alle de voksede klassifikationstræer (Hastie et al., 2017, s. 368). Ligningen vil således være:

$$I_x^2 = \frac{1}{M} \sum_{m=1}^M I_x^2(T_m), \quad (6.3)$$

hvor at  $M$  er summen af klassifikationstræerne  $m$ . I *machine learning* litteraturen, har det vist sig at denne generaliserede *importance* er et bedre mål for variablenes *importance* da det reducerer betydningen af dårlige segmenteringer i det enkelte klassifikationstræ (Hastie et al., 2017, s. 368). Når man kigger på disse generaliserede *importance* værdier, er det vigtigt at holde sig for øje, at værdierne ikke indeholder information om hvordan at variabelen påvirker prædiktionen, og værdierne kan dermed ikke bruges til at understøtte kausale forklaringer om sammenhængen mellem variablene. Værdierne kan dog godt bruges til at sandsynliggøre kausale forklaringer, da en variabel der har stor betydning for en prædiktion, formentlig også har et kausalt om ikke andet så indirekte forhold til responsvariablen.

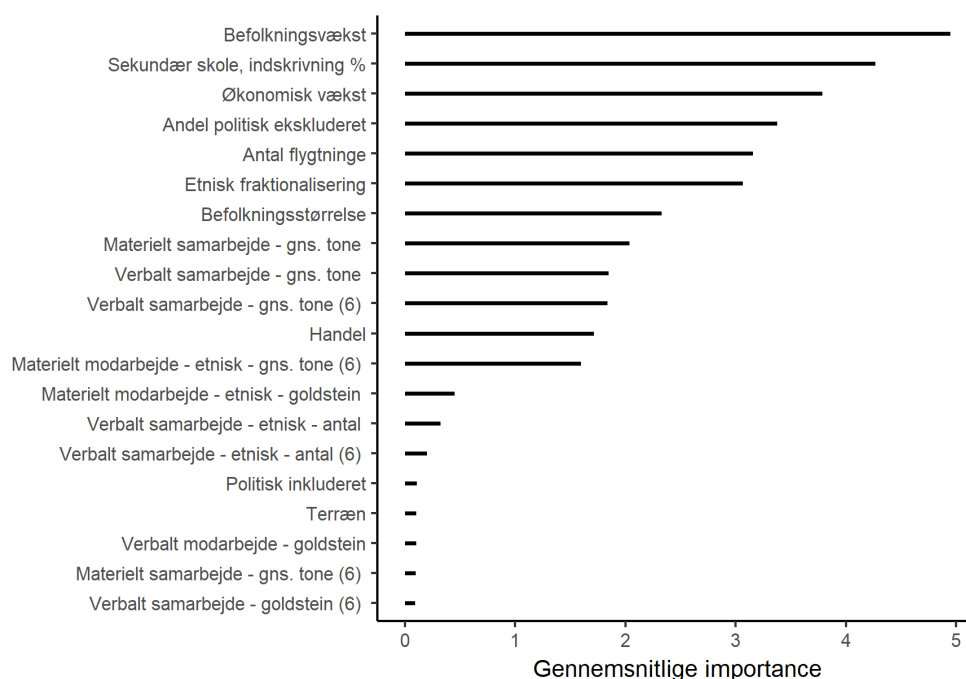
**Figur 17:** Gennemsnitlige *variable importance*

\*Tallene i parentes indikerer det antal af forgående måneder gennemsnittet er udregnet på baggrund af

Da jeg i min analyse har prædikeret 4 forskellige forhold omkring borgerkrig, har jeg i afsnittes figurer og tabeller taget udgangspunkt i den gennemsnitlige *importance* af variablene på tværs af modellerne. Dette har den fordel, at det er muligt at udlede hvilke variable, der samlet set bidrager til prædiktionernes præcision. Ulempen er, at det ikke er muligt at se, hvad variabelens *importance* for den enkelte prædiktionsmodel er. I figur 17 ses en oversigt over de 20 variable der gennemsnitligt har den største *importance* på tværs af de fire prædiktionsmodeller.

Af figur 17 fremgår det, at de 5 variable, der har den allerstørste betydning for prædiktionernes performance og præcision, er et lands konflikthistorie. Dette fund suppleres med, at 10 ud af de 20 vigtigste variable udgøres variable, der er direkte relateret til et lands konflikthistorie. Særligt det gennemsnitlige antal døde i månederne op til prædiktionen har en stor betydningen, når klassifikationstræerne foretager deres segmentering af træningsdatasættet. Dette er ikke det mest overraskende resultat, da man i borgerkrig som i andre konflikter, må forvente en vis stiafhængighed i borgerkrigens forløb og udvikling.

Selvom dette resultat ikke er overraskende, er det meget interessant, da et lands konflikthistorie i den teoretiske del af litteraturen fylder utroligt lidt. Denne problemstilling vil jeg vende tilbage til i diskussionen og perspektivering af mine analysens resultater i afsnit 7.2. I figur 18 har jeg fraseret de variable, kan relateres til et lands konflikthistorie, da figur 17 allerede har cementeret, at et lands konflikthistorie udgør de vigtigste variable i prædiktionen af borgerkrig.

**Figur 18:** Gennemsnitlige *variable importance* - uden konfliktvariable

Når man ser bort fra de konflikt orienterede variabel, kommer en række andre indsigter til syne. Af figur 18 fremgår det at de variable, der har den største betydning for prædiktionsmodellerne, spredt sig ud over de forskellige teoretiske perspektiver på borgerkrig. Eksempelvis optræder *befolkningsstørrelsen* og *indskrivning på sekundært skoleniveau*, der kan ses som et udtryk for, at omkostningerne forbundet med rekruttering er vigtige i prædiktionen af borgerkrig. Samtidig ses der klare elementer, der understøtter politisk *grievance* som motiv f.eks. *andel af befolkningen, der er politisk ekskluderet* og den *etniske fraktionisering* i et land. Ud fra vækst- og velfærdsperspektivet bekræftes det igen af økonomisk vækst ikke blot er en central variabel i kausalestimation, men også når det kommer til prædiktion af borgerkrig.

I udarbejdelsen af prædiktionsmodellerne har det måske største selvstændige bidrag været det datasæt over politiske begivenheder, jeg har introduceret til litteraturen med udgangspunkt i GDELT-databasen (Leetaru and Schrodt, 2013). Af de 20 variable med størst *importance* stammer 10 af dem fra dette datasæt. Her viser det sig, at aggregeringen af nyhedsartikler, der udtrykker politisk verbalt- og materielt samarbejde, både i form af den gennemsnitlige tone i artiklerne og den forventede samfundseffekt af begivenheder udtrykt ved *goldstein*-værdien, bidrager væsentligt til prædiktionerne.

Af figuren ses det ligeledes, at aggregeringen af nyhedsartikler, der udtrykker materielt modarbejde og hvor at en af aktørerne har en direkte etnisk affiliation, har en stor betydning for prædiktionerne. Hvis man skulle tilskrive disse variable til en af de teoretiske motiver for borgerkrig, er det oplagt at variablene bedst understøtter motiverne

## 6. Analyse

i *grievance* litteraturen. Argumentet for dette er, at verbalt såvel som materielt politisk modarbejde netop udtrykker en *grievance* i befolkningen. Samtidig udtrykker verbalt og materielt politisk samarbejde det modsatte. Netop den teoretiske indskrivning af data over faktisk politiske begivenheder vil jeg vende tilbage til, i den heraf afledte diskussion i afsnit 7.2.

**Tabel 10:** Variablenes importance vurderet efter deres rang

Variable	Udbrud	Tilstand	Døde	Kampe	Gns. rang
Økonomisk vækst	3	1	17	11	8.00
Verbalt modarbejde - goldstein	18	11	11	3	10.75
Antal flygtninge	5	39	12	1	14.25
Andel politisk ekskluderet	4	3	19	36	15.50
Verbalt modarbejde - religiøs - goldstein	21	13	27	5	16.50
Sekundær skole, indskrivning %	2	2	18	49	17.75
Verbalt samarbejde - goldstein (6)	20	7	22	24	18.25
Materielt samarbejde - gns. tone (6)	19	6	9	45	19.75
Verbalt modarbejde - gns. tone	23	10	25	22	20.00
Verbalt samarbejde - etnisk - antal	15	45	2	26	22.00
Politisk inkluderet	16	4	20	50	22.50
Terræn	17	5	21	51	23.50
Verbalt modarbejde - gns. tone (6)	38	34	6	20	24.50
Materielt samarbejde - antal (6)	28	18	8	44	24.50
Materielt modarbejde - goldstein	40	31	10	18	24.75
Materielt modarbejde - etnisk - goldstein	14	44	1	41	25.00
Materielt samarbejde - gns. tone	8	35	13	46	25.50
Handel	11	12	26	54	25.75
Materielt samarbejde - goldstein	27	32	44	2	26.25
Bruttonationalprodukt	22	8	23	52	26.25

En anden måde at opgøre variablenes indflydelse på, er ved at se bort fra den relative *importance*, men i stedet se på hvilken rang af vigtighed variablene har. Rangen 1 tildeles til den variable, der har den største *importance*, 2 til den variable, der har den næsthøjeste *importance* og så fremdeles. I tabel 10 ses de variable med den gennemsnitligt højeste rang fraserteret landenes konflikthistorie

De variable, der optræder på listen, er overvejende gengangere fra figur 18. Tabellen viser nogle interessante mønstre omkring, at variable der står centralt i prædiktionen af visse responsvariable, er mere ubetydeligt for andre. Eksempelvis score *økonomisk vækst* rangen 3 for *udbrud* og 1 for *borgerkrigslignende tilstand*, men "kun" rang 17 i prædiktionen af antallet af døde og 11 for antallet af kamphandlinger. Et andet informativt resultat af rangordningen er, at politiske begivenheder der er et udtryk for materielt modarbejde, hvor at en af aktørerne tilskrives en etnicitet er den vigtigste variable, når det kommer til prædiktionen af antallet af døde, mens variabelen har en markant lavere rang for de 3 andre responsvariable.



Hvis formålet med identifikationen af variabelnes relative *importance* er at reducere den anvendte datamængde er den mest centrale indsigt af denne delanalyse, at hver forhold omkring borgerkrig er et udtryk for forskellige dynamikker, der må tages højde for i modelopbygningen. Indtil dette har været under gennemgående behandling, må man se sig til takke med at anvende al den potentielt relevante data.

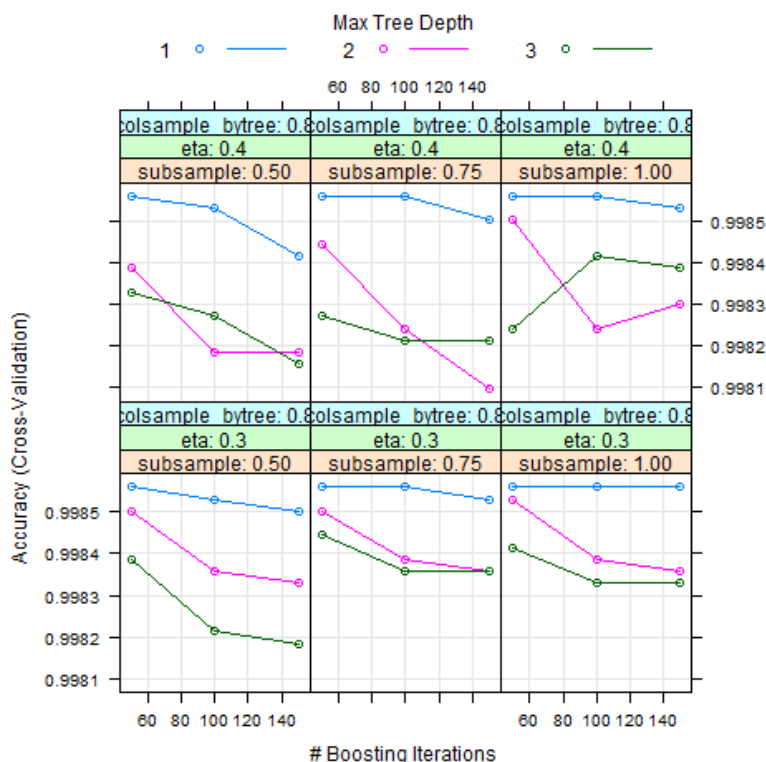
Ovenstående analyse viser, at de teoretiske funderet variable, der udspringer af kausalestimations studier, bidrager substantielt i prædiktionsmodellernes performance. *Importance*-analysen understreger samtidig, at tidsaktuel data omhandlende *det politiske landskab* er med til at skabe mere præcise, og dermed anvendelige, prædiktions. Dette resultat understøtter dermed specialet problemstilling og forhåbningerne for det introducerede datasæt.

#### 6.4.2. Tuning af modellen

*Tuning* af en algoritme har stor betydning for dennes performance og i nærtliggende eksempel evnen til korrekt at prædiktere *udbruddet af borgerkrig*. Når en model skal *tunes*, findes der ikke en på forhånd givet korrekt eller optimal *tuning*, da processen er afhængig af det underliggende data, som modelleringen sker på baggrund af (James et al., 2013, s. 321-322). Som beskrevet i afsnit 4.4 rummer EGBtree algoritmen særligt 3 parametre, der egner sig til *tuning*. Det drejer sig om *shrinkage* parameteret  $\eta$ , der udtrykker algoritmens læringsrate, antallet af sekventielle klassifikationstræer samt dybden af klassifikationstræerne. Disse parametre kaldes også de frie parametre. Ud over *tuning* af de frie parametre, er det muligt at *tune* en models performance, ved at ændre på den krydsvalidering, der er indbygget i modelleringen.

I min analyse byggede jeg værdisætningen af de frie parametre på en tilfældig afsøgning af parametersammensætningen. En proces der understøttes af *caret*-pakken (Wing et al., 2018) til R. Her har funktionaliteten forsøgt at finde den "gyldne middelvej" for sammensætningen af parametrene. En delmængde af denne "tilfældige" afsøgning er illustreret i figur 19

Figur 19: Indledende tuning af XGBtree modellen for udbruddet af borgerkrig

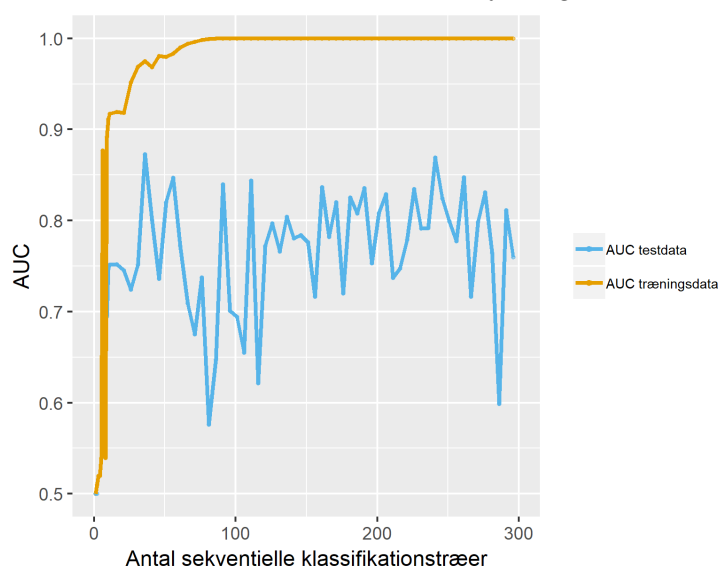


Som det fremgår af figuren, har den indledende *tuning*-proces afsøgt forskellige kombinationer af *shrinkage* parameteret  $\eta$  og de maksimale trædybder til forskellige samples af observationerne og antallet sekventielle klassifikationstræer. Mønsteret i figurerne er ikke fast, men det alligevel udledes, at algoritmerne med en maksimal trædybde på 1 performer bedre end de alternative trædybder. Samtidig er prædiktionernes performance mere stabil, når alle observationer inddrages i hvert enkelt klassifikationstræ. Evalueringen af denne "usystematiske" værdisætning af de frie parametre er foregået på baggrund af modellens nøjagtighed (*accuracy*), der udtrykker den procentuelle korrekte klassifikation i træningsdatasættet.

I min evaluering af prædiktionernes resultater, har jeg ikke lagt vægt på nøjagtighedsmålet. Det har jeg ikke gjort, da det eksempelvis for *udbruddet af borgerkrig* som responsvariable, vil være muligt at få en nøjagtighed på 99,99% ved at prædiktere, at der aldrig vil udbryde borgerkrig. Som det ses af figur 19 opnår forskellige værdier af de frie parametre sin maksimale nøjagtighed ved forskellige antal *boosting iterationer*. Altså hvor mange sekventielle klassifikationstræer, der skal plantes. En ting er at antallet af iterationer er styrende for nøjagtigheden i træningsdatasættet, der er afgørende for valget af den endelige model. Det samme gør sig gældende for testdatasættet. Her kender man dog ikke *nøjagtigheden* af modellen på forhånd, da denne først kommer til udtryk, når modellen anvendes.

For at håndtere denne problematik, har jeg i figur 20 nedenfor udregnet AUC-værdierne for trænings- og testdatasættet for forskellige antal sekventielle klassifikationstræer. I figuren har jeg anvendt de samme værdier for de frie parametre, som den model der lå til grund for resultaterne i afsnit 6.1.1. Værdierne af de frie parametre er opsummeret i tabel 11. Som det ses af figur 20 stiger AUC-værdien for træningsdatasættet jævnt op til omkring 70 sekventielle klassifikationstræer, hvorefter at yderligere træer ikke forbedrer modellens performance i træningsdatasættet. For AUC-værdien i testdatasættet ser resultatet noget anderledes ud. Den visuelle fremstilling er "plaget" af støj, der er resultatet af at AUC-værdierne for testdatasættet varierer med antallet af klassifikationstræer. En del af denne støj kan forklares med den resampling, der sker mellem hver modellering. Samtidig ses det at AUC-værdien for testdatasættet ikke når sit maksimum ved 50 sekventielle klassifikationstræer, der var det antal som blev benyttet i evalueringen af algoritmen. Denne indsigt giver anledning til en yderligere tuning af modellen for at maksimere dens performance.

**Figur 20:** Antallet af iterationers betydning for AUC



Antallet af sekventielle klassifikationstræer har stor betydning for AUC-værdien for modelleringen i testdatasættet, mens antallet mister sin betydning i træningsdatasættet efter et vist punkt. Den optimale model, der performer lige godt i træningsdatasættet som i testdatasættet, ville være udtrykt som et sammenfald mellem AUC-værdierne. Afstanden mellem AUC-værdien for modellens performance i henholdsvis træningsdatasættet og testdatasættet er med andre ord et udtryk for den usikkerhed, der er ved at anvende modellen på ukendt data. Denne indsigt har jeg brugt til, at afsøge et højere antal sekventielle klassifikationstræer, end den tilfældige afsætning gav anledning til den videre *tuning*.

I den videre *tuning* af min model, har jeg afsøgt andre sammensætningen er på de frie parametre gennem et *grid-search*. *Grid-searchet* består basalt set af, at man definere lister med forskellige værdier af de frie parametre. For hver kombination af de frie parametre udregnes der en prædiktionsmodel, der så evalueres op mod hinanden. Denne proces er udregningsmæssig meget *tung* for en computer, hvis alle tænkelige kombinationer af de frie parametre defineres. For at lette på udregningens størrelse og tyngde, har jeg i *grid-searchet* taget udgangspunkt i den indledende *tuning* i forhold til maksimale trædybde, gamma-værdi og den mindste bladvægtning. I *grid-searchet* har jeg således "kun" *tunet* på antallet af klassifikationstræer, læringsraten  $\eta$ , samt andelen af medtagne variable og observationer i hvert klassifikationstræ.

På baggrund af *tuningen* af disse 4 frie parameter opnåede jeg samlet set en bedre prædiktionsmodel. Den resulterende *nye* bedste model ses i tabel 11, hvor den sammenholdes med den *oprindelige* bedste model. Den nye *tunede* model adskiller sig markant fra den oprindelige model, hvilket indikerer at den tilfældige indledende *tuning*-proces ikke har fundet frem til bedste værdisætning af de frie parametre i *out-of-sample* evalueringen.

**Tabel 11:** EGBtree-modellen med systematisk og tilfældig tuning

Parametre	EGBtree	
	Systematisk tuning	Tilfældig tuning
klassifikationstræer	441	50
$\eta$	0,25	0,3
Max Trædybde	1	1
Mindste blad vægt	1	1
gamma	0	0
Variable sample	0,8	0,6
Observations sample	0,9	0,5
AUC-værdi	0,88	0,79

I den nye *tunede* prædiktionsmodel er  $\eta$  0,25 mod 0,3 i den oprindelige model. Denne lavere læringsrate er akkompagneret med et markant højere antal sekventielle klassifikationstræer. Dette er ikke overraskende, da forholdet mellem de to frie parametre er modsatrettede, jævnfør afsnit 4.4. Samtidig er andelen af variable og antallet af observationer medtaget i hvert klassifikationstræ steget til henholdsvis 80 % og 90 % fra 60 % og 50%. Denne værdisætning er lidt overraskende, da højere værdisætninger for disse frie parametre burde give anledning til *overfitting* af træningsdatasættet. Ikke desto mindre har netop denne kombination af de frie parametre forbedret prædiktionsmodellens performance af alle grænseværdier af  $\delta$ , og øget AUC-værdien fra 0,79 til 0,88.

*Tuningen* af modellen har ikke kun betydning for modellens samlede performance udtrykt gennem AUC-værdien. *Tuningen* har også betydning for resultatet af prædiktionen. Resultatet er vist i tabel 12. I tabellen er det samme *omkostnings-framework* anvendt, som oprindeligt blev anvendt på modellen i afsnit 6.1.1. Som det ses af tabellen er

EGBtree algoritmens prædiktive performance forbedret, når man isoleret ser på sand positiv-raten. Den *tuned* model er således i stand til at prædiktere 50% af *udbruddene af borgerkrig* i testdatasæt, hvor den oprindelige model prædikterede 37,5%. Denne forbedring i sand positiv-raten er sket på bekostning af en højere falsk positiv-rate, hvor at modellen prædikterer flere falske positive.

Tabel 12: Confusion-matrix

Systematisk tuning				Indledende tuning			
	$\hat{y}_i = 1$	$\hat{y}_i = 0$			$\hat{y}_i = 1$	$\hat{y}_i = 0$	
$y_i = 1$	4	4	SPR = 0,50	$y_i = 1$	3	5	SPR = 0,375
$y_i = 0$	330	9922	FPR = 0,032	$y_i = 0$	42	10210	FPR = 0,004

Som jeg har illustreret i dette afsnit, har *tuning* af XGBtree algoritmen stor betydning for modellens generelle performance, samt de resulterende prædiktioner. Her kan *tuning* af antallet af sekventielle klassifikationstræer først og fremmest tilskrives en stor værdi for prædiktionsmodellens *out-of sample* performance. I den endelige *tuning* af XGBtree algoritmen for *udbruddet af borgerkrig*, valgte jeg at fokusere på fire af de frie parametre. *Tuning* af disse parametre, resulterede i en prædiktionsmodel, der udkonkurrerede den oprindelige model, både i forhold til dens generelle *performance* og i de faktiske prædiktioner. *Tuning* betød af AUC-værdien for modellen steg fra 0,79 til 0,88, hvilket vidner om en generel forøgelse i modellens performance. Modellen har dog stadig lang vej til at blive en perfekt prædiktionsmodel.

Foruden en forbedring af modellens generelle performance har *tuning* processen haft den substantielle betydning at sand positiv raten for modellen steg til 0,5 mens falsk positiv raten steg til 0,025. Forbedringen af sand positiv raten kan ikke understreges nok. Fra at korrekt at prædiktere 37,5 % af udbruddene af borgerkrig, kan man med en minimumsindsats af *tuning* forbedrer modellens performance, således at 50% af udbruddene af borgerkrig prædikteres. Dette resultat er igen med til at konsolidere *machine learnings* potentiale i tilretningen af politiske tiltag, når man anskuer borgerkrig som et prædiktionsproblem.

Det fulde potentiale for *tuning* af denne og de tre øvrige prædiktionsmodeller er ikke endeligt indfriet med ovenstående proces. Fremtidige studier af prædiktionen af borgerkrig, bør i højere grad fokusere på *tuning* af algoritmerne, for at skabe de bedste vilkår for den praktiske anvendelse af prædiktionsmodellerne. *Tuning* af XGBtree algoritmen her, understreger denne vigtighed, og kan danne udgangspunktet for en videre afsøgning.

### 7. DISKUSSION

Af min analyse kan der udledes en række forhold, der kan diskuteres og som kan bruges i en perspektivering, i anskuelsen af borgerkrig som et prædiktionsproblem. I dette kapitel vil jeg begrænse denne diskussion til to problemstillinger, der tager udgangspunkt i spørgsmålene: *Har prædiktionsmodellerne en praksis anvendelse?* og *Hvilke teoretiske implikationer resulterer analysen i og rummer nyhedsdata et overset potentiale i samfundsvidenskaben?* I de følgende afsnit vil jeg med udgangspunkt i analysen diskutere netop disse spørgsmål.

#### 7.1. Resultaternes anvendelse og implementering af prædiktionsmodellen

Før at en prædiktionsmodel kan have en praktisk anvendelse, er det for det første nødvendigt at vurdere om resultaterne er anvendelige og dernæst overveje, hvordan at modellerne kan implementeres i en politisk kontekst. Resultaterne er, som jeg har skitseret i afsnit 6.2.3, anvendelige i et vist omfang. Her kan afvejningen mellem *falske-positive* og *falske-negative* justeres gennem de forbundne omkostninger, således at modellen passer til den politiske kontekst. Resultaterne viste, at det er muligt at prædiktere 50% af udbruddene af borgerkrig en måned før de finder sted, om der vil være en borgerkrigslignende tilstand i et land med en måneds varsel, samt at få en indikation af intensiteten af konflikten. Hvis man vælger at acceptere disse resultater som anvendelige for beslutningstagere, vil det næste skridt være at implementere prædiktionsmodellerne i en politisk kontekst. Her rejser der sig tre problemstillinger: *Hvordan kan modellen praktisk implementeres? Hvad er implikationen af dataets beskaffenhed? Efterlader tidsdimensionen af prædiktionerne nok handlerum for beslutningstagerne?*

Det er min overbevisning, at de prædiktionsmodeller jeg har introduceret, har en praksis anvendelse i f.eks. de *early warning* og *risk-assessment* programmer, der findes herunder FNs udviklingsprogram *Conflict and Development* (United Nations Development Program, 2018) værktøj eller som analytisk redskab til udforskning i udviklingen af konflikter f.eks. *Armed Conflict Location & Event Data Project* og *The Uppsala Conflict Data Program* (Armed Conflict Location & Event Data Project, 2018 Allansson, 2018). Der findes med andre ord potentielle aftagere af prædiktionsmodellerne, der kan anvende prædiktionerne i deres arbejde, til bedre at forstå borgerkrig og forebygge de negative følger som borgerkrige har.

Hvis man anskuer anvendelsen af prædiktionsmodellerne og det system jeg har opsat for hele projektet i et praktisk perspektiv, vil det være simpelt og omkostningslet at implementere prædiktionsmodellerne. Det skyldes først og fremmest at jeg har bygget specialet op som et fuldautomatiseret system, hvor at det er muligt for alle med adgang til en computer at eksekvere projektet. Dette er resultatet af, at kodningen i specialet er skrevet i open-source softwaret R (Foundation, 2018), der er let at anvende uden

særlige datalogiske kompetencer. Projektet er sat op således, at man ved at eksekvere *master-scriptet* automatisk indhenter al data, strukturerer det meningsfuldt, udfører selve analyserne og får returneret de fremførte resultater - blandt andre. Den eneste forudsætning er, at man opsætter en PostgreSQL database og tilpasser *argumenterne*, der skriver til denne database forud for eksekveringen.

Det største *men* i denne implementering relaterer sig til dataets beskaffenhed. Som udgangspunkt tager projektet udgangspunkt i tre typer af datakilder: Replikationsdata fra andre studier, GDELT og UCDP-databaserne og data fra *World development indicators*. Hvor at man for de to første datakilder kan forvente en vis stabilitet er sidst nævnte mere ustadig i sin karakter. Selv skriver de:

Many factors affect data availability, reliability, and comparability. Statistical systems in many of the poorest countries are limited; statistical methods, coverage, practices, and definitions differ widely; and cross-country and intertemporal comparisons involve complex technical and conceptual problems that cannot be resolved unequivocally (World Bank, 2018)

Citatet tiltaler kernen i den usikkerhed, der er for datakilden. Denne problematik kan ses som mere alvorlig i studiet af borgerkrig, da netop konflikt og politiske uroligheder kan påvirke datakvaliteten, afrapporteringen af data og dermed også adgangen til data. Denne problematik kan ses som værende iboende borgerkrig og konflikt som genstandsfelt, og der findes umiddelbart ikke nogen god måde at omgå problematikken. Det er derfor min vurdering af man må acceptere denne problematik og anerkende dataet som den bedst mulige løsning til at inddrage relevant data. Som jeg har vist har dataet jo netop en anvendelighed i prædiktionsmodellerne, mens anvendeligheden i kausal-inferens studier er en helt anden.

Indrapporteringen og offentliggørelsen af data til *World development indicators* rummer en anden problematik. I mine modeller bygger prædiktionen af borgerkrig i tiden  $t$  på data fra tiden  $t_{-1}$ . Hvis der er store forsinkelser i offentliggørelsen af data, f.eks. *bruttonationalproduktet* for 2018, vil modellen ikke kunne inddrage dette data i prædiktionen af borgerkrig i 2019, før dataet er tilgængeligt. Prædiktionsmodellen vil fortsat fungere, men i lyset af at flere af variablene med den største *importance* stammer fra *World development indicators* er der en fare for, at modellerne ikke vil kunne *performe* lige så godt, hvis offentliggørelsen af data er forsinket.

Disse to data-relaterede problemstillinger udgør den væsentligste udfordring for implementeringen af prædiktionsmodellerne. Omvendt er problemstillingen håndgribelig og udvidede internationale dataprogrammer, kan således ses som en løsningsmodel i håndteringen af problemstillingen. En løsning, der vil have værdi langt ud over *prædiktionen af borgerkrig* som genstandsfelt.

Den sidste problematik i forhold til prædiktionsmodellernes anvendelse er i forhold til den tidslige dimension modellerne opererer under. For at imødekomme Cederman and Weidmann kritik af de eksisterende prædiktionsstudier temporale opløsning, valgte jeg at strukturere mit data lande-måned niveau. For en observation i tiden  $t$  har jeg således forsøgt at prædiktere responsvariablene for borgerkrig til tiden  $t + 1$ ). Denne model kan siges at have den fordel, at den fokuserer på et snævert tidsinterval i den meget nære fremtid. Ulempen er, at modellen kun giver beslutningstager 1 måned til at forberede og eksekvere forebyggende indsatser. Som en perspektivering analyse vil det i fremtidige studier, være en fordel at se nærmere på prædiktioner, der bruger samme temporale opløsning, men som forsøger at prædiktere længere ud i fremtiden f.eks. 3 eller 6 måneder, da dette ville give beslutningstagerne et større spillerum for at tilrette indsatserne.

I implementeringen af modellen er det nødvendigt at have den institutionelle kontekst i baghovedet. Selvom det internationale samfund har en række redskaber til rådighed, f.eks. fredsmægling og intervention, kan de internationale institutioner ses som den forhindring for tidsaktuelle handlinger. Dette gør sig særligt gældende, når man ser på internationale interventioner fra FNs side. Her har man på den ene side vedtaget en *responsibility to protect*, der bør aktiveres når stater begår overgreb på egen befolkning. Forpligtelsen til at beskytte mennesker bliver ofte problematisk, hvor at der særligt i sikkerhedsrådet er en modvilje for at kompromittere andre landes suverænitet (Luck, 2009; Assembly, 2005).

Her kan prædiktionerne af særligt konflikternes intensitet ses som en væsentlig del af argumentationen. Hvis en ellers troværdig model prædikterer at antallet af døde i den kommende måned vil være enorme, vil det være svært at argumentere imod en intervention, da netop omfanget af spørgsmålet om intervention ofte er en afhængig af, hvor seriøs en trussel der foreligger (Luck, 2009). Jo voldsommere kamphandlinger og overgrebene på befolkningen er, desto sværere er det for sikkerhedsrådet, der sidder med beslutningskompetencen for intervention, at ignorerer begivenhederne. Prædiktionerne kan dermed bruges som et kerneargument i denne afvejning. I min analyse var prædikterede jeg det samlede antal døde, mens det på baggrund af datasættet også ville være muligt at prædiktere antallet civile ofre, der spiller en særlig rolle i *responsibility to protect*.

På baggrund af disse overvejelser må det konkluderes at prædiktionsmodeller, der bygger på offentlig tilgængeligt data og *open-source* kodning, har en praktisk anvendelse i forebyggelsen og deeskaleringen af borgerkrige, mens at det en udnyttelse af det fulde potentiale kræver yderligere undersøgelse. Prædiktionerne kan bruges i argumenta-



tionen for forbyggende indsatser og intervention, hvor særligt et fokus på konflikstens intensitet kan bruges som pejlemærke for interventioner.

## 7.2. Teoretiske implikationer

Når man ser på resultaterne af prædiktionsmodellerne, giver det anledning til særligt to spørgsmål, der relaterer sig til den eksisterende teori. For det første må der spørges, hvordan resultaterne hænger sammen med teorien: *Bliver vi be- eller afkræftet i teoriernes forventninger og bidrager teoriens byggeklodser til prædiktionerne?* Det andet spørgsmål, der kan rejses, er om resultaterne giver anledning til et behov for revision af teorierne og de datakilder, der lægges stor værdi i litteraturen.

Som svar på det første spørgsmål, viste *importance*-analysen, at elementer inden for de 3 præsenterede teoretiske perspektiver, selvstændigt bidrager til prædiktionen af de 4 responsvariable. Hvor at teoriretningerne i kausal-estimations studier kan ses som konkurrerende, kan teorierne forenes i prædiktionsstudier. Her bekræfter mit resultat, at teoriretningerne ikke bør ses som konkurrerende, da de hver i sær bidrager til prædiktionernes præcision illustreret ved variablenes *importance*. Resultaterne af analysen kan hverken bekræfte eller afkræfte de kausale forhold mellem de teoretisk funderede variable og responsvariablene, men vigtigheden af disse variable i prædiktionsmodellerne vidner om, at de kausale forhold som teorien understreger, også har en værdi i prædiktionen af borgerkrig, hvilket må ses som en understøttelse af teori-syntesens argumentation.

Denne påstand understøttes bedst af et par eksempler fra tabel 10. Den *økonomiske vækst* i land tog en gennemsnitlig 8. plads i variablenes *importance* og var den vigtigste ikke konflikt relaterede variable i prædiktionen af *borgerkrigslignende tilstande* og den tredje vigtigste i prædiktionen af *udbrud af borgerkrig*. Dette faktum tjener som argumentation for, at *vækst og konflikt nexus* er en gangbar teoretisk tilgang til studiet af borgerkrig. I relation til det omkostningsframework som *greed og opportunity* litteraturen ligger op til optrådte blandt andet terrænet og indskrivning i sekundærskole som *vigtige variable*. Begge variable, der har betydning for den omkostning, der er forbundet med borgerkrig. For *grievance*-motivet i borgerkrig kan variablene *andel politisk ekskluderet* (gns. 4. plads) og *antal politisk inkluderede grupper* (gns. 11. plads)

En anden relevant indsigt, der kan gøres på baggrund af de netop fremhævede variable er, at de alle har en større betydning i prædiktionen af *udbrud af borgerkrig* og *borgerkrigslignende tilstande*, end de har for de to mål for intensiteten. Omvendt ses det at den række af variable, der udspringer af nyhedsdata og som samlet set udtrykker det politiske landskab, har en lavere rang i prædiktionsmodellerne for *antallet af døde* og *antallet af kamphandlinger*. På baggrund af de teoretiske perspektiver, er det ikke overraskende at netop de strukturelle variable, når det kommer til prædiktionen af udbrud-

og eksistensen af borgerkrigslignede tilstande. Resultatet nuancerer dette billede ved at påvise, at de samfundsstrukturelle variable ikke er gode prædiktorer for intensiteten af borgerkrige.

Hvis man skal gå den eksisterende litteratur og lovpriste studier såsom Fearon and Laitin, 2003; Collier and Hoeffler, 2004 på klingen, er en implikation af min analyse, at disse studier har et videnshul i deres afsøgning af forudsætningerne og motiverne bag borgerkrig. Dette hul i litteraturens afsøgning af motiverne og forudsætningerne skal ses i lyset af den *importance* som henholdsvis et lands nære konflikthistorie og informationen om det politiske landskab har for prædiktionerne.

Først og fremmest fylder et lands nære konflikthistorie forsvindende lidt i litteraturen. Typisk inddrages der et eller to *dummy-variable*, der udtrykker om et land har oplevet konflikt inden for de seneste 5-10 år eller en kontinuert variabel, der udtrykker antallet af forløbende år med fred. Disse variable har en relevans, men bidrager ikke med yderligere information om mere partikulære forhold, der kan være afgørende for borgerkrig. Når det kommer til prædiktion af borgerkrig, har min analyse vist, at den nære konflikthistorie udgør de vigtigste variable for prædiktionernes performance og præcision.

Den nære konflikthistorie kan tænkes at spille en særlig rolle i en *proces tracing* af transitionen mellem forskellige stadier af konflikt, der kan indarbejdes i kausal-estimationsstudier. En sådan *proces tracing* ville skabe et stærkt basis for kausal inferens ved at illustrere ubrudte links mellem forløbende voldelige oprør og borgerkrig. *Proces tracingen* af transitionen mellem stadier af konflikt ville dermed kunne specificere forskellige forhold, der bør indarbejdes i kausal inferens studierne. Den største hæmsko for at indarbejde *den nære konflikthistorie* i studier af kausal inferens er variabelenes endogene karakter. Før man kan identificere den kausale sammenhæng mellem en uafhængig variabel og en afhængig variabel, er det nemlig nødvendigt at ingen af de faktorer, der korrelerer med de uafhængige variable, korrelerer med den afhængige variabel (Hariri, 2012, s. 185).

Det datasæt jeg har dannet og introduceret på baggrund af GDELT-databasen, kan ses som en præcisering af de forhold som særligt *grievance*-motivet for borgerkrig fokusere på. Som beskrevet har litteraturen i operationaliseringen af *grievance* hovedsageligt anvendt samfundsstrukturelle og demografiske forhold til at måle fænomenet. Det har haft den ulempe, at variablerne altid har udtrykket en større sandsynlighed for *grievance*-motivet i mere diverse samfund, uden at denne sammenhæng har kunnet fastslås.

GDELT-databasen og det introducerede aggregerede datasæt er her en markant bidrag,

da dataet blandt andet udtrykker den faktiske *grievance* i et samfund frem for den forventede. Ved at tage udgangspunkt i begivenheder af politisk karakter, der blandt andet kan udtrykke politisk utilfredshed dannes der et faktisk mål for den politiske utilfredshed og den afledte *grievance*, der er i et land. Dette konkrete mål behøves ikke nødvendigvis anses som at være de demografiske og samfundsstrukturelle variable overlegne, men må som minimum anerkendes som et nødvendigt supplement til disse, hvis man for alvor er interesseret i at forstå de motiver, der er for borgerkrig og oprør.

I min analyse af variablenes *importance* stod flere af de mål, der udtrykker en faktisk *grievance* i befolkning frem som centrale variable for prædiktionernes *performance*. Samtidig var disse *grievance* orienterede variables modsætning, der udtrykker et politisk samarbejde, ligeledes centrale variable i analysen. Variablenes vigtighed i prædiktionen har den implikation for teorien, at man bør tage teoriens primære datakilder op til revision.

Dataet har den fordel, at det eksisterer på det mest disaggregerede niveau, nemlig observationer for hver enkelt politiske begivenhed og rummer mange forskellige muligheder for sammensætningen. I min analyse lavede jeg således en 4 felts gruppering af data, mens at andre sammensætninger, der specificerer forhold omkring motivet i de politiske modarbejde er mulig på baggrund af typen af begivenheder<sup>24</sup>. Her vil det blandt andet være muligt at sammensætte et datasæt, der fokuserer på en oppositionens eller et mindretals afvisning af politiske tiltag, protester, trusler for at indskærpe *grievance* motivet for borgerkrig. Med udgangspunkt i GDELT-databasen anvendelse af nyhedsdata er det med andre ord muligt at videreføre teorien og undersøge de nærmere links mellem *grievance* motivet og borgerkrig.

---

<sup>24</sup>Se liste over begivenhedstyper i appendiks D.

### 8. KONKLUSION

Hvilket potentiale har *machine learning* til at målrette tiltag mod borgerkrig, hvis man anskuer dette som et prædiktionsproblem? Dette spørgsmål har været omdrejningspunktet for nærliggende speciale. Gennem en teorifunderet og metodisk eksplorativ tilgang, har jeg i min besvarelse udarbejdet en række prædiktionsmodeller, der har forsøgt at prædiktere *udbruddet af borgerkrig, eksistensen af borgerkrigslignende tilstande, antallet af døde i interne konflikter og antallet af kamphandlinger i interne konflikter*.

Undersøgelsen har haft et todelt formål. For det første har undersøgelsen haft til formål at evaluere prædiktionsmodellernes potentiale, gennem en afdækning af *machine learning* algoritmernes *performance* sammenholdt med en række baseline modeller. Undersøgelsens andet formål har været på den praktiske anvendelse af prædiktionsmodeller i forebyggelsen og deeskaleringen af borgerkrige. Som et særskilt fokus i min analyse, har jeg introduceret et nyt datasæt over politiske begivenheder, der danner udgangspunktet for en diskussion af analysens teoretiske implikationer.

Prædiktion som tilgang medfører et skift i fokus væk fra estimation og inferens. Formålet i estimationsstudier er at udforske *unbiased* kausale forhold gennem en maksimering af en models *in sample performance*, mens det i prædiktionsstudier er at maksimere modellens *out-of sample performance*. I den forbindelse har jeg diskuteret, hvordan at prædiktionsmodeller bør evalueres. Her introducerede jeg henholdsvis *out-of sample*  $R^2$  og *Root Mean Square Error* som evalueringskriterier for kontinuerte variable, og AUC-værdier, confusion-matrixer og densitetshistogrammer som evalueringskriterier i evalueringen af binære variable.

Som resultat af min analyse fandt jeg, at den træ-baserede ensemble algoritme *Extreme Gradient Boosted Trees* generelt udkonkurrerede både baseline- og Random Forest modellen. Med en AUC-værdi på 0,79 prædikterede EGBtree algoritmen 37,5 % af *udbruddene af borgerkrig* i testdatasættet, mod 12,5% og 25 % i henholdsvis baseline- og Random Forest algoritmen. De falske positive i EGBTree algoritmes prædiktioner udgjorde ved nærmere eftersyn ikke en alvorlig fejlkilde, da observationerne bestod af lande, der enten førhen eller sidenhen har oplevet borgerkrige eller internationaliserede interne konflikter. I prædiktionen af *borgerkrigslignende tilstande* var AUC-værdien for modellen 99,3%, hvilket kun var en marginal forbedring af de øvrige modeller. Resultatet vidnede om, at det ikke kun er muligt at prædiktere udbruddet af borgerkrige, men også sporadiske tilfælde af voldelige interne uroligheder, samt kontinuerte interne borgerkrigslignende tilstande, der ikke nødvendigvis udvikler sig til regulære borgerkrige.

Prædiktionen af intensiteten i interne konflikter havde det formål, at prædiktionerne

---

kunne bruges som informationsgrundlag i valget af “præventive initiativer”, herunder en skalering af internationale interventioner. Her prædikterede XGBtree modellen i gennemsnit  $\pm 65$  døde forkert, og forklarede på baggrund heraf 82 % af variationen i antallet af døde i test-datasættet. I prædiktionen af antallet af kamphandlinger prædikterede algoritmen  $\pm 4,2$  kamphandlinger forkert med en forklaringskraft på 94% i test-datasættet. Resultaterne kan på trods af deres upræcighed bruges som argument og datagrundlag for interventioner og skaleringen af selvsamme.

Min analyse viste yderligere, at *tuningen* af XGBtree algoritmens frie parametre havde vidtrækkende konsekvenser for dennes performance for prædiktionen af *udbrud af borgerkrig*. *Tuning*-processen forbedrede prædiktionsmodellen således, at 50% af udbruddene af borgerkrig blev korrekt prædikteret. Et resultat, der ikke kan understreges nok.

På baggrund af den samlede analyse konkluderede jeg, at *machine learning* har et potentiale i tilretningen af tiltag mod borgerkrig, når man anskuer dette som et prædiktionsproblem. Dette skyldes særligt, at modellerne i vidt omfang er i stand til at prædiktere *udbrud* og *borgerkrigslignende tilstande* med en måneds varsel. Samtidig kan prædiktionsmodellerne give et konkret bud på konflikternes intensitet, med en kendt usikkerhed.

Anvendelsen af prædiktionsmodeller, som et værktøj i forhindringen af borgerkrig, er først og fremmest afhængig af validiteten af resultaterne, og dernæst af den praktiske implementering af modellerne i en institutionel beslutningsproces. For at imødekomme en let og omkostningsfri implementering af prædiktionsmodellerne, har jeg i min fremgangsmåde haft fokus på offentlig tilgængelige data og åben kodning. Resultatet er et samlet *projekt* for prædiktion af borgerkrige, der omkostningsfrit kan integreres overalt og som ikke kræver specialiseret viden eller yderligere kodning. Den konkrete anvendelse af prædiktionsmodellerne rummer en række udfordringer, der særligt relaterer sig til indretningen af det internationale system, herunder aktørernes indbyrdes relationer, sammensætningen og dynamikkerne i FNs sikkerhedsråd og en moralsk afvejning mellem staters suverænitet og det internationale samfunds *responsibility to protect*.

Fremfor at anlægge et rent datadrevet fokus i tilrettelæggelsen af analysen, har jeg taget udgangspunkt i den eksisterende litteratur i indsamlingen af data. Denne tilgang understreger teoriens fortsatte centrale placering, også indenfor *machine learning*, der ellers ofte bliver beskrevet som en teoretisk *black box*. Teoriens værdi og bidrag til prædiktionsmodellerne blev for alvor tydelig i analysen af de mest betydningsfulde variable. Her stod en række af de teoretisk funderede variable frem, og illustrerede vigtigheden af de motiver og forudsætninger, som den teoretiske litteratur lægger vægt

på.

Analysen af variabelnes betydning gav anledning til afledt diskussion om de teoretiske implikationer af analysen. Analysen viste først og fremmest, at et lands nære konflikthistorie har den største betydning for prædiktionernes præcision. Konflikthistorien for et land fylder forsvindende lidt i de kvantitative studier og borgerkrig. Resultatet giver dermed anledning til revision af måden, hvorpå at studierne udføres således at man ikke risikerer en *omitted variable bias* grundet udeladelsen af disse variable.

I analysen af variabelnes relative vigtighed var det tydeligt, at det introducerede datasæt over politiske begivenheder også bidrag substantielt til analysen. Dette faktum i samspil datasættets teoretiske indpasning i *grievance* litteraturen, giver en forhåbning om, at datasættet i fremtiden kan være med til at udforske de kausale forhold og understøtte teoriretningen på en konstruktiv måde.

Selvom prædiktion af borgerkrige og andre politiske begivenheder ikke kan ses som en eksakt videnskab, har *machine learning*, nye datakilder og voksende datamængder været med til at flytte feltet. Fra førhen at være forudsigelser, er prædiktion af politiske begivenheder, såsom borgerkrig, i dag et troværdigt redskab, der kan implementeres i en institutionel *setting* og være med til at guide beslutninger og sandsynliggøre udfald.

Borgerkrig er fortsat et alvorligt politisk problem i kraft af dets direkte effekt på befolkningens liv og velstand. Borgerkrig udgør desuden en fare for stabiliteten i det internationale system og har negative internationale eksternaliteter, i form af transnationale oprørsgrupper og flygtningestrømme. Ved at anskue borgerkrig som et prædiktionsproblem, har jeg i dette speciale forsøgt at give et bud på, hvordan det internationale samfund kan blive bedre til at håndtere og forebygge oprør og borgerkrige.

---

## LITTERATURLISTE

- Acemoglu, Daron, Simon Johnson, and James A. Robinson (2005). "INSTITUTIONS AS A FUNDAMENTAL CAUSE OF LONG-RUN GROWTH". In: *Handbook of economic growth*. Ed. by Daron Acemoglu, Simon Johnson, and James A. Robinson. 1st ed. Handbooks in economics 22. Amsterdam: Elsevier.
- Allansson, Marie (2018). *About UCDP - Department of Peace and Conflict Research - Uppsala University, Sweden*. URL: <http://www.pcr.uu.se/research/ucdp/about-ucdp/> (visited on 12/18/2018).
- Amin, Samir (1976). *Unequal development: an essay on the social formations of peripheral capitalism*. New York, N.Y.: Monthly Review Press.
- Andersen, Jorgen and Silje Aslaksen (2008). "Constitutions and the resource curse". In: *Journal of Development Economics* 87.2.
- Andrews, Molly (2014). "Narrating Moment of Political Change". In: *The Palgrave handbook of global political psychology*.
- Armed Conflict Location & Event Data Project, Armed Conflict Location & Event Data Project (2018). *About ACLED*. URL: <https://www.acleddata.com/about-acled/> (visited on 12/18/2018).
- Assembly, UN The General (2005). *Resolution adopted by the General Assembly - 60/1. 2005 World Summit Outcome*.
- Bøgh Andersen, Lotte, Anne Skokjær Binderkrantz, and Kasper Møller Hansen (2010). "Forskningsdesign". In: *Metoder i statskundskab*. Ed. by Lotte Bøgh Andersen, Robert Klemmensen, and Kasper Møller Hansen. 2nd. Hans Reitzels Forlag.
- Bühlmann, Peter Lukas and Bin Yu (2000). *Explaining bagging*.
- Blair, Robert, Christopher Blattman, and Alexandra Hartman (2016). "Predicting Local Violence: Evidence from a Panel Survey in Liberia". In: *SSRN Electronic Journal*.
- Brandt, Patrick T., John R. Freeman, and Philip A. Schrodtt (2011). "Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict". In: *Conflict Management and Peace Science* 28.1, pp. 41-64.
- Breiman, Leo (2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statist. Sci.* 16.iss.3, pp. 199-231. ISSN: 0883-4237.
- Caselli, Francesco and Tom Cunningham (2017). *Political Decision Making in Resource Abundant Countries*. Tech. rep. Paper presented at the Oxcarre Conference, Oxford University: Oxford University.
- Cederman, Lars-Erik and Nils B. Weidmann (2017). "Predicting armed conflict: Time to adjust our expectations?" en. In: *Science* 355.6324, pp. 474-476.
- Chen, Tianqi et al. (2018). *xgboost: Extreme Gradient Boosting*. URL: <https://CRAN.R-project.org/package=xgboost> (visited on 12/07/2018).
- Collier, P. and Anke Hoeffler (2004). "Greed and grievance in civil war". In: *Oxford Economic Papers* 56.4, pp. 563-595.

- Collier, Paul (2007). *The bottom billion: why the poorest countries are failing and what can be done about it*. Oxford: Oxford University Press.
- Collier, Paul, Anke Hoeffler, and Måns Söderbom (2008). "Post-Conflict Risks". In: *Journal of Peace Research* 45.4, pp. 461–478.
- Croicu, Mihai and Ralph Sundberg (2017). *UCDP GED Codebook version 18.1*.
- Dahl, Robert Alan (1991). *Democracy and its critics*. 12. print. New Haven: Yale Univ. Press.
- Dicken, Peter (2011). *Global shift: mapping the changing contours of the world economy*. 6th ed. New York: Guilford Press.
- Esteban, Joan-María and Debraj Ray (1994). "ON THE MEASUREMENT OF POLARIZATION". eng. In: *Econometrica* (1986-1998) 62.4.
- Fearon, James D. and David D. Laitin (2003). "Ethnicity, Insurgency, and Civil War". In: *American Political Science Review* 97.01, pp. 75–90. (Visited on 08/09/2018).
- Foundation, The R (2018). *R: What is R?* URL: <https://www.r-project.org/about.html> (visited on 11/27/2018).
- GDELT (2013). *GDELT – DATA FORMAT CODEBOOK V 1.03 8/25/2013*. URL: [http://data.gdeltproject.org/documentation/GDELT-Data\\_Format\\_Codebook.pdf](http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf).
- (2018). *The GDELT Story: About the GDELT Project*. URL: <https://www.gdeltproject.org/about.html#creation> (visited on 08/30/2018).
- Gerner, Deborah J. et al. (2002). "Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions". en. In: URL: <http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=87C5EF2966F21BC26BF50734B66D95C4?doi=10.1.1.6.5097> (visited on 08/02/2018).
- Goodman, John B. and Louis W. Pauly (1993). "The Obsolescence of Capital Controls?: Economic Management in an Age of Global Markets". In: *World Politics* 46.01, pp. 50–82. (Visited on 02/05/2018).
- Gorman, Ben (2017). *A Kaggle Master Explains Gradient Boosting*. URL: <http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/> (visited on 12/03/2018).
- Hammack, Phillip H. and Andrew Pilecki (2014). "Methodological Approaches in political psychology: Discourse and narrative". In: *The Palgrave handbook of global political psychology*. ISBN: 978-1-137-29118-9.
- Hariri, Jacob Gerner (2012). "Kausal inferens i statskundskaben". In: *Politica* 44.2, pp. 184–201.
- Harvard University, Education (2018). *Harvard WorldMap*. URL: <http://worldmap.harvard.edu/> (visited on 11/26/2018).
- Hastie, Trevor, Robert Tibshirani, and Jerome H Friedman J. H. (2017). *The elements of statistical learning, data mining, inference, and prediction*. 2. ed. Springer series in statistics. New York: Springer.
- Hegre, Håvard et al. (2013). "Predicting Armed Conflict, 2010-2050: Predicting Armed Conflict". In: *International Studies Quarterly* 57.2, pp. 250–270.



- Hobbes, Thomas (2012). *Leviathan*. Minola, N.Y.: Dover Publications.
- Hviid Pedersen, Søren (2015). "Thomas Hobbes". In: *Klassisk og moderne politisk teori*. Ed. by Lars Bo Kaspersen and Jørn Loftager. Kbh.: Hans Reitzel, pp. 190–203.
- James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York.
- King, Gary, Robert O Keohane, and Sidney Verba (1995a). *Designing social inquiry: scientific inference in qualitative research*. Princeton University Press. (Visited on 11/02/2018).
- King, Gary, Robert O. Keohane, and Sidney Verba (1995b). "The Importance of Research Design in Political Science". In: *American Political Science Review* 89.02, pp. 475–481.
- Klemmensen, Robert, Lotte Bøgh Andersen, and Kasper Møller Hansen (2010). "At lave undersøgelser inden for statskundskab". In: *Metoder i statskundskab*. Ed. by Lotte Bøgh Andersen, Robert Klemmensen, and Kasper Møller Hansen. 2nd. Kbh: Hans Reitzels Forlag.
- Knox, Steven W (2018). *Machine learning, a concise introduction*. eng. Wiley series in probability and statistics. Hoboken, NJ: John Wiley and Sons, Inc.
- Kuhn, Thomas S (2012). *The structure of scientific revolutions*. Fourth ed. University of Chicago Press.
- Lahiri, Bidisha and Feroz K Masjidi (2012). "Landlocked Countries". In: *Journal of Economic Integration* 27.4, pp. 505–519. ISSN: 1225651X. URL: <http://search.proquest.com/docview/1266134632/>.
- Leetaru, Kalev and Philip A. Schrodt (2013). *GDELT: Global Data on Events, Location and Tone*. URL: <http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf>.
- Liaw, Andy and Matthew Wiener (2002). *Classification and Regression by randomForest*.
- Luck, Edward (2009). "Sovereignty, Choice, and the Responsibility to Protect". In: *Global Responsibility to Protect* 1.1, pp. 10–21.
- Lujala, Päivi, Nils Gleditsch, and Elisabeth Gilmore (2005a). "A Diamond Curse? CIVIL WAR AND A LOOTABLE RESOURCE". In: *The Journal of Conflict Resolution* 49.4, pp. 538–562. ISSN: 00220027.
- Lujala, Päivil, Nils Petter Gleditsch, and Elisabeth Gilmore (2005b). "A Diamond Curse? Civil War and a Lutable Resource". In: *The Journal of Conflict Resolution* 49.4, pp. 538–562.
- Marshall, Monty G., Ted Robert Gurr, and Keith Jagers (2016). *POLITY IV PROJECT - Dataset Users' Manual. Political Regime Characteristics and Transitions, 1800-2016*. URL: <http://www.systemicpeace.org/>.
- McLean, Iain and Alistair McMillan (2009). *The concise Oxford dictionary of politics*. 3. ed. Oxford University Press.
- McNamara, Kathleen (2002). "Rational Fictions: Central Bank Independence and the Social Logic of Delegation". In: *West European Politics* 25.1, pp. 47–76.
- Mesquita, Bruce Bueno de (2011). "A New Model for Predicting Policy Choices: Preliminary Tests". In: *Conflict Management and Peace Science* 28.1, pp. 65–87.

- Montalvo, Jos G and Marta Reynal-Querol (2005). "Ethnic Polarization, Potential Conflict, and Civil Wars". In: *American Economic Review* 95.3, pp. 796–816.
- Muchlinski, David et al. (2016). "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data". In: *Political Analysis* 24.01, pp. 87–103.
- Mueller, Hannes and Christopher Rauh (2018). "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text". en. In: *American Political Science Review* 112.02, pp. 358–375.
- Murshed, Syed Mansoob (2009). *Explaining civil war: a rational choice approach*. Cheltenham, UK ; Northampton, MA: Edward Elgar.
- Perry, Chris (2013). "Machine Learning and Conflict Prediction: A Use Case". en. In: *Stability: International Journal of Security & Development* 2.3, p. 56.
- Pitcher, M. Anne (2017). "Entrepreneurial Governance and the Expansion of Public Investment Funds in Africa". In: *Africa in world politics: constructing political and economic order*. Ed. by John W. Harbeson and Donald S. Rothchild. Sixth edition. Boulder, CO: Westview Press.
- Prebisch, Raúl (1950). *The Economic Development of Latin America – and its principal problems*. Lake Success, New York: United Nations – Department of Economic Affairs.
- Rüegger, Seraina and Heidrun Bohnet (2018). "The Ethnicity of Refugees (ER): A new dataset for understanding flight patterns<sup>1</sup>". In: *Conflict Management and Peace Science* 35.1, pp. 65–88.
- Reno, William (2017). "The International Factor in African Warfare". In: *Africa in world politics: constructing political and economic order*. Ed. by John W. Harbeson and Donald S. Rothchild. Sixth edition. Westview Press.
- Rost, Nicolas, Gerald Schneider, and Johannes Kleibl (2009). "A global risk assessment model for civil wars". In: *Social Science Research* 38.4, pp. 921–933.
- Rostow, W. W (1991). *The Stages of Economic Growth, A Non-Communist Manifesto*. 3. ed. Cambridge: Cambridge University Press.
- Sandvik, Bjørn (2018). *World Borders Dataset*. URL: [http://thematicmapping.org/downloads/world\\_borders.php](http://thematicmapping.org/downloads/world_borders.php) (visited on 12/15/2018).
- Selimovic, Johanna Mannergren (2014). "Narrative Constructions of Conflict and Coexistence: The Case of Bosnia-Herzegovina". English. In: *The Palgrave handbook of global political psychology*, pp. 388–402.
- Sidnik, D and S Condor (2014). "Social Identity Theory and Self-Categorization Theory". In: *The Palgrave handbook of global political psychology*, pp. 39–54.
- Soysa, Indra De and Eric Neumayer (2007). "Resource Wealth and the Risk of Civil War Onset: Results from a New Dataset of Natural Resource Rents, 1970–1999". In: *Conflict Management and Peace Science* 24.3, pp. 201–218.
- Stubager, Rune and Kim Mannemar Sønderskov (2011). *Forudsætninger for lineær regression og variansanalyse efter mindste kvadraters metode*. Danish. Aarhus: Institut for Statskundskab, Aarhus Universitet.

- Sundberg, Ralph and Erik Melander (2013). "Introducing the UCDP Georeferenced Event Dataset". In: *Journal of Peace Research* 50.4, pp. 523–532.
- Taylor, Ian (2017). "The Sad Story of "Africa Rising" and the Continent's Romance with the BRICS". In: *Africa in world politics: constructing political and economic order*. Ed. by John W. Harbeson and Donald S. Rothchild. Sixth edition. Westview Press.
- The World Bank, The World Bank (2018). "World development indicators". In: URL: <http://wdi.worldbank.org/tables>.
- United Nations Development Program, United Nations Development Program (2018). *Conflict analysis and risk assessment*. URL: <http://www.undp.org/content/undp/en/home/ourwork/democratic-governance-and-peacebuilding/conflict-prevention-and-peacebuilding/conflict-analysis-and-risk-assessment.html> (visited on 12/18/2018).
- United Nations Global Pulse, United Nations Global Pulse (2018). *Annual Report 2017 - Harnessing Big Data for Development and Humanitarian Action*.
- Varian, Hal R. (2014). "Big Data: New Tricks for Econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.
- Volkan, Vamik (2005). "Politics and international Relationships". In: *The American Psychiatric Publishing textbook of psychoanalysis*. Ed. by Ethel Spector Person, Arnold M. Cooper, and Glen O. Gabbard. 1st ed. American Psychiatric Pub, pp. 525–533.
- W Beck, Marcus et al. (2018). "R Package imputeTestbench to Compare Imputation Methods for Univariate Time Series". English. In: *The R Journal* 2018.10, pp. 218–233. URL: <https://journal.r-project.org/archive/2018/RJ-2018-024/RJ-2018-024.pdf>.
- Wickham, Hadlet et al. (2018). *Smoothed density estimates*. URL: [https://ggplot2.tidyverse.org/reference/geom\\_density.html](https://ggplot2.tidyverse.org/reference/geom_density.html) (visited on 12/10/2018).
- Wig, Tore (2016). "Peace from the past: Pre-colonial political institutions and civil wars in Africa". In: *Journal of Peace Research* 53.4, pp. 509–524.
- Wing, Max Kuhn Contributions from Jed et al. (2018). *caret: Classification and Regression Training*. URL: <https://CRAN.R-project.org/package=caret>.
- World Bank, World Bank (2018). *WDI - Sources and Methods*. URL: <http://datatopics.worldbank.org/world-development-indicators/sources-and-methods.html> (visited on 12/18/2018).

---

## A. APPENDIKS

### B. DANNELSEN AF GLOBAL DATABASE OF EVENTS, LANGUAGE AND TONE DATASÆTTET

Dette appendix beskriver hvorledes at GDELT datasættet er blevet dannet og nogle af de metodiske spørgsmål, der kan opstå i forbindelse med behandlingen af nyheder som datakilde. Jeg har opdelt denne gennemgang i tre dele: (1) Udvælgelse af nyhedsskilder og nyheder, (2) Behandling og geokodning, (3) Tekstanalysen. Beskrivelsen tager udgangspunkt i Leetaru and Schrodtt, 2013; Gerner et al., 2002; GDELT, 2013

#### B.1. Udvælgelse af hentning af nyhedsdata

Metoden hvormed at GDELT-projektet danner sit datasæt er en kompliceret proces, der anvender flere af de metodikker som *data science* kan bidrage til samfundsvidenskaben med.

Det første skridt på vejen til at danne GDELT datasættet er, at indhente nyheder. Her har projektet fokuseret sin indhentning af nyhedsartikler på 13 medier. De 13 medier er udvalgt grundet deres store internationale afdelinger og det brede fokus de lægger i nyhedsdækning. Blandt medierne optræder blandt andre Washington Post, AfricaNews og BBC Monitoring (Leetaru and Schrodtt, 2013, s. 2) Nyhedsdataet er hentet gennem LexisNexus, hvor at forskellige SQL-forespørgsler er benyttet til at fokusere på de 13 mediers internationale sektioner.

Der kan utvivlsomt stilles spørgsmålstejn ved hvilken betydning det har for identifikationen af begivenheder, at netop disse 13 medier er udvalgt. Andre nyhedsmedier ville givetvis have ført til identifikationen af andre politiske begivenheder eller en anden repræsentation af selvsamme. Hvorvidt medierne er over- og underrepræsentere bestemte dele af verden er blevet undersøgt til projektets fordel. Eksempelvis kan det forventes at nyhedsmediet Agence France Presse ville over-repræsentere "franske affærer". Dette viser sig dog ikke at være tilfældet, når man sammenligner med, hvordan at andre nyhedsmedier for eksempel Reuters dækker afrikanske lande (Leetaru and Schrodtt, 2013, s. 4). Samme resultat findes når man internt kigger på, hvor det geografiske fokus ligger. Her korrelerer Xinhua og Associated Press med et  $r = 0,95$ , hvilket indikerer at der ikke er nogen signifikant forskel på, hvordan nyhedsdækningen optræder (Leetaru and Schrodtt, 2013, s. 11).

#### B.2. Efterbehandling og Geokodning

Da terminologien i sportsnyheder & finansverdenen benytter samme termer som optræder i den politiske sfære, eksempelvis er ordet *kampe* almindelige i både sportsny-

heder samt politiske nyheder, har der været et behov for at frasortere disse. Hvis ikke at disse nyheder bliver frasorteret ville de i den efterfølgende kodning af artiklerne være med til at give et misvisende billede af det politiske landskab. Det skyldes blandt andet at rivaliserende sportsholds møder ofte typisk beskrives som *kampe* frem for spil, mens at en virksomheds aktiekurs kan være under *belejring*. Dette er blot for at sige, at nogen finansielle nyheder, der kunne have en politisk relevans fejlagtigt, er blevet frasorteret, mens at sportsnyheder, der generelt må siges at have ringe relevans for nærliggende speciales er frasorteret.

Det næste skridt i dannelsen af datasættet har været at geokode nyhederne på baggrund af referencer til steder og aktører. Alt efter hvor specifikke nyhederne har været i deres beskrivelser har dette ført til meget forskellige grader af præcision i geokodningen af data. Nogle af geokodningen vil være med en meget høj præcision i sit mål for længde- og breddegrad, mens andre begivenhederne kun kan siges at være foregået inden for et land.

### B.3. Tekstanalyse og dannelsen af observationer

Gennem tekstanalyseprogrammet TABARI analysere GDELT projektet de hentede artiklers indhold og skaber et nyt datasæt, der følge det gængse CAMEO codebook. Processen, ordbøgerne og processorkraften bag tekstanalysen er ganske omfangsrigt og det ligger uden for dette speciales formål at redegøre for hver enkelt detalje i processen. I grove træk finder TABARI frem til aktørerne i en artikel, hvad handlingen er mellem aktørerne er, og hvornår denne er foregået. Som et eksempel på den politiske utilfredshed som jeg søger i datasættet kan eksempelvis fremhæves eventkoden 1412, der dækker over en begivenhed, hvor der demonstreres for politisk forandring. Som eksempel på, hvordan denne tekstanalyse foretages eksemplificeres med følgende uddrag af en artikel:

Tens of thousands of **university students throughout Indonesia** staged **mass demonstrations** Saturday to **demand political reforms** by **President Suharto's government** (Gerner et al., 2002, s. 67)

Her vil det være de studerende, der er den handlende aktør som laver en handling, i dette tilfælde en demonstration for politiske reformer. Handlingen kan siges at være mod regeringen og denne identificeres således som den anden aktør i begivenheden. Et andet givende eksempel findes i eksemplet omhandlende handlingen "afvisning af at ændre det politiske lederskab":

**Vice-President Moody Awori** has **declined to resign** despite growing pressure by the **Kenya Anti-Corruption Commission** after he was implicated in a major scandal. (Gerner et al., 2002, s. 57)

På samme hvis finder den automatiske tekstanalyse frem til her, at handlingen "declined to resign" udføres af Moody Awai, at han tilhører regeringen og det er The Kenya Anti-Corruption Commission, der har lagt pres på ham. Tekstanalysen er ikke kun i stand til at identificere hvem gør hvad mod hvem, hvor og hvornår. Tekstanalysen har også inkorporeret et værktøj, der måler hvilken tone der er i artiklen, målt på en skala der går fra meget negativ (-100) til meget positiv (+100) alt efter hvilke ord der optræder i teksten. Tonen kan bruges til at differentiere mellem begivenheder af den samme type. Her må det forventes at begivenheder med en meget negativ tone har en større politisk slagkraft end begivenheder der beskrives i mere neutrale toner (Leetaru and Schrodt, 2013, s. 46)

Selvsagt findes der et hav af forskellige typer af begivenheder som tekstanalysen finder frem til. Nogle er ganske specifikke som dem jeg har eksemplificeret ovenfor, mens andre er af en mere generel karakter. I min anvendelse af data har jeg fravalgt alle de typer af begivenheder, der er direkte forbundet med vold, da en medtagelse af disse ville give et tautologisk forhold til min afhængige variable, der netop er den afhængige variable.

### C. GRUPPERING AF GDELT DATA

Nedenstående SQL-kommando har været blevet brugt til at danne det grupperede GDELT datasæt

```
1 CREATE TABLE public.gdelt_group AS
2 SELECT
3     "ActionGeo_CountryCode" as Country
4     , "year"
5     , "month"
6
7     , AVG(cast(NULLIF("q1at", o) AS BIGINT)) as q1at
8     , SUM(CASE WHEN "QuadClass" = 1 THEN 1 ELSE 0 END) as q1cnt
9     , AVG(cast(NULLIF("q2at", o) AS BIGINT)) as q2at
10    , SUM(CASE WHEN "QuadClass" = 2 THEN 1 ELSE 0 END) as q2cnt
11    , AVG(cast(NULLIF("q3at", o) AS BIGINT)) as q3at
12    , SUM(CASE WHEN "QuadClass" = 3 THEN 1 ELSE 0 END) as q3cnt
13    , AVG(cast(NULLIF("q4at", o) AS BIGINT)) as q4at
14    , SUM(CASE WHEN "QuadClass" = 4 THEN 1 ELSE 0 END) as q4cnt
15
16    —Politiske begivenheder med religis betoning
17    , AVG(cast(NULLIF("relq1at", o) AS BIGINT)) as relq1at
18    , SUM(CASE WHEN "QuadClass" = 2 AND "rel"= 1 THEN 1 ELSE 0 END) as relq1cnt
19    , AVG(cast(NULLIF("relq2at", o) AS BIGINT)) as relq2at
20    , SUM(CASE WHEN "QuadClass" = 2 AND "rel"= 1 THEN 1 ELSE 0 END) as relq2cnt
21    , AVG(cast(NULLIF("relq3at", o) AS BIGINT)) as relq3at
22    , SUM(CASE WHEN "QuadClass" = 3 AND "rel"= 1 THEN 1 ELSE 0 END) as relq3cnt
23    , AVG(cast(NULLIF("relq4at", o) AS BIGINT)) as relq4at
24    , SUM(CASE WHEN "QuadClass" = 4 AND "rel"= 1 THEN 1 ELSE 0 END) as relq4cnt
25
26    —Politiske begivenheder med etnisk betoning
27    , AVG(cast(NULLIF("ethq1at", o) AS BIGINT)) as ethq1at
28    , SUM(CASE WHEN "QuadClass" = 2 AND "eth"= 1 THEN 1 ELSE 0 END) as ethq1cnt
29    , AVG(cast(NULLIF("ethq2at", o) AS BIGINT)) as ethq2at
30    , SUM(CASE WHEN "QuadClass" = 2 AND "eth"= 1 THEN 1 ELSE 0 END) as ethq2cnt
31    , AVG(cast(NULLIF("ethq3at", o) AS BIGINT)) as ethq3at
32    , SUM(CASE WHEN "QuadClass" = 3 AND "eth"= 1 THEN 1 ELSE 0 END) as ethq3cnt
33    , AVG(cast(NULLIF("ethq4at", o) AS BIGINT)) as ethq4at
34    , SUM(CASE WHEN "QuadClass" = 4 AND "eth"= 1 THEN 1 ELSE 0 END) as ethq4cnt
35
36 FROM public.gdelt_data
37 GROUP BY "ActionGeo_CountryCode"
38     , "year"
39     , "month"
40 ORDER BY "ActionGeo_CountryCode"
41     , "year"
42     , "month"
```

---

## D. GDELT BEGIVENHEDSTYPER

Liste over begivenhedstyper inkluderet i GDELT-datasættet

01: MAKE PUBLIC STATEMENT	025: Appeal to yield, not specified below
010: Make statement, not specified below	0251: Appeal for easing of administrative sanctions
011: Decline comment	0252: Appeal for easing of political dissent
012: Make pessimistic comment	0253: Appeal for release of persons or property
013: Make optimistic comment	0254: Appeal for easing of economic sanctions, boycott, or embargo
014: Consider policy option	0255: Appeal for target to allow international involvement (non-mediation)
015: Acknowledge or claim responsibility	0256: Appeal for de-escalation of military engagement
016: Deny responsibility	026: Appeal to others to meet or negotiate
017: Engage in symbolic act	027: Appeal to others to settle dispute
018: Make empathetic comment	028: Appeal to engage in or accept mediation
019: Express accord	
02: APPEAL	
020: Make an appeal or request, not specified below	
021: Appeal for material cooperation, not specified below	
0211: Appeal for economic cooperation	03: EXPRESS INTENT TO COOPERATE
0212: Appeal for military cooperation	030: Express intent to cooperate, not specified below
0213: Appeal for judicial cooperation	031: Express intent to engage in material cooperation, not specified below
0214: Appeal for intelligence	0311: Express intent to cooperate economically
022: Appeal for diplomatic cooperation (such as policy support)	0312: Express intent to cooperate militarily
023: Appeal for aid, not specified below	0313: Express intent to cooperate on judicial matters
0231: Appeal for economic aid	0314: Express intent to cooperate on intelligence
0232: Appeal for military aid	032: Express intent to engage in diplomatic cooperation (such as policy support)
0233: Appeal for humanitarian aid	033: Express intent to provide material aid, not specified below
0234: Appeal for military protection or peacekeeping	0331: Express intent to provide economic aid
024: Appeal for political reform, not specified below	0332: Express intent to provide military
0241: Appeal for change in leadership	
0242: Appeal for policy change	
0243: Appeal for rights	
0244: Appeal for change in institutions, regime	



## D. GDELT begivenhedstyper

---

aid	ERATION
0333: Express intent to provide humanitarian aid	050: Engage in diplomatic cooperation, not specified below
0334: Express intent to provide military protection or peacekeeping	051: Praise or endorse
034: Express intent to institute political reform, not specified below	052: Defend verbally
0341: Express intent to change leadership	053: Rally support on behalf of
0342: Express intent to change policy	054: Grant diplomatic recognition
0343: Express intent to provide rights	055: Apologize
0344: Express intent to change institutions, regime	056: Forgive
035: Express intent to yield, not specified below	057: Sign formal agreement
0351: Express intent to ease administrative sanctions	06: ENGAGE IN MATERIAL COOPERATION
0352: Express intent to ease popular dissent	060: Engage in material cooperation, not specified below
0353: Express intent to release persons or property	061: Cooperate economically
0354: Express intent to ease economic sanctions, boycott, or embargo	062: Cooperate militarily
0355: Express intent to allow international involvement (non-mediation)	063: Engage in judicial cooperation
0356: Express intent to de-escalate military engagement	064: Share intelligence or information
036: Express intent to meet or negotiate	07: PROVIDE AID
037: Express intent to settle dispute	070: Provide aid, not specified below
038: Express intent to accept mediation	071: Provide economic aid
039: Express intent to mediate	072: Provide military aid
04: CONSULT	073: Provide humanitarian aid
040: Consult, not specified below	074: Provide military protection or peacekeeping
041: Discuss by telephone	075: Grant asylum
042: Make a visit	08: YIELD
043: Host a visit	080: Yield, not specified below
044: Meet at a "third" location	081: Ease administrative sanctions, not specified below
045: Mediate	0811: Ease restrictions on political freedoms
046: Engage in negotiation	0812: Ease ban on political parties or politicians
05: ENGAGE IN DIPLOMATIC COOP-	0813: Ease curfew
	0814: Ease state of emergency or martial law
	082: Ease political dissent

---

083: Accede to requests or demands for political reform, not specified below	as policy support)
0831: Accede to demands for change in leadership	103: Demand material aid, not specified below
0832: Accede to demands for change in policy	1031: Demand economic aid
0833: Accede to demands for rights	1032: Demand military aid
0834: Accede to demands for change in institutions, regime	1033: Demand humanitarian aid
084: Return, release, not specified below	1034: Demand military protection or peacekeeping
0841: Return, release person(s)	104: Demand political reform, not specified below
0842: Return, release property	1041: Demand change in leadership
085: Ease economic sanctions, boycott, embargo	1042: Demand policy change
086: Allow international involvement, not specified below	1043: Demand rights
0861: Receive deployment of peacekeepers	1044: Demand change in institutions, regime
0862: Receive inspectors	105: Demand that target yields, not specified below
0863: Allow humanitarian access	1051: Demand easing of administrative sanctions
087: De-escalate military engagement	1052: Demand easing of political dissent
0871: Declare truce, ceasefire	1053: Demand release of persons or property
0872: Ease military blockade	1054: Demand easing of economic sanctions, boycott, or embargo
0873: Demobilize armed forces	1055: Demand that target allows international involvement (non-mediation)
0874: Retreat or surrender militarily	1056: Demand de-escalation of military engagement
09: INVESTIGATE	106: Demand meeting, negotiation
090: Investigate, not specified below	107: Demand settling of dispute
091: Investigate crime, corruption	108: Demand mediation
092: Investigate human rights abuses	
093: Investigate military action	
094: Investigate war crimes	
10: DEMAND	11: DISAPPROVE
100: Demand, not specified below	110: Disapprove, not specified below
101: Demand material cooperation, not specified below	111: Criticize or denounce
1011: Demand economic cooperation	112: Accuse, not specified below
1012: Demand military cooperation	1121: Accuse of crime, corruption
1013: Demand judicial cooperation	1122: Accuse of human rights abuses
1014: Demand intelligence cooperation	1123: Accuse of aggression
102: Demand diplomatic cooperation (such	1124: Accuse of war crimes
	1125: Accuse of espionage, treason

113: Rally opposition against	pute
114: Complain officially	128: Defy norms, law
115: Bring lawsuit against	129: Veto
116: Find guilty or liable (legally)	
12: REJECT	13: THREATEN
120: Reject, not specified below	130: Threaten, not specified below
121: Reject material cooperation	131: Threaten non-force, not specified below
1211: Reject economic cooperation	1311: Threaten to reduce or stop aid
1212: Reject military cooperation	1312: Threaten with sanctions, boycott, embargo
122: Reject request or demand for material aid, not specified below	1313: Threaten to reduce or break relations
1221: Reject request for economic aid	132: Threaten with administrative sanctions, not specified below
1222: Reject request for military aid	1321: Threaten with restrictions on political freedoms
1223: Reject request for humanitarian aid	1322: Threaten to ban political parties or politicians
1224: Reject request for military protection or peacekeeping	1323: Threaten to impose curfew
123: Reject request or demand for political reform, not specified below	1324: Threaten to impose state of emergency or martial law
1231: Reject request for change in leadership	133: Threaten with political dissent, protest
1232: Reject request for policy change	134: Threaten to halt negotiations
1233: Reject request for rights	135: Threaten to halt mediation
1234: Reject request for change in institutions, regime	136: Threaten to halt international involvement (non-mediation)
124: Refuse to yield, not specified below	137: Threaten with repression
1241: Refuse to ease administrative sanctions	138: Threaten with military force, not specified below
1242: Refuse to ease popular dissent	1381: Threaten blockade
1243: Refuse to release persons or property	1382: Threaten occupation
1244: Refuse to ease economic sanctions, boycott, or embargo	1383: Threaten unconventional violence
1245: Refuse to allow international involvement (non mediation)	1384: Threaten conventional attack
1246: Refuse to de-escalate military engagement	1385: Threaten attack with WMD
125: Reject proposal to meet, discuss, or negotiate	139: Give ultimatum
126: Reject mediation	14: PROTEST
127: Reject plan, agreement to settle dis-	140: Engage in political dissent, not specified below
	141: Demonstrate or rally, not specified below

---

1411: Demonstrate for leadership change	15: EXHIBIT FORCE POSTURE
1412: Demonstrate for policy change	150: Demonstrate military or police power, not specified below
1413: Demonstrate for rights	151: Increase police alert status
1414: Demonstrate for change in institutions, regime	152: Increase military alert status
142: Conduct hunger strike, not specified below	153: Mobilize or increase police power
1421: Conduct hunger strike for leadership change	154: Mobilize or increase armed forces
1422: Conduct hunger strike for policy change	155: Mobilize or increase cyber-forces
1423: Conduct hunger strike for rights	16: REDUCE RELATIONS
1424: Conduct hunger strike for change in institutions, regime	160: Reduce relations, not specified below
143: Conduct strike or boycott, not specified below	161: Reduce or break diplomatic relations
1431: Conduct strike or boycott for leadership change	162: Reduce or stop material aid, not specified below
1432: Conduct strike or boycott for policy change	1621: Reduce or stop economic assistance
1433: Conduct strike or boycott for rights	1622: Reduce or stop military assistance
1434: Conduct strike or boycott for change in institutions, regime	1623: Reduce or stop humanitarian assistance
144: Obstruct passage, block, not specified below	163: Impose embargo, boycott, or sanctions
1441: Obstruct passage to demand leadership change	164: Halt negotiations
1442: Obstruct passage to demand policy change	165: Halt mediation
1443: Obstruct passage to demand rights	166: Expel or withdraw, not specified below
1444: Obstruct passage to demand change in institutions, regime	1661: Expel or withdraw peacekeepers
145: Protest violently, riot, not specified below	1662: Expel or withdraw inspectors, observers
1451: Engage in violent protest for leadership change	1663: Expel or withdraw aid agencies
1452: Engage in violent protest for policy change	17: COERCE
1453: Engage in violent protest for rights	170: Coerce, not specified below
1454: Engage in violent protest for change in institutions, regime	171: Seize or damage property, not specified below
	1711: Confiscate property
	1712: Destroy property
	172: Impose administrative sanctions, not specified below
	1721: Impose restrictions on political freedoms
	1722: Ban political parties or politicians
	1723: Impose curfew

D. GDELT begivenhedstyper

---

- 1724: Impose state of emergency or martial law

173: Arrest, detain, or charge with legal action

174: Expel or deport individuals

175: Use tactics of violent repression

176: Attack cybernetically

18: ASSAULT

180: Use unconventional violence, not specified below

181: Abduct, hijack, or take hostage

182: Physically assault, not specified below

1821: Sexually assault

1822: Torture

1823: Kill by physical assault

183: Conduct suicide, car, or other non-military bombing, not specified below

1831: Carry out suicide bombing

1832: Carry out vehicular bombing

1833: Carry out roadside bombing

1834: Carry out location bombing

184: Use as human shield

185: Attempt to assassinate

186: Assassinate

19: FIGHT
- 190: Use conventional military force, not specified below

191: Impose blockade, restrict movement

192: Occupy territory

193: Fight with small arms and light weapons

194: Fight with artillery and tanks

195: Employ aerial weapons, not specified below

1951: Employ precision-guided aerial munitions

1952: Employ remotely

20: USE UNCONVENTIONAL MASS VIOLENCE

200: Use unconventional mass violence, not specified below

201: Engage in mass expulsion

202: Engage in mass killings

203: Engage in ethnic cleansing

204: Use weapons of mass destruction, not specified below

2041: Use chemical, biological, or radiological weapons

2042: Detonate nuclear weapons

## E. KODNING AF ANALYSE I R

### E.1. Analyse: Udbrud af borgerkrig

```
1 #-----
2 # Dannelselse af forudsigelsesmodeller for udbrud af borgerkrig
3 #
4 #load af data:
5 load(Data_cwstart.Rda)
6
7 #Oprettelse af traeningsdatasaet og testdatasaet
8 train_data <- data %>%
9   filter(year <= 2011)
10 test_data <- data %>%
11   filter(year >=2012)
12
13 # Fjernelse af unoeedvendig data og data, der ikke skal indgaa i analysen grundet dets "fremtidskarakter"
14 train_data <- train_data %>%
15   select(-cwy,-deathyear)
16 test_data <- test_data %>%
17   select(-cwy,-deathyear)
18
19 #-----
20 #                               Oprettelse af forudsigelsesmodeller
21 #-----
22
23 #-----Baseline model-----
24 model_baseline_cwstart <- glm(cwstart ~ . -country - year -month,family=binomial(link='logit'),data=train_data)
25
26 #forudsigelse
27 test_data$cwstart_pred_baseline = predict(model_baseline_cwstart, newdata = test_data, type="response")
28
29 #-----Random forest model-----
30 library("randomForest")
31 model_rf_cwstart <- randomForest(cwstart ~ . -country - year -month, data=train_data ,ntree=250, importance=TRUE)
32
33 #Predictions
34 test_data$cwstart_pred_rf = predict(model_rf_cwstart, newdata = test_data, type="response")
35
36 #-----extreme gradient boosted trees -----
37 library("caret")
38
39 fitControl <- trainControl(method = "cv", number = 5)
40
41 model_egbt_cwstart <- train(cwstart ~ . -country - year -month,
42   data = train_data,
43   method = "xgbTree",
44   trControl = fitControl,
45   search = "random",
46   objective = "binary:logistic")
47
48 #Predictions
49 test_data$cwstart_pred_egbt = predict(model_egbt_cwstart, newdata = test_data)
50
51
52 #-----
53 #                               Visualiseringer af forudsigelser
54 #-----
55
56 #-----Densitets histogrammer-----
57
58 #-----Baseline model-----
59 plot_pred_type_distribution <- function(df, threshold) {
60   v <- rep(NA, nrow(df))
61   v <- ifelse(df$cwstart_pred_baseline >= threshold & df$cwstart == 1, "TP", v)
62   v <- ifelse(df$cwstart_pred_baseline >= threshold & df$cwstart == 0, "FP", v)
63   v <- ifelse(df$cwstart_pred_baseline < threshold & df$cwstart == 1, "FN", v)
64   v <- ifelse(df$cwstart_pred_baseline < threshold & df$cwstart == 0, "TN", v)
65 }
66 # Tilfoerelse af FP,FN,SP,SN til datasaet
67 test_data$pred_type <- plot_pred_type_distribution(test_data, 0.1)
68
69 #Plotting af densitet histogram med jitter og linje
70 densitet_baseline_cwstart <- test_data %>%
71   ggplot(aes(as.factor(cwstart),cwstart_pred_baseline)) +
72   geom_violin(mapping = NULL, data = test_data, stat = "ydensity",
73     position = "dodge", draw_quantiles = NULL, trim = TRUE,
```

## E. Kodning af analyse i R

```
74     scale = "area", na.rm = TRUE, show.legend = TRUE,
75     inherit.aes = TRUE)+
76   geom_jitter(aes(color=pred_type),data= test_data, alpha=0.2, width = 0.4)+
77   geom_hline(yintercept=0.1, color="red", alpha=0.6)+
78   labs(y = "Sandsynlighed", x="", colour = "Praediktionstype") +
79   scale_x_discrete(breaks=c("0","1"),
80     labels=c("ikke-borgerkrig \n y = 0", "Borgerkrig \n y = 1")) +
81   guides(color = guide_legend(override.aes = list(size=1, alpha = 1))) +
82   ggtitle("Baseline model") +
83   theme(plot.title = element_text(hjust = 0.5, size =12, face = "plain"))
84
85
86 #----- Random forest model-----
87 plot_pred_type_distribution <- function(df, threshold) {
88   v <- rep(NA, nrow(df))
89   v <- ifelse(df$cwstart_pred_rf >= threshold & df$cwstart == 1, "TP", v)
90   v <- ifelse(df$cwstart_pred_rf >= threshold & df$cwstart == 0, "FP", v)
91   v <- ifelse(df$cwstart_pred_rf < threshold & df$cwstart == 1, "FN", v)
92   v <- ifelse(df$cwstart_pred_rf < threshold & df$cwstart == 0, "TN", v)
93 }
94 # Tilfoerelse af FP,FN,SP,SN til datasaet
95 test_data$pred_type <- plot_pred_type_distribution(test_data, 0.1)
96
97
98 #Plotting af densitet histogram med jitter og linje
99 densitet_rf_cwstart <- test_data %>%
100   ggplot(aes(as.factor(cwstart),cwstart_pred_rf)) +
101   geom_violin(mapping = NULL, data = test_data, stat = "ydensity",
102     position = "dodge", draw_quantiles = NULL, trim = TRUE,
103     scale = "area", na.rm = TRUE, show.legend = TRUE,
104     inherit.aes = TRUE)+
105   geom_jitter(aes(color=pred_type),data= test_data, alpha=0.2, width = 0.4)+
106   geom_hline(yintercept=0.1, color="red", alpha=0.6)+
107   labs(y = "Sandsynlighed", x="", colour = "Praediktionstype") +
108   scale_x_discrete(breaks=c("0","1"),
109     labels=c("ikke-borgerkrig \n y = 0", "Borgerkrig \n y = 1")) +
110   guides(color = guide_legend(override.aes = list(size=1, alpha = 1, default.unit = ""))) +
111   ggtitle("Random forest model") +
112   theme(plot.title = element_text(hjust = 0.5, size =12, face = "plain"))
113
114 #-----extreme gradient boosting model-----
115 plot_pred_type_distribution <- function(df, threshold) {
116   v <- rep(NA, nrow(df))
117   v <- ifelse(df$cwstart_pred_egbt >= threshold & df$cwstart == 1, "TP", v)
118   v <- ifelse(df$cwstart_pred_egbt >= threshold & df$cwstart == 0, "FP", v)
119   v <- ifelse(df$cwstart_pred_egbt < threshold & df$cwstart == 1, "FN", v)
120   v <- ifelse(df$cwstart_pred_egbt < threshold & df$cwstart == 0, "TN", v)
121 }
122 # Tilfoerelse af FP,FN,SP,SN til datasaet
123 test_data$pred_type <- plot_pred_type_distribution(test_data, 0.05)
124
125
126 #Plotting af densitet histogram med jitter og linje
127 densitet_egbt_cwstart <- test_data %>%
128   ggplot(aes(as.factor(cwstart),cwstart_pred_egbt)) +
129   geom_violin(mapping = NULL, data = test_data, stat = "ydensity",
130     position = "dodge", draw_quantiles = NULL, trim = TRUE,
131     scale = "area", na.rm = TRUE, show.legend = TRUE,
132     inherit.aes = TRUE)+
133   geom_jitter(aes(color=pred_type),data= test_data, alpha=0.2, width = 0.4)+
134   geom_hline(yintercept=0.05, color="red", alpha=0.6)+
135   labs(y = "Sandsynlighed", x="", colour = "Praediktionstype") +
136   scale_x_discrete(breaks=c("0","1"),
137     labels=c("ikke-borgerkrig \n y = 0", "Borgerkrig \n y = 1")) +
138   guides(color = guide_legend(override.aes = list(size=1, alpha = 1))) +
139   ggtitle("EXGT model") +
140   theme(plot.title = element_text(hjust = 0.5, size =12, face = "plain"))
141
142
143 #-----Samling af densitetsfigurer til 1 figur-----
144 setwd("Latexfigure")
145 pdf("density_cwstart.pdf")
146 grid.arrange(densitet_baseline_cwstart, densitet_rf_cwstart,densitet_egbt_cwstart)
147 dev.off()
148
149 #-----AREA UNDER CURVE-----
150 library("pROC")
151
152 ROC_logit <- roc(test_data$cwstart, test_data$cwstart_pred_baseline)
```

```

153 ROC_rf <- roc(test_data$cwstart, test_data$cwstart_pred_rf)
154 ROC_egbt <- roc(test_data$cwstart, test_data$cwstart_pred_egbt)
155
156 auc(ROC_logit)
157 auc(ROC_rf)
158 auc(ROC_egbt)
159
160 #-----
161 # Advanced ROC
162 #-----
163 #Oprettelse af genereisk plot function:
164
165 plot_roc <- function(roc, threshold, cost_of_fp, cost_of_fn) {
166   norm_vec <- function(v) (v - min(v))/diff(range(v))
167   idx_threshold = which.min(abs(roc$threshold-threshold))
168   col_ramp <- colorRampPalette(c("green","orange","red","black"))(100)
169   col_by_cost <- col_ramp[ceiling(norm_vec(roc$cost)*99)+1]
170   p_roc <- ggplot(roc, aes(fpr,tpr)) +
171     geom_line(color=rgb(0,0,1,alpha=0.6)) +
172     geom_point(color=col_by_cost, size=1, alpha=1) +
173     coord_fixed() +
174     geom_line(aes(threshold,threshold), color=rgb(0,0,1,alpha=0.5)) +
175     xlab("FPR") +
176     ylab("SPR") +
177     theme(axis.text.x=element_text(size=rel(0.8), angle=90))+
178     theme(axis.text.y=element_text(size=rel(0.8)))+
179     theme(plot.margin = unit(c(0, 0, 0, 0), "cm"))+
180     geom_hline(yintercept=roc[idx_threshold,"tpr"], alpha=0.5, linetype="dashed") +
181     geom_vline(xintercept=roc[idx_threshold,"fpr"], alpha=0.5, linetype="dashed")+
182     theme(plot.title = element_text(size=12, hjust = 0.5))
183   p_roc
184 }
185
186 #-----Udregning af ROC vaerdi for baselinemodel-----
187 calculate_roc <- function(df, cost_of_fp, cost_of_fn, n=100) {
188   tpr <- function(df, threshold) {
189     sum(test_data$cwstart_pred_baseline >= threshold & test_data$cwstart == 1) / sum(test_data$cwstart == 1)
190   }
191
192   fpr <- function(df, threshold) {
193     sum(test_data$cwstart_pred_baseline >= threshold & test_data$cwstart == 0) / sum(test_data$cwstart == 0)
194   }
195
196   cost <- function(df, threshold, cost_of_fp, cost_of_fn) {
197     sum(test_data$cwstart_pred_baseline >= threshold & test_data$cwstart == 0) * cost_of_fp +
198     sum(test_data$cwstart_pred_baseline < threshold & test_data$cwstart == 1) * cost_of_fn
199   }
200
201   roc <- data.frame(threshold = seq(0,1,length.out=n), tpr=NA, fpr=NA)
202   roc$tpr <- sapply(roc$threshold, function(th) tpr(df, th))
203   roc$fpr <- sapply(roc$threshold, function(th) fpr(df, th))
204   roc$cost <- sapply(roc$threshold, function(th) cost(df, th, cost_of_fp, cost_of_fn))
205   return(roc)
206 }
207
208 #Udregning og plot
209 roc <- calculate_roc(test_data, 1, 4, n = 1000)
210 roc_baseline_cvm <- plot_roc(roc, 0.1, 1,4)
211
212 #-----Udregning af ROC vaerdi for random forest model-----
213
214
215 calculate_roc <- function(df, cost_of_fp, cost_of_fn, n=100) {
216   tpr <- function(df, threshold) {
217     sum(test_data$cwstart_pred_rf >= threshold & test_data$cwstart == 1) / sum(test_data$cwstart == 1)
218   }
219
220   fpr <- function(df, threshold) {
221     sum(test_data$cwstart_pred_rf >= threshold & test_data$cwstart == 0) / sum(test_data$cwstart == 0)
222   }
223
224   cost <- function(df, threshold, cost_of_fp, cost_of_fn) {
225     sum(test_data$cwstart_pred_rf >= threshold & test_data$cwstart == 0) * cost_of_fp +
226     sum(test_data$cwstart_pred_rf < threshold & test_data$cwstart == 1) * cost_of_fn
227   }
228
229   roc <- data.frame(threshold = seq(0,1,length.out=n), tpr=NA, fpr=NA)
230   roc$tpr <- sapply(roc$threshold, function(th) tpr(df, th))
231   roc$fpr <- sapply(roc$threshold, function(th) fpr(df, th))

```



## E. Kodning af analyse i R

```
232 roc$cost <- sapply(roc$threshold, function(th) cost(df, th, cost_of_fp, cost_of_fn))
233 return(roc)
234 }
235
236 #Udregning og plot
237 roc <- calculate_roc(test_data, 1, 4, n = 1000)
238 roc_rf_cwstart <- plot_roc(roc, 0.1, 1,4)
239
240
241
242 #-----Udregning af ROC vaerdi for egbt model-----
243
244 calculate_roc <- function(df, cost_of_fp, cost_of_fn, n=100) {
245   tpr <- function(df, threshold) {
246     sum(test_data$cwstart_pred_egbt >= threshold & test_data$cwstart == 1) / sum(test_data$cwstart == 1)
247   }
248
249   fpr <- function(df, threshold) {
250     sum(test_data$cwstart_pred_egbt >= threshold & test_data$cwstart == 0) / sum(test_data$cwstart == 0)
251   }
252
253   cost <- function(df, threshold, cost_of_fp, cost_of_fn) {
254     sum(test_data$cwstart_pred_egbt >= threshold & test_data$cwstart == 0) * cost_of_fp +
255     sum(test_data$cwstart_pred_egbt < threshold & test_data$cwstart == 1) * cost_of_fn
256   }
257
258   roc <- data.frame(threshold = seq(0,1,length.out=n), tpr=NA, fpr=NA)
259   roc$tpr <- sapply(roc$threshold, function(th) tpr(df, th))
260   roc$fpr <- sapply(roc$threshold, function(th) fpr(df, th))
261   roc$cost <- sapply(roc$threshold, function(th) cost(df, th, cost_of_fp, cost_of_fn))
262   return(roc)
263 }
264
265 #Udregning og plot
266 roc <- calculate_roc(test_data, 1, 4, n = 1000)
267 roc_egbt_cwstart <- plot_roc(roc, 0.1, 1,4)
268
269 #----- Samling af ROC plot -----
270
271 setwd(Latexfigure)
272 pdf("roc_cwstart.pdf")
273 grid.arrange(roc_baseline_cvm, roc_rf_cwstart, roc_egbt_cwstart, ncol=3, nrow=1)
274 dev.off()
```

### E.2. Analyse: Borgerkrigslignende tilstande

```
1
2 #-----
3 # Dannelse af forudsigelsesmodeller for borgerkrigslignende tilstande
4
5 #load af data:
6 load(Data_cwstart.Rda)
7
8 #Oprettelse af traeningsdatasaet og testdatasaet - Bloksplit
9 train_data <- data %>%
10   filter(year <= 2011)
11 test_data <- data %>%
12   filter(year >=2012)
13
14 # Fjernelse af unoedvendig data og data, der ikke skal indgaa i analysen grundet dets "fremtidskarakter"
15 train_data <- train_data %>%
16   select(-cwy,-deathyear)
17 test_data <- test_data %>%
18   select(-cwy,-deathyear)
19
20 #-----
21 # Oprettelse af forudsigelsesmodeller
22
23 #-----Baseline model-----
24 model_baseline_cwm <- glm(cwm ~ . -country - year -month,family=binomial(link='logit'),data=train_data)
25
26 #forudsigelse
27 test_data$cwm_pred_baseline = predict(model_baseline_cwm, newdata = test_data, type="response")
```

```

28
29
30 #-----Random forest model-----
31 library("randomForest")
32 model_rf_cwm <- randomForest(cwm ~ . -country - year -month, data=train_data, ntree=250, importance=TRUE)
33
34 #Predictions
35 test_data$cwm_pred_rf = predict(model_rf_cwm, newdata = test_data, type="prob")[,2]
36
37
38 #-----extreme gradient boosted trees -----
39 library("caret")
40
41 fitControl <- trainControl(method = "cv", number = 5)
42
43 model_egbt_cwm <- train(cwm ~ . -country - year -month,
44                        data = train_data,
45                        method = "xgbTree",
46                        trControl = fitControl,
47                        search = "random",
48                        objective = "binary:logistic")
49
50 #Predictions
51 test_data$cwm_pred_egbt = predict(model_egbt_cwm, newdata = test_data, type="prob")[,2]
52
53 #-----
54 # Visualisering af forudsigelser
55
56 #-----
57 # Optimaltaerskelvaerdier for delta findes gennem en analyse af ROC elementerne
58 #
59 breaks <- c(seq(0, 0.00001, by = 0.000000001),
60             seq(0.00001, 0.0001, by = 0.00000001),
61             seq(0.0001, 0.001, by = 0.000001),
62             seq(0.001, 0.01, by = 0.00001),
63             seq(0.01, 0.1, by = 0.0001),
64             seq(0.1, 1, by = 0.001))
65 # bruges til at oprette hvor mange varitioner for af delta der skal udregnes
66
67 #Oprettelse af genereisk ROC-plot function:
68
69
70 #Oprettelse af genereisk plot function:
71 plot_roc <- function(roc, threshold, cost_of_fp, cost_of_fn) {
72   norm_vec <- function(v) (v - min(v))/diff(range(v))
73   idx_threshold = which.min(abs(roc$threshold-threshold))
74   col_ramp <- colorRampPalette(c("green", "orange", "red", "black"))(100)
75   col_by_cost <- col_ramp[ceiling(norm_vec(roc$cost)*99)+1]
76   p_roc <- ggplot(roc, aes(fpr, tpr)) +
77     geom_line(color=rgb(0,0,1,alpha=0.6)) +
78     geom_point(color=col_by_cost, size=1, alpha=1) +
79     coord_fixed() +
80     geom_segment(aes(x = 0, xend = 1, y = 0, yend = 1), color="blue", size = 0.5) +
81     xlab("FPR") +
82     ylab("SPR") +
83     xlim(0,1) +
84     ylim(0,1) +
85     theme(axis.text.x=element_text(size=rel(0.8), angle=90))+
86     theme(axis.text.y=element_text(size=rel(0.8)))+
87     theme(plot.margin = unit(c(0, 0, 0, 0), "cm"))+
88     geom_hline(yintercept=roc[idx_threshold, "tpr"], alpha=0.5, linetype="dashed") +
89     geom_vline(xintercept=roc[idx_threshold, "fpr"], alpha=0.5, linetype="dashed")+
90     theme(plot.title = element_text(size=12, hjust = 0.5))
91   p_roc
92 }
93
94
95 #-----
96 #Udregning af ROC vaerdi for baselinemodel
97 calculate_roc <- function(df, cost_of_fp, cost_of_fn) {
98   tpr <- function(df, threshold) {
99     sum(test_data$cwm_pred_baseline >= threshold & test_data$cwm == 1) / sum(test_data$cwm == 1)
100   }
101
102   fpr <- function(df, threshold) {
103     sum(test_data$cwm_pred_baseline >= threshold & test_data$cwm == 0) / sum(test_data$cwm == 0)
104   }
105
106   cost <- function(df, threshold, cost_of_fp, cost_of_fn) {

```

## E. Kodning af analyse i R

```
107   sum(test_data$cwm_pred_baseline >= threshold & test_data$cwm == 0) * 1 +
108   sum(test_data$cwm_pred_baseline < threshold & test_data$cwm == 1) * 10
109 }
110
111 roc <- data.frame(threshold = breaks, tpr=NA, fpr=NA)
112 roc$tpr <- sapply(roc$threshold, function(th) tpr(df, th))
113 roc$fpr <- sapply(roc$threshold, function(th) fpr(df, th))
114 roc$cost <- sapply(roc$threshold, function(th) cost(df, th, 1, 10))
115
116 return(roc)
117 }
118
119 #Udregning og plot
120 roc <- calculate_roc(test_data, 1, 10)
121 roc_baseline_cwm <- plot_roc(roc, 0.337, 1,10)
122
123
124 #-----
125 #Udregning af ROC vaerdi for Random Forest
126 calculate_roc <- function(df, cost_of_fp, cost_of_fn) {
127   tpr <- function(df, threshold) {
128     sum(test_data$cwm_pred_rf >= threshold & test_data$cwm == 1) / sum(test_data$cwm == 1)
129   }
130
131   fpr <- function(df, threshold) {
132     sum(test_data$cwm_pred_rf >= threshold & test_data$cwm == 0) / sum(test_data$cwm == 0)
133   }
134
135   cost <- function(df, threshold, cost_of_fp, cost_of_fn) {
136     sum(test_data$cwm_pred_rf >= threshold & test_data$cwm == 0) * 1 +
137     sum(test_data$cwm_pred_rf < threshold & test_data$cwm == 1) * 10
138   }
139
140   roc <- data.frame(threshold = breaks, tpr=NA, fpr=NA)
141   roc$tpr <- sapply(roc$threshold, function(th) tpr(df, th))
142   roc$fpr <- sapply(roc$threshold, function(th) fpr(df, th))
143   roc$cost <- sapply(roc$threshold, function(th) cost(df, th, 1, 10))
144
145   return(roc)
146 }
147
148 #Udregning og plot
149 roc <- calculate_roc(test_data, 1, 10)
150 roc_rf_cwm <- plot_roc(roc, 0.1290, 1,10)
151
152 #-----
153 #Udregning af ROC vaerdi for EGBtree
154 calculate_roc <- function(df, cost_of_fp, cost_of_fn) {
155   tpr <- function(df, threshold) {
156     sum(test_data$cwm_pred_egbt >= threshold & test_data$cwm == 1) / sum(test_data$cwm == 1)
157   }
158
159   fpr <- function(df, threshold) {
160     sum(test_data$cwm_pred_egbt >= threshold & test_data$cwm == 0) / sum(test_data$cwm == 0)
161   }
162
163   cost <- function(df, threshold, cost_of_fp, cost_of_fn) {
164     sum(test_data$cwm_pred_egbt >= threshold & test_data$cwm == 0) * 1 +
165     sum(test_data$cwm_pred_egbt < threshold & test_data$cwm == 1) * 10
166   }
167
168   roc <- data.frame(threshold = breaks, tpr=NA, fpr=NA)
169   roc$tpr <- sapply(roc$threshold, function(th) tpr(df, th))
170   roc$fpr <- sapply(roc$threshold, function(th) fpr(df, th))
171   roc$cost <- sapply(roc$threshold, function(th) cost(df, th, 1, 10))
172
173   return(roc)
174 }
175
176 #Udregning og plot
177 roc <- calculate_roc(test_data, 1, 10)
178 roc_egbt_cwm <- plot_roc(roc, 0.0396, 1,10)
179
180
181 #Samling af ROC-kurver
182 setwd(Latexfigure)
183 pdf("roc_cwm.pdf")
184 grid.arrange(roc_baseline_cwm, roc_rf_cwm, roc_egbt_cwm, ncol=3, nrow=1)
185 dev.off()
```

```

186
187 #-----Densitets histogrammer for de 3 forudsigelser
188
189 #-- Baseline model
190 plot_pred_type_distribution <- function(df, threshold) {
191   v <- rep(NA, nrow(df))
192   v <- ifelse(df$cwm_pred_baseline >= threshold & df$cwm == 1, "TP", v)
193   v <- ifelse(df$cwm_pred_baseline >= threshold & df$cwm == 0, "FP", v)
194   v <- ifelse(df$cwm_pred_baseline < threshold & df$cwm == 1, "FN", v)
195   v <- ifelse(df$cwm_pred_baseline < threshold & df$cwm == 0, "TN", v)
196 }
197 # Tilfoerelse af FP,FN,SP,SN til datasæet
198 test_data$pred_type <- plot_pred_type_distribution(test_data, 0.337)
199
200 #Plotting af densitet histogram med jitter og linje
201 densitet_baseline_cwm <- test_data %>%
202   ggplot(aes(as.factor(cwm),cwm_pred_baseline)) +
203   geom_violin(mapping = NULL, data = test_data, stat = "ydensity",
204     position = "dodge", draw_quantiles = NULL, trim = TRUE,
205     scale = "width", na.rm = TRUE, show.legend = TRUE,
206     inherit.aes = TRUE)+
207   geom_jitter(aes(color=pred_type),data= test_data, alpha=0.2, width = 0.4)+
208   geom_hline(yintercept=0.337, color="red", alpha=0.6)+
209   labs(y = "Sandsynlighed", x="", colour = "Praediktionstype") +
210   scale_x_discrete(breaks=c("0", "1"),
211     labels=c("ikke-borgerkrig \n y = 0", "Borgerkrig \n y = 1")) +
212   guides(color = guide_legend(override.aes = list(size=1, alpha = 1))) +
213   ggtitle("Baseline model") +
214   theme(plot.title = element_text(hjust = 0.5, size =12, face = "plain"))
215
216
217 # -----
218 #-- Random forest model
219 plot_pred_type_distribution <- function(df, threshold) {
220   v <- rep(NA, nrow(df))
221   v <- ifelse(df$cwm_pred_rf >= threshold & df$cwm == 1, "TP", v)
222   v <- ifelse(df$cwm_pred_rf >= threshold & df$cwm == 0, "FP", v)
223   v <- ifelse(df$cwm_pred_rf < threshold & df$cwm == 1, "FN", v)
224   v <- ifelse(df$cwm_pred_rf < threshold & df$cwm == 0, "TN", v)
225 }
226 # Tilfoerelse af FP,FN,SP,SN til datasæet
227 test_data$pred_type <- plot_pred_type_distribution(test_data, 0.1290)
228
229 #Plotting af densitet histogram med jitter og linje
230 densitet_rf_cwm <- test_data %>%
231   ggplot(aes(as.factor(cwm),cwm_pred_rf)) +
232   geom_violin(mapping = NULL, data = test_data, stat = "ydensity",
233     position = "dodge", draw_quantiles = NULL, trim = TRUE,
234     scale = "width", na.rm = TRUE, show.legend = TRUE,
235     inherit.aes = TRUE)+
236   geom_jitter(aes(color=pred_type),data= test_data, alpha=0.2, width = 0.4)+
237   geom_hline(yintercept=0.1290, color="red", alpha=0.6)+
238   labs(y = "Sandsynlighed", x="", colour = "Praediktionstype") +
239   scale_x_discrete(breaks=c("0", "1"),
240     labels=c("ikke-borgerkrig \n y = 0", "Borgerkrig \n y = 1")) +
241   guides(color = guide_legend(override.aes = list(size=1, alpha = 1, default.unit = ""))) +
242   ggtitle("Random forest model") +
243   theme(plot.title = element_text(hjust = 0.5, size =12, face = "plain"))
244
245 # -----
246 #-- extreme gradient boosting model
247 plot_pred_type_distribution <- function(df, threshold) {
248   v <- rep(NA, nrow(df))
249   v <- ifelse(df$cwm_pred_egbt >= threshold & df$cwm == 1, "TP", v)
250   v <- ifelse(df$cwm_pred_egbt >= threshold & df$cwm == 0, "FP", v)
251   v <- ifelse(df$cwm_pred_egbt < threshold & df$cwm == 1, "FN", v)
252   v <- ifelse(df$cwm_pred_egbt < threshold & df$cwm == 0, "TN", v)
253 }
254 # Tilfoerelse af FP,FN,SP,SN til datasæet
255 test_data$pred_type <- plot_pred_type_distribution(test_data, 0.0396)
256
257 #Plotting af densitet histogram med jitter og linje
258 densitet_egbt_cwm <- test_data %>%
259   ggplot(aes(as.factor(cwm),cwm_pred_egbt)) +
260   geom_violin(mapping = NULL, data = test_data, stat = "ydensity",
261     position = "dodge", draw_quantiles = NULL, trim = TRUE,
262     scale = "width", na.rm = TRUE, show.legend = TRUE,
263     inherit.aes = TRUE)+
264   geom_jitter(aes(color=pred_type),data= test_data, alpha=0.2, width = 0.4)+

```

## E. Kodning af analyse i R

```
265 geom_hline(yintercept=0.0396, color="red", alpha=0.6)+
266 labs( y = "Sandsynlighed", x="", colour ="Praediktionstype") +
267 scale_x_discrete(breaks=c("0","1"),
268                 labels=c("ikke-borgerkrig \n y = 0", "Borgerkrig \n y = 1")) +
269 guides(color = guide_legend(override.aes = list(size=1, alpha = 1))) +
270 ggtitle("EXGT model") +
271 theme(plot.title = element_text(hjust = 0.5, size =12, face = "plain"))
272
273 #Samling af figurer til 1 figur
274 setwd(Latexfigure)
275 png("density_cwm.png")
276 grid.arrange(densitet_baseline_cwm, densitet_rf_cwm,densitet_egbt_cwm)
277 dev.off()
278
279 # -----
280 # Area under curve
281 library("pROC")
282 ROC_logit <- roc(test_data$cwm, test_data$cwm_pred_baseline)
283 ROC_rf <- roc(test_data$cwm, test_data$cwm_pred_rf)
284 ROC_egbt <- roc(test_data$cwm, test_data$cwm_pred_egbt)
285 auc(ROC_logit)
286 auc(ROC_rf)
287 auc(ROC_egbt)
```

### E.3. Analyse: Intensiteten af borgerkrig

#### E.3.1. Antal Døde

```
1 set.seed(42)
2 # -----
3 # Dannelse af forudsigelsesmodeller for udbrud af borgerkrig
4 #load af data:
5 load(Data_death.Rda)
6
7 #Oprettelse af traeningsdatasaet og testdatasaet - Bloksplit
8 train_data <- data %>%
9   filter(year <= 2011)
10 test_data <- data %>%
11   filter(year >=2012)
12
13 # Fjernelse af unoedvendig data og data, der ikke skal indgaa i analysen grundet dets "fremtidskarakter"
14 train_data <- train_data %>%
15   select(-cwy,-deathyear)
16 test_data <- test_data %>%
17   select(-cwy,-deathyear)
18
19
20 # -----
21 # Oprettelse af forudsigelsesmodeller
22
23 #-----Baseline model-----
24 model_baseline_deaths <- lm(formula = deaths ~.-country -year -month, data=train_data)
25
26 #Forudsigelse
27 test_data$deaths_pred_baseline = predict(model_baseline_deaths, newdata = test_data)
28 test_data$pred_res_baseline = test_data$deaths - test_data$deaths_pred_baseline
29
30 library(Metrics)
31 rmse_baseline <- rmse(test_data$deaths, test_data$deaths_pred_baseline) # 2084
32 Rsquared_baseline_death = cor(test_data$deaths, test_data$deaths_pred_baseline)^2 #0,0023
33
34
35 #-----Random forest model-----
36 library("randomForest")
37 model_rf_deaths <- randomForest(deaths ~ . -country - year -month, data=train_data ,ntree=250)
38
39 #Predictions
40 test_data$deaths_pred_rf = predict(model_rf_deaths, newdata = test_data)
41 test_data$pred_res_rd_rf = test_data$deaths - test_data$deaths_pred_rf
42
43 rmse_rf <- rmse(test_data$deaths, test_data$deaths_pred_rf) #529
44 Rsquared_rf_death = cor(test_data$deaths, test_data$deaths_pred_rf)^2 #0,14
```

```

45 #-----extreme gradient boosted trees -----
46 library("caret")
47
48 fitControl <- trainControl(method = "cv", number = 5)
49
50 model_egbt_deaths <- train(deaths ~ . -country - year -month,
51                           data = train_data,
52                           method = "xgbTree",
53                           trControl = fitControl,
54                           search = "random")
55
56 #Predictions
57 test_data$deaths_pred_egbt = predict(model_egbt_deaths, newdata = test_data)
58 test_data$pred_res_egbt = test_data$deaths - test_data$deaths_pred_egbt
59
60 rmse_egbt <- rmse(test_data$deaths, test_data$deaths_pred_egbt) #105
61 Rsquared_egbt_death = cor(test_data$deaths, test_data$deaths_pred_egbt)^2 #0,665
62
63 #
64 # Visualiseringer af forudsigelser
65
66 deathpointplot <- ggplot(data=test_data) +
67   geom_point(aes(x = deaths, y = deaths_pred_baseline, color = "green"), size = 0.25, alpha = 0.2) +
68   geom_point(aes(x = deaths, y = deaths_pred_rf, color = "blue"), size = 0.25, alpha=0.2) +
69   geom_point(aes(x = deaths, y = deaths_pred_egbt, color = "red"), size= 0.25, alpha=0.2) +
70   geom_abline(slope = 1, intercept = 0) +
71   scale_color_manual(name = "",
72                     values=c("green"="green", "blue"="blue", "red" = "red"),
73                     labels = c("green" = "Baseline", "blue"="Random Forest", "red" = "EGBtree")) +
74   ylab("Forudsagte antal doede") +
75   xlab("Antal doede")+
76   ylim(-100,1500) +
77   xlim(0,1500) +
78   theme(legend.position = "",
79         legend.title=element_blank(),
80         legend.text = element_text(size = 7),
81         legend.background = element_rect(fill=NA)) +
82   guides(color = guide_legend(override.aes = list(size=2, alpha = 1)))
83
84 kerneldensity <- ggplot(data=test_data) +
85   geom_density(aes(x= pred_res_baseline), bw= 12, color = alpha("red", 0.2), size = 0.3, fill= "red", alpha = 0.5 )+
86   geom_density(aes(x= pred_res_rd_rf), bw= 12, color = alpha("blue",0.2), size = 0.3, fill= "blue", alpha = 0.5)+
87   geom_density(aes(x= pred_res_egbt), bw= 12, color = alpha("green", 0.2), size = 0.3, fill = "green", alpha = 0.5)+
88   stat_density(aes(x=pred_res_baseline, colour="red"), geom="line", bw= 12,position="identity", size = 0, alpha=0.1) +
89   stat_density(aes(x=pred_res_rd_rf, colour="blue"), geom="line", bw= 12,position="identity", size = 0, alpha=0.1) +
90   stat_density(aes(x=pred_res_egbt, colour="green"), geom="line", bw= 12,position="identity", size = 0, alpha=0.1) +
91   scale_color_manual(name = "",
92                     values=c("red"="red", "blue"="blue", "green" = "green"),
93                     labels = c("red" = "Baseline", "green"="Random Forest", "blue" = "EGBtree")) +
94   ylab("Densitet") +
95   xlab("Fejl i praediktion") +
96   xlim(-250,250) +
97   theme(legend.position = "bottom",
98         legend.title=element_blank(),
99         legend.text = element_text(size = 7),
100        legend.background = element_rect(fill=NA)) +
101   guides(color = guide_legend(override.aes = list(size=2, alpha = 1)))
102
103 setwd(Latexfigure)
104 png("deathsplot.png")
105 grid.arrange(deathpointplot, kerneldensity)
106 dev.off()

```

### E.3.2. Antal kamphandleringer

```

1 #
2 # Dannelse af forudsigelsesmodeller for udbrud af borgerkrig
3 #load af data:
4 load(Data_incidents.Rda)
5
6 #Oprettelse af traeningsdatasaet og testdatasaet - Bloksplit

```

## E. Kodning af analyse i R

```
7 train_data <- data %>%
8   filter(year <= 2011)
9 test_data <- data %>%
10   filter(year >=2012)
11
12 # Fjernelse af unødvendig data og data, der ikke skal indgaa i analysen grundet dets "fremtidskarakter"
13 train_data <- train_data %>%
14   select(-cwy,-deathyear)
15 test_data <- test_data %>%
16   select(-cwy,-deathyear)
17
18 #-----
19 # Oprettelse af forudsigelsesmodeller
20
21 #-----Baseline model-----
22 model_baseline_incidents <- lm(formula = incidents ~.-country -year -month, data=train_data)
23 summary(model_baseline_incidents)
24 setwd(Models)
25 save(model_baseline_incidents, file = "o_lm_incidents.rda")
26 #forudsigelse
27 test_data$incidents_pred_baseline = predict(model_baseline_incidents, newdata = test_data)
28 test_data$pred_res_baseline = test_data$incidents - test_data$incidents_pred_baseline
29
30 library(Metrics)
31 rmse_baseline <- rmse(test_data$incidents, test_data$incidents_pred_baseline) # 3,8
32 Rsquared_baseline = cor(test_data$incidents, test_data$incidents_pred_baseline)^2 #0,933
33
34 #-----Random forest model-----
35 library("randomForest")
36 model_rf_incidents <- randomForest(incidents ~ . -country - year -month, data=train_data, ntree=250, importance=TRUE)
37 model_rf_incidents
38 setwd(Models)
39 save(model_rf_incidents, file = "2_rf_incidents.rda")
40
41 #Predictions
42 test_data$incidents_pred_rf = predict(model_rf_incidents, newdata = test_data)
43 test_data$pred_res_rd_rf = test_data$incidents - test_data$incidents_pred_rf
44
45 rmse_rf <- rmse(test_data$incidents, test_data$incidents_pred_rf) # 4,785
46 Rsquared_rf = cor(test_data$incidents, test_data$incidents_pred_rf)^2 #0,933
47
48 #-----extreme gradient boosted trees -----
49 library("caret")
50
51 fitControl <- trainControl(method = "cv", number = 5)
52 model_egbt_incidents <- train(incidents ~ . -country - year -month,
53   data = train_data,
54   method = "xgbTree",
55   trControl = fitControl,
56   search = "random"
57 )
58
59 #Predictions
60 test_data$incidents_pred_egbt = predict(model_egbt_incidents, newdata = test_data)
61 test_data$pred_res_egbt = test_data$incidents - test_data$incidents_pred_egbt
62
63 rmse_egbt <- rmse(test_data$incidents, test_data$incidents_pred_egbt) #4,792
64 Rsquared_egbt = cor(test_data$incidents, test_data$incidents_pred_egbt)^2 #0,934
65
66 #-----
67 # Visualiseringer af forudsigelser
68
69 #punkt plot over praedikteret vs faktiske
70 point_incidents <- ggplot(data=test_data) +
71   geom_point(aes(x = incidents, y = incidents_pred_baseline, color ="red"), size = 0.3, alpha = 0.2) +
72   geom_point(aes(x = incidents, y = incidents_pred_rf, color ="green"), size= 0.3, alpha=0.2) +
73   geom_point(aes(x = incidents, y = incidents_pred_egbt, color ="blue"), size= 0.3, alpha=0.2) +
74   geom_segment(aes(x = 0, xend = 300, y = 0, yend = 300), color="black", size =0.5) +
75   scale_color_manual(name = "",
76     values=c("green"="green", "blue"="blue", "red" = "red"),
77     labels = c("red" = "Baseline", "green"="Random Forest", "blue" = "EGBtree")) +
78   ylab("Praedikterede antal begivenheder") +
79   xlab("Antal begivenheder")+
80   xlim(0,300)+
81   ylim(0,300)+
82   theme(legend.position = "",
83     legend.title=element_blank(),
84     legend.text = element_text(size = 7),
85     legend.background = element_rect(fill=NA)) +
```

```

86 guides(color = guide_legend(override.aes = list(size=2, alpha = 1)))
87
88 kerneldensity <- ggplot(data=test_data) +
89   geom_density(aes(x= pred_res_baseline), bw= 0.8, color = alpha("red", 0.2), size = 0.3, fill= "red", alpha = 0.5 )+
90   geom_density(aes(x= pred_res_rd_rf), bw= 0.8, color = alpha("blue",0.2), size = 0.3, fill= "blue", alpha = 0.5)+
91   geom_density(aes(x= pred_res_egbt), bw= 0.8, color = alpha("green", 0.2), size = 0.3, fill = "green", alpha = 0.5)+
92   stat_density(aes(x=pred_res_baseline, colour="red"), geom="line", bw= 0.8, position="identity", size = 0, alpha=0.1) +
93   stat_density(aes(x=pred_res_rd_rf, colour="blue"), geom="line", bw= 0.8, position="identity", size = 0, alpha=0.1) +
94   stat_density(aes(x=pred_res_egbt, colour="green"), geom="line", bw= 0.8, position="identity", size = 0, alpha=0.1) +
95   scale_color_manual(name = "",
96     values=c("green"="green", "blue"="blue", "red" = "red"),
97     labels = c("red" = "Baseline", "green"="Random Forest", "blue" = "EGBtree")) +
98   ylab("Densitet") +
99   xlab("Fejl i praediktion") +
100   xlim(-15,15) +
101   theme(legend.position = "bottom",
102     legend.title=element_blank(),
103     legend.text = element_text(size = 7),
104     legend.background = element_rect(fill=NA)) +
105   guides(color = guide_legend(override.aes = list(size=2, alpha = 1)))
106
107 setwd(Latexfigure)
108 png("incidents_figure.png")
109 grid.arrange(point_incidents, kerneldensity)
110 dev.off()

```