The Effect of Climate Change on Volatility Indicies for the S&P500 Index Modelled by Machine Learning Methods

Master's Thesis

Lise Lønsmann Nielsen, Mathilde Bach
& Signe Denhardt Brandt Nielsen

Aalborg Universitet

2. juni 2022

AALBORG UNIVERSITY

STUDENT REPORT

10. Semester
Matematik-Økonomi
Skjernvej 4A
9220 Aalborg Øst
http://www.aau.dk

Titel:

The Effect of Climate Change on Volatility Indicies for the S&P500 Index Modelled by Machine Learning Methods

Emne:

Master's Thesis

Projekt Periode:

Februar 2022 - juni 2022

Projekt Gruppe:

1.211

Gruppemedlemmer:

Lise Lønsmann Nielsen

Mathilde Bach

Signe Denhardt Brandt Nielsen

Vejleder:

J. Eduardo Vera-Valdés

Sideantal: 103
Antal Appendix: 3
Projekt afsluttet: 2. juni 2022

Synopsis:

This project investigates whether climate change affects the volatility of the S&P500 index. This will be done by modelling the VIX index and the realized variance using autoregressive models and machine learning methods.

The climate changes are represented by Google Trends variables, and to examine whether these variables have an impact on the volatility indices, we first construct models which only consist of previous values of the index and then compare these with models also containing the Google Trends data representing climate changes.

The models will be compared in-sample by adjusted R^2 values and AIC values as well as out-of-sample by MAE and MSE values. One of the main results is that in-sample, we obtain the highest R^2 values for both VIX and realized variance with models containing the Google Trends variables as external regressors. This indicates that climate change contributes to the explanation of the volatility indices.

Forord

Dette projekt er udarbejdet af gruppe 1.211, som studerer Matematik-Økonomi på 10. semester i foråret 2022 ved School of Engineering and Science på Aalborg Universitet. Projektet er et kandidatspeciale, og titlen er *The Effect of Climate Change on Volatility Indicies for the S&P500 Index Modelled by Machine Learning Methods*.

Tak til vejleder J. Eduardo Vera-Valdés for god og konstruktiv vejledning.

Under udarbejdelsen af projektet er følgende programmer anvendt:

- **Overleaf**
- **RStudio**
- **Google Drev**

Læsevejledning

Der vil igennem rapporten fremtræde kildehenvisninger efter Harvard-metoden, så der i teksten henvises til en kilde med [Efternavn, År].

Kilderne er samlet i en litteraturliste, som findes på Side 83. Henvisningerne i teksten fører til litteraturlisten, hvor bøger er angivet med forfatter, titel, forlag og årstal, mens internetsider er angivet med forfatter, titel, link og dato. Kildehenvisningerne vil primært stå i starten af hvert kapitel og afsnit, med undtagelse af nogle få steder hvor de vil stå direkte i teksten.

På Side *vi* ses indholdsfortegnelsen, og appendiks påbegyndes på Side 85.

Underskrifter

Lise Lønsmann Nielsen
llni17@student.aau.dk

Mathilde Bach
mbac17@student.aau.dk

Signe Denhardt Brandt Nielsen
sdbn17@student.aau.dk

Indholdsfortegnelse

Indledning 1

Miljø og klimaforandringerne er emner, som er særligt aktuelle på nuværende tidspunkt, idet det er ved at få alvorlige konsekvenser for livet på jorden. Konsekvenserne indebærer blandt andet stigende temperaturer, stigende hav samt mere ekstremt vejr, som viser sig i form af hyppigere forekomster af naturkatastrofer. Dette kunne være flere oversvømmelser, skovbrænde, skybrud eller storme, hvilket især rammer de fattige lande i de tropiske og subtropiske regioner. Derudover kan klimaforandringer virke som en konfliktforstærker, sådan at nogle konflikter eller endda krige kan være en konsekvens af klimaforandringer, fordi klimaforandringerne netop lægger et stort pres på ressourcer, hvilket kan skabe nød og dermed desperation, [Care-Danmark]. Det vil sige, at klimaforandringerne kan have graverende konsekvenser for menneskene, hvilket kan skabe en usikkerhed om fremtiden. Det er derfor interessant at undersøge, om den usikkerhed klimaforandringerne skaber også afspejles i aktiemarkedet, og om det også her skaber en større usikkerhed. I projektet er vi derfor interesserede i at undersøge om volatiliteten for S&P500 indekset påvirkes af klimaforandringerne, og om der er udsving, som kan forklares herudfra.

Vi vil gennem dette projekt opstille forskellige modeller for både VIX indekset samt den realiserede varians for S&P500 indekset, hvor klimaforandringer i form af Google Trends data indgår som variable. Med disse modeller vil vi undersøge, om det at inddrage klimaforandringerne i modelleringen vil forbedre forklaringsgraden samt prædiktionerne af volatilitetsindekserne. Modellerne vil alle blive baseret på data fra samme periode samt anvendt til at prædiktere data fra samme periode, for på den måde at have bedst mulige forudsætninger for at kunne sammenligne modellerne direkte.

1.1 Afgrænsning

Det kan være kompliceret at måle klimaforandringerne direkte, og derfor vil vi i dette projekt anvende Google Trends søgninger som en proxy for dette, da disse kan give en indikation af befolkningens opmærksomhed og bekymringer om klimaforandringer. Derudover vil vi udelukkende anvende VIX indekset samt den realiserede varians for S&P500 indekset som volatilitetsmål til trods for, at der findes flere forskellige.

1.2 Problemformulering

På baggrund af introduktionen og afgrænsningen er følgende problemformulering opstillet:

Hvordan påvirkes volatilitetsindekser for S&P500 indekset af klimaforandringerne, som beskrives i form af Google Trends data? Og bidrager Google Trends data til forklaringen af volatilitetsindekserne?

Problemformuleringen besvares blandt andet ved hjælp af teori omkring tidsrækkeanalyse og en række machine learning metoder.

Introduktion 2

Det vil i dette projekt undersøges, om klimaforandringerne som repræsenteres i form af Google Trends variable påvirker volatiliteten af S&P500 indekset, hvorfor vi i dette kapitel vil introducere en række begreber og emner, der er nødvendige for den videre analyse.

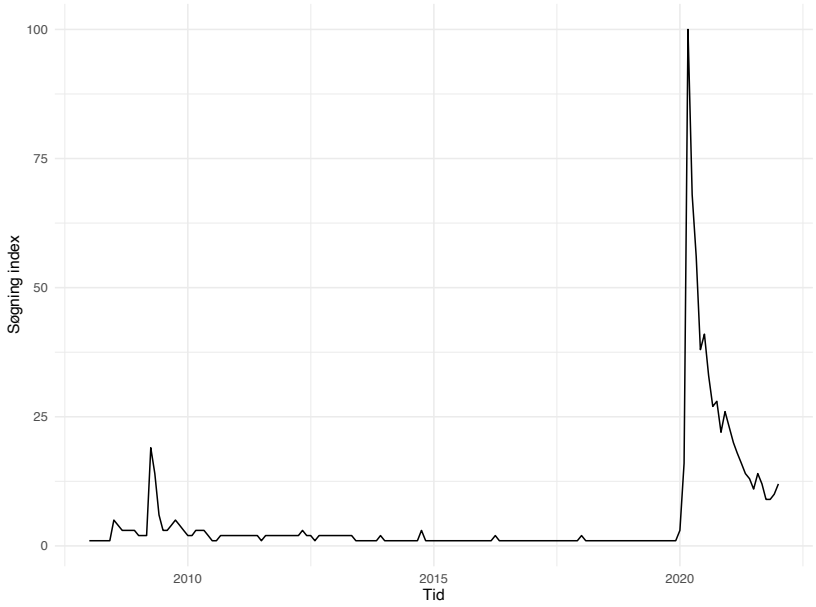
2.1 Google Trends

Dette afsnit er baseret på kilderne [Xiong et al., 2016], [Choi og Varian, 2012] og [Google].

Siden år 2004 har Google indsamlet og gemt antallet af søgninger på hvert enkelt ord, og disse tal er offentlig tilgængelige i databasen Google Trends. Derfor kan Google Trends data anvendes som en repræsentation af den offentlige interesse i forskellige emner, og dermed kan Google Trends data anvendes som variable i matematiske modeller, hvis det eksempelvis ønskes at undersøge sammenhænge mellem forskellige begivenheder og aktiekurser.

Google Trends leverer et tidsrækkeindeks for mængden af forespørgsler, som brugere indtaster på Google i et geografisk areal. Google Trends data er tilgængeligt for hvert land, stat eller mindre region, og søgningerne klassificeres i grupperinger og er normaliseret i forhold til tid og lokation, således at forskellige termer lettere kan sammenlignes. Det vil sige, at hvert datapunkt divideres med det totale antal søgninger i den lokation og det tidsinterval, som repræsenteres. Dette gøres for at sammenligne relativ popularitet, sådan at lokationer med flest søgninger ikke nødvendigvis rangeres højest. De resulterende tal er skaleret i intervallet 0 til 100, sådan at det største antal søgninger i tidsperioden normaliseres til at være 100.

Et eksempel på en tidsrække fra Google Trends er vist på Figur 2.1, hvor det månedlige indeks for antal søgninger i USA på *Pandemic* er vist. Her ses, at antallet stiger voldsomt i marts 2020, hvilket stemmer overens med, at det var i denne måned, hvor de første store nationale nedlukninger skete på grund af corona. Desuden ses et udsving i 2009, hvor der var et udbrud af influenza A H1N1, også kaldet svineinfluenza.



Figur 2.1. Google Trends for *Pandemic*.

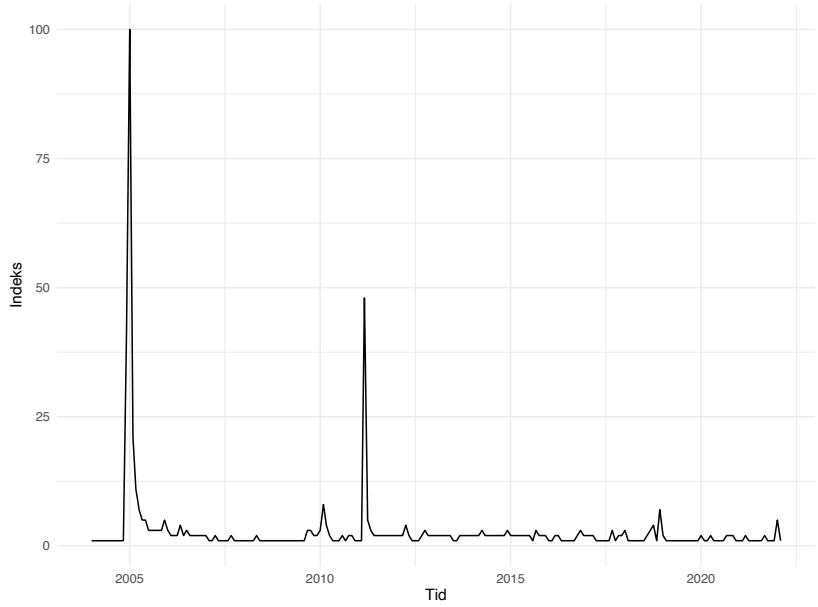
Google Trends data reflekterer altså søgninger, som brugere foretager på Google hver dag, men det kan også indeholde irregular søgningsaktivitet, såsom automatiske søgninger. Dog filtreres nogle typer af søgninger ud, såsom duplikerede søgninger, hvor en person søger på samme term mange gange over en kort periode. Desuden frasorteres observationer, hvis ekstremt få personer har foretaget søgningen.

Eftersom Google Trends data kan anvendes til at afspejle befolkningens interesse i forskellige emner, kan bestemte søgeord anvendes som mål for klimaforandringerne.

2.2 Klimaforandringer

Klimaforandringer er blevet et vigtigt emne, eftersom den menneskelige aktivitet har ført til forhøjede mængder af drivhusgasser i atmosfæren. Gasserne medfører drivhuseffekten, hvilket gør, at temperaturen på jorden stiger, og der sker markante ændringer i vores klima. Siden den industrielle revolution er temperaturerne på jorden steget med 1 grad, og denne globale opvarmning forstyrrer vores klima. Klimaforandringer påvirker hele verden, idet de forårsager hyppigere forekomst af naturkatastrofer. Hvis der ikke ændres retning, risikeres blandt andet, at op i mod 90.000 mennesker vil dø årligt som følge af hedebølger, eller at 2,2 millioner mennesker vil blive ramt af oversvømmelser fra havet, og i sidste ende vil klimaforandringerne kunne ramme vores mad- og vandforsyninger. Det store fokus på klimaforandringerne er essentielt, fordi udviklingen forsætter og problemerne kun bliver vanskeligere og mere bekostelige at løse med tiden, [Union, 2022].

Det er vanskeligt at måle niveauet af disse klimaforandringer, men Google Trends kan afspejle, hvordan verden ser og reagerer på dem. Det vil sige, at Google Trends data kan vise, hvornår der har været naturkatastrofer og give en indikation på, hvor alvorligt det har været. På Figur 2.2 ses månedligt Google Trends data for søgninger på termet *Tsunami*, hvor der på grafen er to spikes, i henholdsvis januar 2005 og marts 2011. Det første udsving henviser til jordskælvet i Det Indiske Ocean den 26. december 2004, som igangsatte flere ødelæggende tsunamier, blandt andet i Indonesien, Sri Lanka, Indien og Thailand. Det andet store udsving skyldes jordskælvet i Japan i 2011, hvilket også medførte alvorlige tsunamier langs den japanske kyst. Det ses dermed, at Google Trends data både afspejler hvornår og hvor kraftige disse tsunamier har været. Dette gælder også for andre typer naturkatastrofer, og derfor kan Google Trends data anvendes videre i projektet som variable, der beskriver interessen for klimaforandringerne og dermed omfanget af dem.



Figur 2.2. Google Trends data for *Tsunami*.

Med disse Google Trends variable, der beskriver klimaforandringerne, kan modeller opstilles for at undersøge, hvordan de påvirker usikkerheden i aktiemarkedet, hvor usikkerheden er målt ved et volatilitetsindeks. Nogle af de mest anvendte volatilitetsindekser er den Realiserede Varians (RV) samt VIX indekset, og disse vil derfor blive beskrevet i næste afsnit.

2.3 Volatilitet for S&P500

Volatilitet er et statistisk begreb, som bruges omkring det finansielle marked og er raten for, hvor meget en aktiepris falder eller stiger over en given periode. Volatilitet beregnes som standardafvigelsen af en akties afkast over en given periode. Hvis prisen på en aktie svinger kraftigt over en kort periode, så siges aktien at have høj volatilitet, modsat siges aktien at have lav volatilitet, hvis prisen ligger stabilt, [Fidelity-International].

Volatilitet i det finansielle marked kan som tidligere nævnt måles på flere måder, og der vil i de næste afsnit blive gennemgået to indekser, som måler volatiliteten, hvilke begge er beregnet på baggrund af S&P500 aktieindekset. Dette aktieindeks er valgt, eftersom det består af de 500 største amerikanske virksomheder og derfor afspejler de generelle tendenser i markedet.

2.3.1 VIX indekset

Dette afsnit er baseret på kilderne [Whaley, 2009] og [Cboe, 2022]. Desuden er data hentet fra [Yahoo-Finance, 2022] og [Nasdaq, 2022].

Et volatilitetsmål for S&P500 er The Market Volatility Index, forkortet VIX indekset, som blev oprettet af Chicago Board Options Exchange (CBOE) i år 1993. VIX indekset beregnes på baggrund af optionspriser fra S&P500, og det repræsenterer markedets forventninger for volatiliteten over de kommende 30 dage. VIX måler derfor den volatilitet, som investorerne forventer at se, og en høj værdi i VIX indekset afspejler en stor frygt blandt investorerne for et potentielt fald i markedet. Indekset er derfor overordnet et udtryk for niveauet af risiko eller frygt i markedet, og kan derfor bruges i investeringsstrategier.

VIX indekset blev oprindeligt introduceret for at have et sammenligningsgrundlag til den kortsigtede forventede volatilitet i markedet, og blev dengang beregnet tilbage til 1986 for at kunne sammenligne med Black Monday krisen i 1987. VIX indekset giver derfor muligheden for at sammenligne det nuværende niveau af frygt i markedet med historiske niveauer, hvor der i dag ofte sammenlignes med finanskrisen i 2008. En anden anvendelsesmulighed er at tegne future- og optionskontrakter på bag-

grund af VIX indekset for dermed at kunne handle med volatilitet. Det er vigtigt for anvendeligheden af VIX indekset, at det beregnes på baggrund af priser fra et aktivt options marked. Det oprindelige VIX indeks blev beregnet på baggrund af S&P100 indekset, grundet at det på daværende tidspunkt var det mest handlede indeks i USA. Sidenhen er S&P500 indekset blevet det mest aktivt handlede indeks, og fra 2003 er VIX indekset blevet beregnet på baggrund af optionspriser herfra. I 2003 blev der desuden også foretaget den ændring, at indekset også beregnes på baggrund af out-of-the-money optioner fremfor udelukkende at-the-money optioner. Forskellen mellem out-of-the-money og at-the-money ligger i, hvordan strike prisen på optionen er i forhold til markedsværdien af det underliggende aktiv. Hvis markedsværdien for en call-option til maturity date er højere end den aftalte strike price, sådan at der er en fortjeneste, siges optionen at være in-the-money. Modsat, hvis markedsværdien er lavere end strikeprisen for en call option, siges den at være out-of-the-money. Hvis markedsprisen er lig strike prisen, siges optionen at være at-the-money. At inkludere en yderligere type af optioner til beregningen af VIX indekset bidrager til, at det er mindre sensitivt og dermed mindre påvirkelig overfor manipulation.

Chicago Board Options Exchange beregner VIX indekset på baggrund af standard S&P500 (SPX) optioner og ugentlige SPX optioner. Der bruges kun optioner, som udløber en fredag. Standard SPX optioner udløber på den tredje fredag i hver måned, og ugentlige SPX optioner kan udløbe alle andre fredage. Derudover bruges kun optioner med mere end 23 dage, og mindre end 37 dage til udløb. Disse optioner vægtes for at give et konstant 30-dages mål af den forventede volatilitet af S&P500 indekset. Intraday VIX indeks beregningen er baseret på snapshots af SPX optioners bid/ask værdier hver 15. sekund, og giver en indikation på den fair markedspris af forventet volatilitet til et bestemt tidspunkt. Derfor er VIX indeks værdierne ofte refereret til som spotværdier, og de beregnes i tidsrummene 3.15-9.15 og 9.30-16.15 opgjort i europæisk tid.

Den generelle formel, som anvendes i beregningen af VIX indekset, er

$$\sigma^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{RT} Q(K_i) - \frac{1}{T}\left(\frac{F}{K_0} - 1\right)^2, \tag{2.1}$$

hvor σ er VIX/100, sådan, at VIX $= \sigma \cdot 100$. Desuden er T tid indtil udløbsdatoen, og F er forward indeks level, hvilket er udledt fra indeks optionspriser som følgende

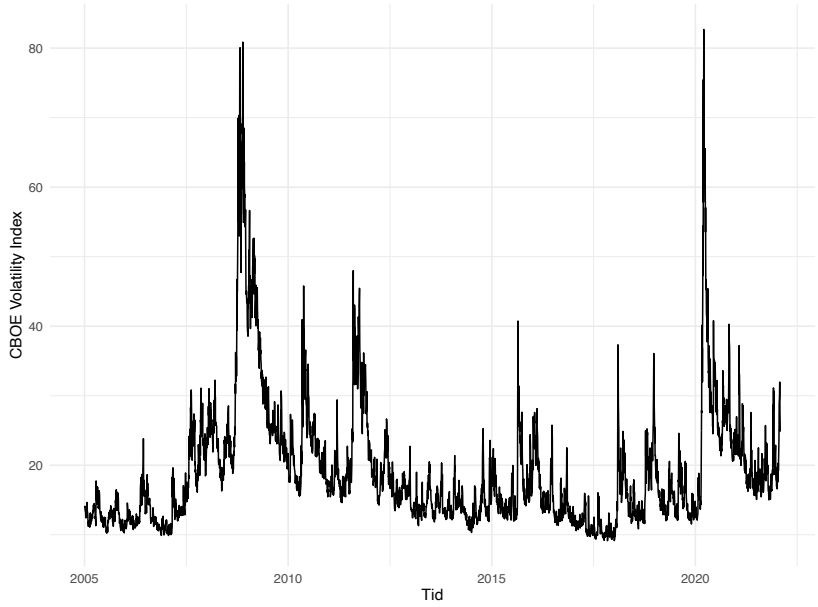
$$F = K + e^{RT} \cdot (\text{call price - put price}),$$

hvor R er den risikofrie interest rate til udløbsdatoen. Derudover er K_0 første strike price under forward indeks level F, mens K_i er strike price for den i'te out-of-the-money option, hvilket vil sige en call, hvis $K_i > K_0$ og en put, hvis $K_i < K_0$, samt både put og call, hvis $K_i = K_0$. Herudover er ΔK_i intervallet mellem strike prices, så

$$\Delta K_i = \frac{K_{i+1} - K_{i-1}}{2}.$$

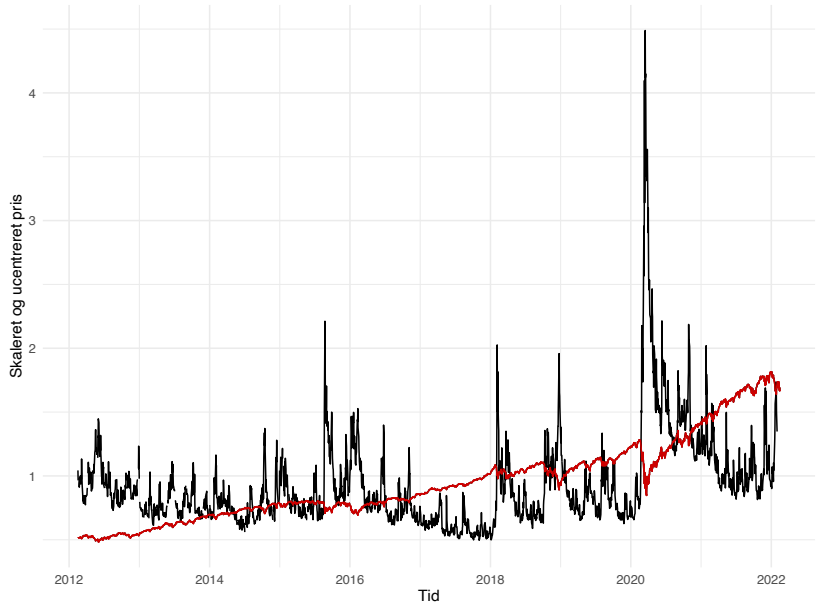
Ydermere er $Q(K_i)$ midtpunktet af bid-ask spread for hver option med strike K_i.

For at vurdere hvornår niveauet af VIX indekset er højt, er det nødvendigt at sammenligne med historiske tal. Figur 2.3 viser det daglige VIX indeks fra januar 2005 til februar 2022. Her ses, hvordan VIX indekset som reaktion på uventede markeds- og verdensbegivenheder har perioder med meget høje udsving. I finanskrisen i oktober 2008 ses et spike, hvor VIX indekset når et højdepunkt med en værdi på 80. Et andet eksempel er marts 2020, hvor coronapandemien tvinger verden til nedlukninger og resulterer i en VIX værdi på 82. Disse tendenser afspejler investorernes usikkerhed og frygt under uforudsigelige kriser og uro i markedet, og dette er grunden til, at VIX indekset betegnes som målet for investorernes frygt.



Figur 2.3. VIX indekset fra januar 2005 til februar 2022.

Figur 2.4 viser det standardiserede VIX indeks samt standardiserede S&P500 indeks fra 2012 til 2022. Det kan ses, at de to grafer har en tendens til at reagere modsat på begivenheder, hvilket betyder, at der er en negativ korrelation mellem S&P500 og VIX indekset. Det vil sige, at et kraftigt fald i S&P500 indekset følges af en stor stigning i VIX indekset, hvilket eksempelvis ses ved coronapandemien i 2020. Modsat ses, at når priserne i S&P500 indekset stiger jævnt, ligger VIX indekset på et lavt niveau. Udviklingen skyldes, at hvis den forventede volatilitet i markedet stiger, så kræver investorerne en højere risikopræmie, hvilket gør, at prisen på aktiverne falder.



Figur 2.4. Standardiseret VIX (sort) og standardiseret S&P500 (rød) fra 2012-2022.

For at kunne afgøre hvad et højt niveau af VIX indekset er, laves deskriptiv statistik for observationerne fra 2005 til 2022. Resultatet af dette ses i Tabel 2.5.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
VIX	9,14	13,16	16,41	19,14	22,10	82,69

Tabel 2.5. Deskriptiv statistik for VIX indekset fra 2005 til 2022.

Over hele perioden er den gennemsnitlige værdi for VIX indekset 19,14. Derudover er den 3. kvartil 22,10, det vil sige, at 75% af værdierne i perioden 2005 til 2022 ligger under denne værdi. Dermed er der kun 25% sandsynlighed for at observere en værdi over 22,10. En sådan værdi kan derfor betragtes som værende høj, og afspejler stor frygt hos investorerne for et potentielt fald i markedet. Derudover er maksimumværdien for VIX indekset 82,69 i den udvalgte periode, hvilken forekom den 16.03.2020 i forbindelse med coronapandemien. Det kan derfor tyde på, at coronapandemien har haft en væsentlig indflydelse på VIX indekset og dermed frygten i forbindelse med aktiemarket.

2.3.2 Realiseret varians

Dette afsnit er baseret på kilderne [Rahimikia og Poon, 2021] og [Oxford-Man-Institute, 2022].

Et andet meget anvendt volatilitetsmål for aktiver, kaldes den realiserede varians, hvilken beregnes fra aktivernes højfrekvens afkast. Antag, at aktivets prisproces P_t følger den stokastiske proces

$$\mathrm{d}\log(P_t) = \mu_t \,\mathrm{d}t + \sigma_t \,\mathrm{d}W_t,$$

hvor μ_t er driften og σ_t er volatilitets processen, som er en cádlág funktion. Desuden er W_t en standard brownsk bevægelse og t angiver dagen. Den integrerede varians er for dag $t-1$ til t defineret som

$$IV_t = \int_{t-1}^{t} \sigma_s^2 \,\mathrm{d}s,$$

og dette er en latent variabel, eftersom den ikke kan observeres direkte. Derfor anvendes den realiserede varians som en estimator for den integrerede varians, og den beregnes for dag t som

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2, \tag{2.2}$$

hvor M er antallet af observationer på en dag. Derudover er $r_{t,i}$ logafkastene, sådan at

$$r_{t,i} = \log(P_{t-1+i\delta}) - \log(P_{t-1+(i-1)\delta}),$$

hvor afstanden mellem observationerne er givet ved δ. Når $\delta \to 0$, så konvergerer den realiserede varians, RV_t, mod den integrerede varians IV_t. Dog medfører det microstructure noise i priserne, når frekvensen for observationerne stiger. At der er microstructure noise betyder, at de sande priser, P_t, observeres med en støj ε_t, således den observerede pris er givet som

$$P_t^* = P_t + \varepsilon_t.$$

Ved at anvende disse observerede priser i analyser, kan det give misvisende resultater, og det er derfor vigtigt at tage højde for microstructure noise, når højfrekvensdata benyttes.

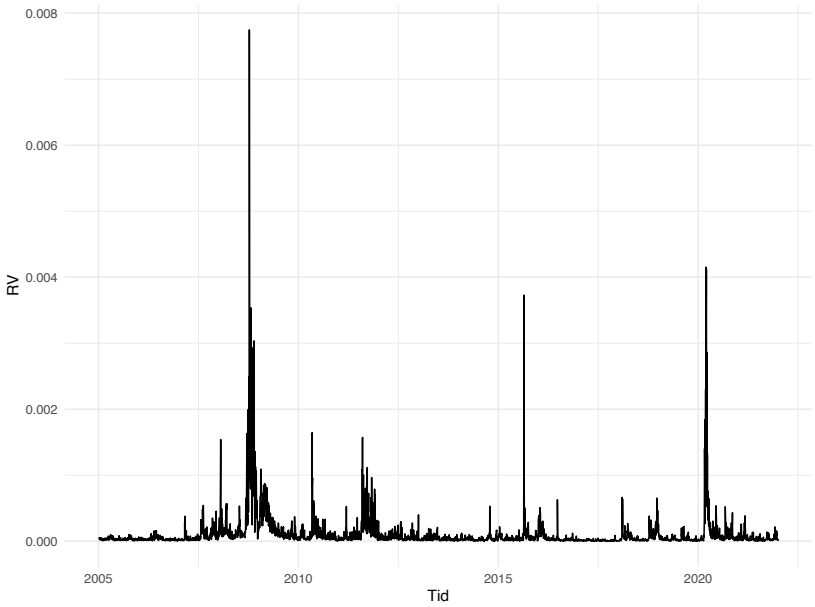
I forbindelse med databehandling og analyse af volatilitet i projektet hentes data for den realiserede varians for S&P500 indekset fra kilden [Oxford-Man-Institute, 2022]. Denne database indeholder daglige afkast

$$r_1, r_2, \ldots, r_T,$$

som er beregnet på baggrund af lukkepriserne, samt de tilhørende følger af daglige realiserede mål, som betegnes

$$RM_1, RM_2, \ldots, RM_T.$$

Disse realiserede mål er høj-frekvens og ikke-parametriske estimatorer af variationen af et aktivs prisproces gennem den tid, hvor aktivet hyppigt handles på en børs. Variationen af priser hen over natten ignoreres, og nogle gange ignoreres variationen også de første få minutter af trading dagen, fordi de observerede priser muligvis indeholder store fejl. De forskellige følger af realiserede mål tager højde for microstructure noise på forskellige måder, eksempelvis ved subsampling eller ved brug af realiserede kernels. Det simpleste af de realiserede mål, som Oxford-Man Institute anvender, er den realiserede varians i (2.2), og eftersom de beregner dette mål på baggrund af afkast, der er samplet hver 5. minut, behøves der ikke at blive taget højde for microstructure noise. På Figur 2.6 ses den realiserede varians for S&P500 indekset over perioden fra januar 2005 til januar 2022.



Figur 2.6. Realiseret varians (5min) for S&P500 indekset fra januar 2005 til februar 2022.

Det kan ses, hvordan den realiserede varians reagerer på markeds- og verdensbegivenheder, hvilket afspejler volatiliteten og dermed usikkerheden i markedet. I 2008 ses et stort udsving i RV, hvilket afspejler usikkerheden grundet finanskrisen i 2008 og tilsvarende i 2020 grundet coronapandemien.

For at kunne afgøre hvornår niveauet af RV er højt, laves deskriptiv statistik over perioden fra 2005 til 2022, som ses i Tabel 2.7.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RV	1,22e-06	1,95e-05	3,70e-05	1,05e-04	8,64e-05	7,75e-03

Tabel 2.7. Deskriptiv statistik over RV indekset for S&P500 fra 2005 til 2022.

Den gennemsnitlige værdi for RV over perioden er 0,0001, mens den 3. kvartil er 0,00008. Det betyder, at 75% af RV-værdierne ligger under denne værdi, hvormed der kun er 25% sandsynlighed for at observere en værdi over 0,00008, og derfor vil sådan en observation betragtes som værende høj. Maksimumsværdien for RV indekset er 0,007 og er observeret i oktober 2008 i forbindelse med finanskrisen, mens maksimumsværdien for VIX indekset var i forbindelse med coronapandemien, hvilket betyder, at disse begivenheder har haft stor indflydelse på usikkerheden i det finansielle marked.

Eftersom coronapandemien har haft en stor indflydelse på volatiliteten i markedet kunne det være interessant at undersøge om det samme gør sig gældende med andre former for naturkatastrofer, som også er forårsaget af klimaforandringerne. Idet katastroferne kan give befolkningen usikkerhed om fremtiden, kan befolkningens reaktion på naturkastrofer afspejle frygten i aktiemarkedet. Derfor vil vi i dette projekt undersøge sammenhængen mellem Google Trends data for klimaforandringerne

og volatiliteten i markedet ved brug af machine learning metoder, hvor vi vil benytte de to netop gennemgået volatilitetsindekser. Derfor vil der i de to følgende kapitler blive introduceret teori om tidsrækkeanalyse og machine learning metoder.

Tidsrækkkeanalyse 3

I dette kapitel introduceres metoder til modellering af tidsrækkedata, idet det ønskes at opstille autoregressive modeller for de to volatilitetsindekser VIX og RV, som blev beskrevet i Kapitel 2. Det er undersøgt og vist af blandt andet [Torelli et al., 2013] samt [Koopman et al., 2005], at volatilitetsindekserne VIX og RV for S&P500 udviser long memory, hvorfor teori omkring dette beskrives i kapitlet. Tilstedeværelsen af long memory er ofte påvist i finansielle aktivers afkastprocesser, hvilket indikerer, at choks i volatilitetsprocesserne kan være persistente, sådan at effekten varer ved i længere tid. Dette medfører, at tidligere afkast for et aktiv kan anvendes til at forudsige de nuværende afkast, hvilket dog er i strid med den efficiente markedshypotese. Markedshypotesen siger, at investorer ikke har mulighed for at udnytte informationen fra de historiske priser til at opnå konsistent høj profit, [Ghosh og Bouri, 2022].

En model som beregner de nuværende værdier på baggrund af tidligere værdier kaldes en autorregressiv model, og defineres som følgende.

Definition 3.1: Autoregressiv model

En autoregressiv proces af orden p, AR(p), er en stationær proces på formen

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + w_t,$$

hvor $w_t \sim \text{wn}(0,\sigma^2)$ er hvid støj, og $\alpha_p \neq 0$, [Shumway og Stoffer, 2017]. Tilføjes eksterne regressorer til modellen kaldes det en autoregressiv distributed lag (ADL) model, hvilken defineres som

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} X_{t-j}\beta_j + w_t,$$

hvor X_{t-j} er den $t-j$'te række i matricen X, som består af de eksterne regressorer, og β_j er vektoren bestående af de tilhørende koefficienter. Eftersom der er p lags på y_t og q lags på X_t, kaldes det en ADL(p,q)-model, [Davison og MacKinnon., 2009].

3.1 Long memory

Dette afsnit er baseret på kilderne [Baillie, 1996] og [Bobeica og Bojesteanu, 2008]. Hvis der er long memory i en tidsrække, vil autokorrelationerne være meget længere tid om at aftage end med en ARMA model, hvor autokorrelationen aftager med en eksponentiel rate. Autokorrelationensfunktionen (ACF) vil vise persistens, som hverken stemmer overens med en I(1) eller en I(0) proces, da der er for meget afhængighed på lang sigt.

En tidsrække proces y_t med autokorrelationsfunktionen γ_j til lag j har egenskaben long memory, hvis

$$\lim_{n\to\infty} \sum_{j=-n}^{n} |\gamma_j| = \infty, \tag{3.1}$$

hvor

$$\gamma_k \approx ck^{2d-1},$$

for $d \in (0{,}1/2)$ samt en konstant c, [Haldrup og Valdés, 2015]. En stationær og invertibel ARMA proces har autokorrelationer, som er begrænset, hvilket vil sige, at for et stort k og $0 < m < 1$ gælder

$$|\gamma_k| \leq cm^{-k},$$

hvilket medfører, at (3.1) ikke er opfyldt, og disse processer kaldes derfor short memory processer.

For at undersøge om en tidsrække udviser long memory, kan Hurst koefficienten H beregnes på baggrund af R/S statistikken, hvor R/S står for range over standardafvigelse. Denne statistik beregnes som R_T/S_T, hvor R_T er defineret som

$$R_T = \max_{0\leq j\leq T} \left\{ \sum_{j=1}^{T} (y_j - \mathbb{E}[y]) \right\} - \min_{0\leq j\leq T} \left\{ \sum_{j=1}^{T} (y_j - \mathbb{E}[y]) \right\},$$

hvilket er spændet af summerne af tidsrækkens afvigelse fra middelværdien. Derudover er S_T tidsrækkens sample standardafvigelse defineret som

$$S_T = \left\{ 1/T \sum_{j=1}^{T} (y_j - \bar{y})^2 \right\},$$

hvor $\bar{y}$ er tidsrækkens sample mean. Hurst koefficienten beregnes dernæst som

$$H = \frac{\log(R_T/S_T)}{\log(T)}. \tag{3.2}$$

Statistikken H ligger i intervallet [0,1], og der er fire forskellige tilfælde, som kan beskrive tidsrækkens egenskaber. Hvis $0{,}5 < H \leq 1$, så er tidsrækken persistent, og den udviser evidens for long memory, hvilket modstrider den efficiente markedshypotese. Jo højere værdien af H er, jo mere persistent er tidsrækken og jo mere long memory udviser den. Hvis $0 < H < 0{,}5$, så er tidsrækken anti-persistent, hvilket vil sige, at den har short memory. Det indikerer hurtige ændringer i tidsrækkens retninger. Hvis $H = 0$, så udviser tidsrækken ikke afhængighed på lang sigt, og hvis $H = 0{,}5$, så følger tidsrækken en random walk, hvilket understøtter den efficiente markedshypotese. For tidsrækker uden nogen afhængighed ses ofte en Gaussisk fordeling, dog er det ikke ofte, at en proces har $H = 0{,}5$, [Ghosh og Bouri, 2022].

Fraktionelt integrerede processer er long memory, så en process y_t siges at være integreret af orden d, eller I(d), hvis

$$(1-L)^d y_t = u_t,$$

hvor L er lag operatoren, $-0{,}5 < d < 0{,}5$ og u_t er en stationær og ergodisk proces med et begrænset spektrum, som har positive værdier til alle frekvenser. For $0 < d < 0{,}5$ defineres

$$(1-L)^d = 1 + \sum_{j=1}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)j!} L^j,$$

hvor

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t \quad \text{for } x \neq 0, -1, -2, \ldots,$$

er gammafunktionen. Modelleringen af fraktionelt integrerede processer sker ved ARFIMA-modeller, hvilke defineres som følgende.

Definition 3.2: ARFIMA

En proces kaldes en ARFIMA(p,d,q) proces, hvis dens d'te differens er en stationær og invertibel ARMA(p,q) model, hvor $-0{,}5 < d < 0{,}5$. Det vil sige, at y_t er en ARFIMA(p,d,q) hvis

$$\phi(L)(1-L)^d(y_t-\mu) = \theta(L)\varepsilon_t, \tag{3.3}$$

hvor $\phi(z) = 1-\phi_1 z - \cdots - \phi_p z^p$ og $\theta(z) = 1+\theta_1 z + \cdots + \theta_q z^q$ er lagpolynomier af henholdsvis orden p og q, som begge har rødder udenfor enhedscirklen. Derudover er $\varepsilon_t \sim \text{IID}(0,\sigma^2)$, [Ørregaard Nielsen og Frederiksen, 2007].

Parameteren d i ARFIMA processen afgør, hvor meget memory der er i processen. Hvis $d > -0{,}5$, så er processen invertibel, og hvis $d < 0{,}5$, så er den stationær. Hvis $0 < d < 0{,}5$, så udviser processen long memory, og dens autokorrelationer er alle positive, og de aftager med en hyperbolsk rate. Hvis $-0{,}5 < d < 0$, så summer alle absolutværdierne af processens autokorrelationer til en konstant, hvormed processen har short memory og er stationær. I dette tilfælde siges processen at være antipersitent og alle dens autokorrelationer, undtagen til lag 0, er negative og går mod 0 med en hyperbolsk rate. Hvis $d = 0$ svarer det til, at der ikke skal differenses, hvormed det i forvejen er en stationær og invertibel process, som udviser short memory.

3.1.1 Maksimum Likelihood estimation af d

Dette afsnit er baseret på kilden [Ørregaard Nielsen og Frederiksen, 2007]. For at opstille ARFIMA modeller er det nødvendigt at estimere d, hvilket kan gøres med maksimum likelihood estimation. Den eksakte gaussiske likelihood funktion for modellen i (3.3) er for $-0{,}5 < d < 0{,}5$ givet ved

$$L_E(d,\phi,\theta,\sigma^2,\mu) = -\frac{T}{2}\ln|\Omega| - \frac{1}{2}(Y-\mu l)^T\Omega^{-1}(Y-\mu l), \tag{3.4}$$

hvor $l = (1,\ldots,1)^T$, $Y = (y_1,\ldots,y_T)^T$, ϕ og θ er parametrene af $\phi(L)$ og $\theta(L)$, μ er middelværdien af Y og Ω er varians-kovarians matricen for Y, som er en funktion af d og resten af parametrene i modellen. Hvis alle parametrene samles i vektoren $\gamma = (d,\phi^T,\theta^T,\sigma^2,\mu)^T$, så opnås den eksakte maksimum likelihood (EML) estimator ved at maksimere likelihood funktionen i (3.4) med hensyn til γ.

[Sowell, 1992] viste, at EML estimatoren af d er $\sqrt{T}$ konsistent og asymptotisk normal, så

$$\sqrt{T}(\hat{d}_{\text{EML}} - d) \xrightarrow{d} N(0,(\pi^2/6 - C)^{-1}),$$

hvor $C = 0$, når $p = q = 0$ og ellers $C > 0$.

Der findes også andre metoder til estimering af d, men dette er fremgangsmåden, som vil blive anvendt i projektet. I næste kapitel vil machine learning metoder blive introduceret for blandt andet at kunne reducere dimensionen af parametre i regressionsmodeller.

Machine learning 4

Når lineær regression anvendes for flere forklarende variable er der to formål, hvor det første af disse er modelforståelse. Her er ideen at fjerne irrelevante variable, så vi kan opnå en model, som er let forståelig og anvendelig. Det andet formål er at opnå en model, som prædikterer med høj nøjagtighed. Dette kan opnås ved at begrænse eller krympe koefficienterne, så variansen reduceres, dog på bekostning af en smule bias. Der er forskellige metoder til at opnå disse to formål:

1. Subset selektion
2. Shrinkage
3. Dimensions reduktion

Ved subset selektion identificeres en delmængde af variablene, som forklarer responsvariablen bedst, hvorefter en model opstilles på baggrund af denne delmængde. Ved shrinkage opstilles derimod en model med alle variablene, og så krympes koefficienterne mod 0, hvilket er en måde at reducere variansen på. Dimensions reduktion projekterer variablene ned på et rum med lavere dimension og anvender projektionerne som forklarende variable i modellen. Vi har valgt at anvende shrinkage og dimensions reduktion metoder, som derfor gennemgås i det følgende.

4.1 Shrinkage

Shrinkage udføres ved at introducere en straf i modellen, som er givet ved størrelsen af de estimerede koefficienter. Der er to forskellige måder at udføre shrinkage, hvilke kaldes *Ridge regression* og *Lasso*. Forskellen mellem disse metoder er måden, hvorpå størrelsen af de estimerede parameter måles, og dermed hvordan straffen måles. Ved begge metoder centreres og skaleres data inden brug, idet metoderne lægger begrænsninger på størrelsen af koefficienterne, hvilket vil afhænge af hver variabels størrelse.

4.1.1 Ridge regression

Når en lineær regressionsmodel opstilles, estimeres koefficienterne β_0,$\beta_1, \ldots, \beta_d$ ved at finde de værdier, som minimerer *Residual Sum of Squares*, der er givet ved

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} \beta_j x_{ij} \right)^2.$$

Når Ridge regression i stedet anvendes, udvides dette, således der tilføjes en straf til minimeringsligningen. Derfor vil Ridge regression koefficienterne, $\hat{\beta}^R$, være givet ved

$$\hat{\beta}_\lambda^R = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2 \right\}, \tag{4.1}$$

hvor det ekstra led $\lambda \sum_{j=1}^{d} \beta_j^2$ kaldes shrinkage straffen, og $\lambda \geq 0$ er en tuning parameter, der bestemmes seperat.

Hvis der er mange korrelerede variable i en lineær regressionsmodel, ses det ofte, at koefficienterne kan være misvisende og have en høj varians. Dette skyldes, at en stor positiv koefficient på én variabel kan gå ud med en lige så stor negativ koefficient på en af dens korrelerede variable. Dette problem løses ved at tilføje en betingelse for, hvor store koefficienterne må blive, hvilket shrinkage straffen i (4.1) netop tager højde for. Ligesom ved OLS, søger Ridge regression altså at estimere koefficienter, så der opnås et godt fit ved at minimere RSS, hvilket sker når shrinkage straffen er lille. Dette opnås, hvis $\beta_1, \ldots, \beta_d$ er tæt på 0, og derfor vil effekten af Ridge regression være, at estimaterne af β_j for $j = 1, \ldots, d$ bliver krympet mod 0. Tuning parameteren λ kontrollerer, hvor meget koefficienterne krympes. Hvis $\lambda = 0$ indgår der ikke længere en straf, og vi udfører blot almindelig lineær regression. Når $\lambda \to \infty$, vil indflydelsen af shrinkage straffen stige, og Ridge regression koefficientestimaterne vil gå mod 0. Når λ stiger, falder fleksibiliteten af regressionen, hvilket fører til en faldende varians, men stigende bias. For hver værdi af λ vil der være et tilhørende sæt af koefficienter $\hat{\beta}_\lambda^R$, som løser minimeringsproblemet, og for at vælge den optimale værdi af λ anvendes cross-validation. Dette udføres ved at vælge en mængde af mulige λ værdier, hvorefter cross-validation fejlen beregnes for hver regression opstillet med det tilhørende λ. Dermed kan tuning paramteren λ vælges givet den laveste cross-validation fejl.

Det ses i (4.1), at skæringen β_0 ikke er inkluderet i straffen, hvorfor løsningen kan opdeles i to skridt, først findes skæringsestimatet og dernæst de resterende estimater. Dette gøres ved først at centrere data, så hvert x_{ij} er givet som $x_{ij} - \bar{x}_j$, hvorefter skæringsestimatet beregnes som $\hat{\beta}_0 = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. De resterende koefficienter estimeres ved Ridge regression uden skæring, hvor de centrerede værdier, x_{ij}, er anvendt. Derfor kan problemet opstilles på matrix-vektor form som

$$\text{RSS} = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta, \tag{4.2}$$

hvor X her angiver matricen, som udelukkende består af de d variable, og som dermed ikke inkluderer skærringssøjlen bestående 1-taller. Minimeringsproblemet ved Ridge regression kan løses analytisk ved at opstille RSS som Lagrange funktionen og differentiere denne med hensyn til β, hvorefter funktionen sættes lig 0 og β isoleres. Ligning (4.2) minimeres for

$$\hat{\beta}_\lambda^R = (X^T X + \lambda I)^{-1} X^T y.$$

Eftersom Ridge regression udtrykker koefficientestimaterne på lukket form, kan variansen af koefficientestimatet findes som følgende

$$\begin{aligned} \text{Var}\big(\hat{\beta}_\lambda^R\big) &= (X^T X + \lambda I)^{-1} X^T \text{Var}(y) \big((X^T X + \lambda I)^{-1} X^T\big)^T \\ &= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}, \end{aligned} \tag{4.3}$$

hvor σ^2 er variansen af y.

4.1.2 Lasso

Dette afsnit er baseret på kilden [James et al., 2017]. Ridge regression kan have den ulempe, at den inkluderer alle d prædiktorer i modellen, eftersom ingen af koefficienterne bliver præcis 0. Dermed vil ingen variable blive ekskluderet fra modellen, uanset hvor stor værdien af λ er. Lasso er et alternativ til Ridge regression, som løser denne problemstilling. Lasso koefficienterne findes ved det følgende minimeringsproblem

$$\hat{\beta}_\lambda^L = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{d} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right\}. \tag{4.4}$$

Den eneste forskel mellem denne ligning og udtrykket for Ridge regression i (4.1) er, at β_j^2 nu er erstattet med $|\beta_j|$. Derfor adskiller de to metoder sig ved måden, hvorpå de straffer, eftersom Lasso bruger en ℓ_1 straf i stedet for en ℓ_2 straf. At Lasso anvender en ℓ_1 straf gør, at koefficienterne kan tvinges til at være præcis 0, når tuning parameteren λ er stor nok. Dermed udfører Lasso variabel selektion, eftersom den genererede model ikke nødvendigvis indeholder alle d prædiktorer, men blot en delmængde, hvilket medfører, at modellen er nemmere at fortolke. Der er ikke en løsning på lukket form

for minimeringsproblemet i Lasso, eftersom absolutværdien ikke kan differentieres. Det vil sige, at der ikke er et udtryk på lukket form for parametrene, og derfor skal de estimeres ved numerisk optimering. Dette kan gøres med metoden coordinate descent, hvor én parameter estimeres ad gangen, mens alle andre parametre holdes konstant. Derefter fortættes med næste parameter, og processen itereres indtil et konvergenskriterie er opnået. Ligesom i Ridge regression er valget af λ afgørende i Lasso, og den vælges på tilsvarende vis som ved Rigde regression med cross-validation.

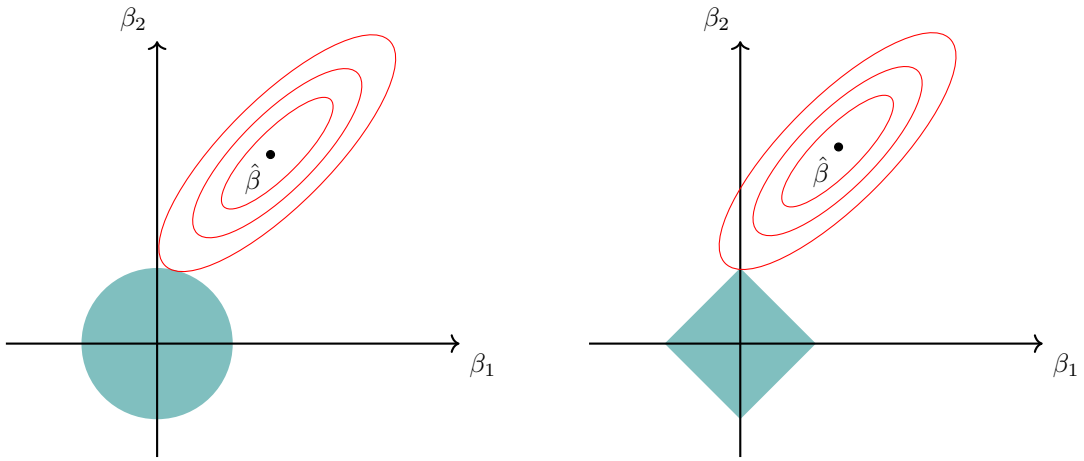
Det kan vises, at Lasso koefficienterne løser følgende minimeringsproblem

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} \beta_j x_{ij} \right)^2, \quad \text{u.b.b} \sum_{j=1}^{d} |\beta_j| \leq s. \tag{4.5}$$

Det vil sige, at for ethvert λ findes et s, så (4.4) og (4.5) vil give de samme Lasso koefficientestimater. Når Lasso udføres, ønskes det at finde det sæt af koefficientestimater, som medfører den mindste RSS værdi givet bibetingelsen, som bestemmer størrelsen af estimaterne. Når s er stor, er betingelsen ikke særlig streng, og dermed tillader vi koefficienterne at være store, så hvis s er stor nok til at OLS løsningen opfylder betingelsen, så vil (4.5) blot give OLS løsningen. Tilsvarende hvis s er lille, så tvinges koefficienterne til at være små for at overholde betingelsen.

Figur 4.1 er baseret på [James et al., 2017], og viser det to-dimensionale tilfælde, hvor betingelsesområderne for henholdsvis $\beta_1^2 + \beta_2^2 < s$ og $|\beta_1| + |\beta_2| < s$ er vist som de turkise områder. Derudover udgør de røde ellipser kurverne for RSS værdierne, som stiger, jo længere væk elipsen er fra $\hat{\beta}$. Hvis s er tilstrækkelig stor, så vil betingelsesområdet indeholde $\hat{\beta}$, og koefficientestimaterne vil derfor være de samme som OLS estimaterne.

Ligning (4.4) samt tilsvarende udtryk for Ridge regression i (4.1) medfører, at koefficientestimaterne er givet ved det første punkt, hvor elipsen har kontakt med betingelsesområdet. Eftersom Ridge regression har et rundt betingelsesområde, så vil denne skæring generelt ikke ligge på en af akserne, og dermed vil Ridge regression koefficienterne ikke være præcis 0. Modsat har Lasso's betingelsesområde hjørner på hver af akserne, og på denne måde vil skæringen mellem området og elipsen ofte ligge på en akse. Når dette sker, vil en af koefficienterne være lig 0, og i højre dimensioner er dette muligt for flere koefficientestimater samtidig. På Figur 4.1b sker skæringen i $\beta_1 = 0$, og på denne måde indeholder den endelige model kun β_2. Hvis der eksempelvis betragtes tre dimensioner, så bliver betingelsesområdet for Ridge regression en kugle og for Lasso et polyeder, men argumenterne fra Figur 4.1 gælder stadig.



(a) Kurver for RSS samt bibetingelsesfunktionen for Ridge.

(b) Kurver for RSS samt bibetingelsesfunktionen for Lasso.

Figur 4.1

Begge shrinkage metoder indeholder et bias/varians trade-off, eftersom en stigning i λ medfører, at variansen falder, men at bias vil stige. De to metoder har hver deres fordele og ulemper, men generelt forventes, at Lasso præsterer bedst, når et lille antal af prædiktorer har betydelig store koefficienter, mens de andre har koefficienter tæt på 0. Tilsvarende forventes, at Ridge regression præsterer bedst, når alle prædiktorerene har omtrent samme størrelse koefficienter.

4.2 Dimension reduktions metoder

Dette afsnit er baseret på kilderne [James et al., 2017], [Sankar, 2021] og [Aggarwal, 2015].

Dimension reduktion er transformationen af data med en høj dimension til en lavere dimension, således antallet af variable i et datasæt reduceres. Før anvendelsen af dimensions reduktions metoder bør prædiktorerne standardiseres, så alle variable er på samme skala. Metoderne transformerer prædiktorerne og derefter fittes en least squares model, hvor de transformerede variable anvendes. Lad X være datamatricen bestående af de originale prædiktorer $X_1, X_2, \ldots, X_d$, og lad Z være datamatricen bestående af $Z_1, Z_2, \ldots, Z_M$, som repræsentere $M < d$ linearkombinationer af de d prædiktorer. Det vil sige, at

$$Z_m = \sum_{j=1}^{d} \phi_{jm} X_j \tag{4.6}$$

for konstanter $\phi_{1m}, \phi_{2m}, \ldots, \phi_{dm}$ og $m = 1, \ldots, M$. Så kan følgende lineære regressionsmodel fittes ved brug af least squares

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4.7}$$

hvor regressionskoefficienterne dermed er $\theta_0, \theta_1, \ldots, \theta_M$. Hvis konstanterne i de linearkombinationer i (4.6) er valgt fornuftigt, kan dimension reduktions metoderne præstere bedre end en least squares regression.

Metoden reducerer problemet med at estimere $d + 1$ koefficienter til et simplere problem med blot $M + 1$ koefficienter, hvor $M < d$. Fra (4.6) gælder, at

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{d} \phi_{jm} x_{ij} = \sum_{j=1}^{d} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{d} \beta_j x_{ij},$$

hvor

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}. \tag{4.8}$$

Det viser, at den lineære regressionsmodel i (4.7) er et specialtilfælde af den originale lineære regressionsmodel

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_d + \varepsilon.$$

Dimension reduktions metoder begrænser dermed de estimerede koefficienter, β_j, fordi de nu er restringeret til at være på formen (4.8). Der kan dog opstå bias i forbindelse med de restringerede koefficienter, men hvis d er stor i forhold til antal observationer n, så kan variansen af de fittede koefficienter reduceres markant, hvis M vælges meget mindre end d. Hvis $M = d$, og søjlerne i Z alle er lineære uafhængige, så pålægger (4.8) ingen restriktioner, og dermed sker der ikke reduktion i dimensionen. Hvis modellen i (4.7) opstilles for dette tilfælde, er det ækvivalent med at anvende least squares på de d originale prædiktorer.

Alle dimension reduktions metoder har to steps. Først opnås de transformerede prædiktorer, $Z_1, Z_2, \ldots, Z_M$, og derefter fittes en regressions model ved brug af disse. Selvom antallet af prædiktorer, der indgår i regressionen, er reduceret til M, er det ikke en variabel selektion metode, idet der anvendes linearkombinationer af alle d originale variable. Beregningen af $Z_1, Z_2, \ldots, Z_M$ kan udføres på forskellige måder, og i de følgende afsnit beskrives to af dem, hvilke kaldes *principal components analyse* og *partial least squares*.

4.2.1 Principal components analyse

I datasæt ses ofte en korrelation mellem variablene, hvilket indikerer at der er signifikant afhængighed mellem dem. Dette medfører, at en delmængde af variablene kan anvendes til at prædiktere værdierne af de resterende variable.

Principal components analyse (PCA) anvendes generelt på data, hvor middelværdien er trukket fra observationerne, således at det er centreret i origo. Ideen med PCA er at rotere datapunkterne i et akse-system, hvor mest mulig varians er beskrevet ved få variable. Variansen af et datasæt langs en bestemt retning kan beskrives direkte ved den tilhørende kovariansmatrix.

Lad Σ være en symmetrisk $d \times d$ kovariansmatrix for $n \times d$ datamatricen X. Så noterer den i,j'te indgang i Σ, σ_{ij}, kovariansen mellem den i'te og den j'te søjle i X. Lad x_{km} betegne indgang k i den m'te søjle, så kan kovariansen σ_{ij} beregnes som

$$\sigma_{ij} = \frac{\sum_{k=1}^{n} x_{ki}x_{kj}}{n}, \qquad \forall i,j \in \{1,\ldots,d\}. \tag{4.9}$$

Så kan (4.9) opskrives som $d \times d$ kovariansmatricen for X:

$$\Sigma = \frac{X^T X}{n}.$$

For projektionen af datasættet X på enhver d-dimensional søjlevektor $\bar{v}$, $X\bar{v}$, er variansen givet ved $\bar{v}^T\Sigma\bar{v}$. At $\bar{v}^T\Sigma\bar{v}$ er variansen gør, at den nødvendigvis er positiv, hvormed kovariansmatricen Σ er positiv semidefinit. Det vil sige, at

$$\bar{v}^T\Sigma\bar{v} = \frac{(X\bar{v})^T X\bar{v}}{n} \geq 0. \tag{4.10}$$

Formålet med PCA er at bestemme de ortonormale vektorer $\bar{v}$, som maksimerer variansen af $X\bar{v}$. For at løse dette maksimeringsproblem opstilles en langrange funktionen som

$$\mathcal{L}(\bar{v}) = \bar{v}^T\Sigma\bar{v} - \lambda(1 - \bar{v}^T\bar{v}), \tag{4.11}$$

hvor λ er lagrange multiplieren. Ved at sætte den afledte af (4.11) lig 0, fås

$$\frac{\partial\mathcal{L}(\bar{v})}{\partial\bar{v}} = 2\Sigma\bar{v} - 2\lambda\bar{v} = 0,$$

hvormed

$$\Sigma\bar{v} = \lambda\bar{v}.$$

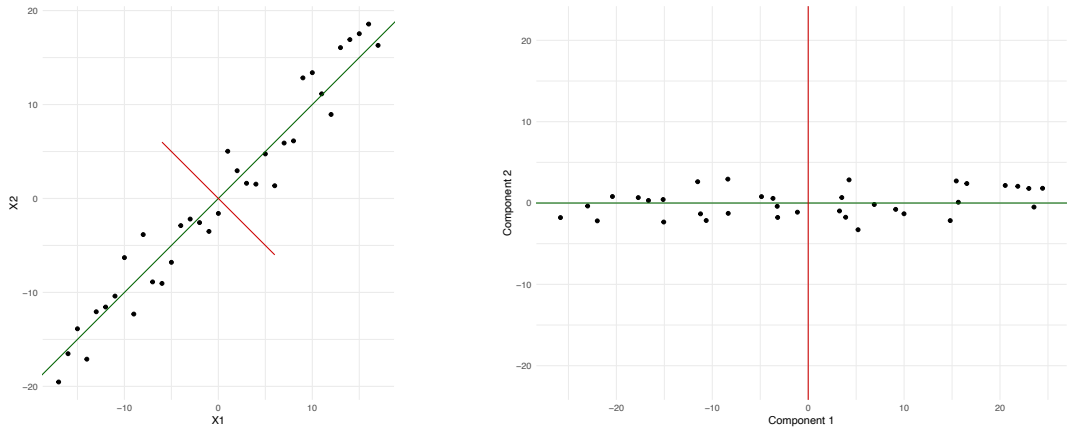
Af dette ses, at $\bar{v}$ er egenvektoren for Σ med tilhørende egenværdi λ. Disse egenvektorer repræsenterer dermed ortogonale løsninger til maksimeringsproblemet af variansen $\bar{v}^T\Sigma\bar{v}$. Egenvektorerne kan findes ved først at diagonalisere kovariansmatricen som

$$\Sigma = P\Lambda P^T,$$

hvilket kan gøres, eftersom den er symmetrisk og positiv semidefinit. Søjlerne af matricen P indeholder de ortonormale egenvektorer af Σ, og Λ er en diagonalmatrix, som indeholder alle de ikke-negative egenværdier. Dermed er indgang Λ_{ii} egenværdien tilhørende den i'te egenvektor af matricen P.

Hvis akserne i et datasæt roteres til det ortogonale sæt af egenvektorer, kan det vises, at alle kovarianser af dette transformerede data er 0. Det vil sige, at den retning med størst varians også er retningen, som fjerner korrelationen. Derudover repræsenterer en egenværdi variansen af datasættet langs den tilhørende egenvektor, og Λ er dermed den nye kovariansmatrix af det transformerede data. Egenvektorer med store egenværdier beskriver dermed mest varians, og disse kaldes principal komponenter. Vi ønsker derfor at opnå et system med nye akser, som kun indeholder de egenvektorer med store egenværdier, det vil sige, at vi søger at forklare mest mulig varians i datasættet med så få principal komponenter som muligt.

På Figur 4.2a ses et eksempel med to variable, hvor de to tilhørende egenvektorer er illustreret ved den røde og grønne linje. Den grønne linje viser den første principal komponent, altså den retning, hvor datapunkterne varierer mest. Den røde viser den anden principal komponent. Ved at rotere akserne, sådan at den første principal komponent udgør x-aksen, fjernes korrelationen mellem variablene, hvormed dimensionen kan reduceres. På Figur 4.2b vises datapunkterne, når akserne er roteret, og her ses, at der ikke er meget variation i koordinaten for y-aksen. På denne måde kan data repræsenteres langs en 1-dimensional linje uden tab af meget information.



(a) **(b)**

Figur 4.2. Illustration af PCA med to dimensioner.

Uden tab af generalitet kan det antages, at søjlerne af P og den tilhørende diagonalmatrix Λ er arrangeret fra venstre mod højre sådan, at det svarer til at have aftagende egenværdier. Så kan den transformerede datamatrix X', i et nyt koordinatsystem efter akserotation til de ortonormale søjler af P, beregnes som den følgende lineære transformation

$$X' = XP.$$

Datamatricen X' har størrelse $n \times d$, men kun de første M søjler fra venstre, hvor $M \ll d$, viser en signifikant variation i værdierne. Disse M søjler svarer til $Z_1, \ldots, Z_M$, som tidligere beskrevet. Hver af de resterende $(d - M)$ søjler af X' vil approksimativt være lig middelværdien af datasættet i det roterede aksesystem, hvilken i et centreret datasæt er lig 0. På denne måde kan dimensionen af datasættet reduceres, mens man samtidig bevarer forklaringen af variationen i datasættet. Variansen af datasættet defineret ved projektionerne langs de første M egenvektorer er lig summen af de tilhørende M egenværdier. I mange anvendelser ses et markant fald i egenvædierne efter de første få, og hovedideen er, derfor at et lille antal principal komponenter er tilstrækkelige til at forklare det meste af variabiliteten i datasættet og sammenhængen med responsvariablen y i en regressionsmodel. Det antages derfor, at de retninger, hvor $X_1, \ldots, X_d$ viser mest variation, er de retninger, som er associeret med responsvariablen Y. Antagelsen er ikke altid opfyldt, men ofte er det en fornuftig nok approksimation til at give gode resultater. Hvis antagelsen er opfyldt, opnås bedre resultater ved at fitte en least squares model på $Z_1, \ldots, Z_M$, hvilket kaldes *principal component regression* (PCR), fremfor at fitte en least squares model på $X_1, \ldots, X_d$, Dette gælder, eftersom $Z_1, \ldots, Z_M$ indeholder det meste eller alt informationen i datasættet, som relateres til responsvariablen og ved kun at estimere $M \ll d$ koefficienter, reduceres risikoen for overfitting. Derfor præsterer PCR godt i tilfælde, hvor de første få principal komponenter er tilstrækkelige til at fange det meste af variationen i prædiktorerene samt forholdet mellem prædiktorerne og responsvariablen. Desuden gælder at jo flere principal komponenter, der anvendes i regressionsmodellen, jo højere bliver variansen, mens bias falder. Antallet af principal komponenter i PCR, M, vælges typisk ved brug af cross-validation.

4.2.2 Partial least squares

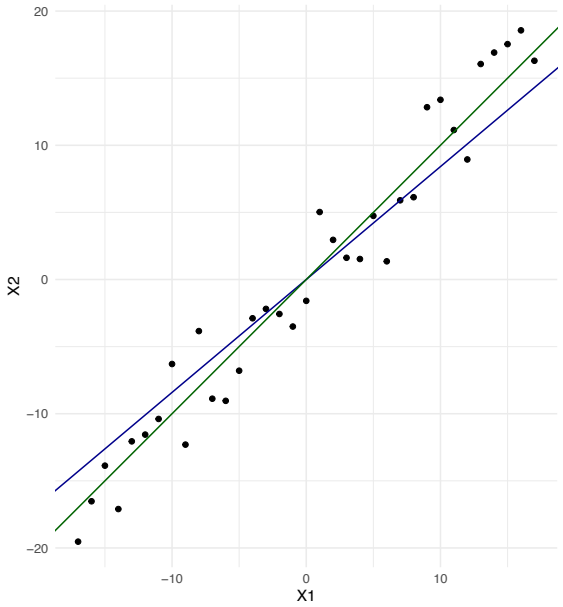
I PCR vælges linearkombinationer ved en unsupervised metode, hvor responsvariablen Y ikke anvendes til at bestemme komponenterne. Det betyder, at der ved PCR ikke er garanti for, at retningerne, som bedst beskriver prædiktorerne også er de bedste retninger til at prædiktere responsvariablen. Derfor introduceres nu et supervised alternativ til PCR, som kaldes *partial least squares* (PLS).

Partial least squares er som PCR også en dimension reduktions metode, hvor der først findes nye variable $Z_1, \ldots, Z_M$, som er linearkombinationer af de originale variable, og dernæst fittes en regressionsmodel med least squares ved brug af de M nye variable. Dog findes $Z_1, \ldots, Z_M$ i stedet ved en supervised metode, hvor der gøres brug af responsvariablen Y. Dette medfører, at PLS komponenterne ikke kun konstrueres til at beskrive prædiktorerne bedst, men også til at kunne prædiktere responsvariablen.

PLS teknikken virker ved at udtrække komponenter fra både de forklarende variable X og responsvariablen Y, sådan at kovariansen mellem dem er maksimeret. På en tilsvarende vis som i PCA findes principal komponenterne med PLS metoden ved at betragte kovariansmatricen mellem de forklarende variable og responsvariablen. Kovariansen mellem X og Y er givet ved X^TY, og for at finde den første principal komponent findes egenvektoren af X^TYY^TX. Når den første komponent er fundet, kan proceduren gentages med opdaterede værdier af X og Y for at finde de næste principal komponenter. Opdateringerne er givet ved

$$X_1 = X - tt^TX,$$
$$Y_1 = Y - tt^TY.$$

Her er $t = Xp$, hvor p er egenvektoren, der tilhører den første egenværdi af X^TYY^TX, [Maitra og Yan, 2008].



Figur 4.3. Illustration af PLS komponent (blå) sammenlignet med PCA komponent (grøn).

Figur 4.3 viser et eksempel, hvor PLS metoden er anvendt på de to variable fra tidligere. Den grønne linje viser den første principal komponent med PCA, som tidligere blev vist på Figur 4.2a, mens den blå linje indikerer den første PLS retning. Det ses, at PLS vælger en anden retning end PCA. Dette skyldes, at PCA udelukkende betragter de forklarende variable, mens PLS også betragter responsvariablen.

Som i PCR bestemmes M typisk fra cross-validation, og prædiktorerne og responsvariablen standardiseres typisk inden anvendelse. Da PLS er supervised kan metoden reducere bias, men variansen kan dog stige, så der ingen overordnede fordel er fra PLS i forhold til PCR.

4.3 Neurale Netværk

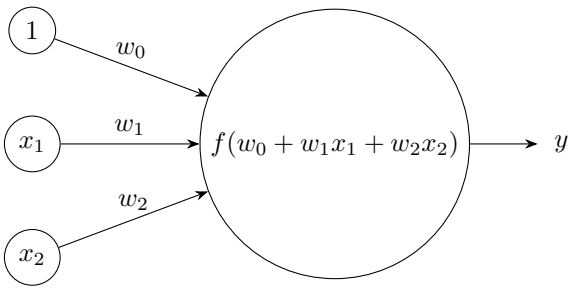
Dette afsnit er baseret på kilderne [Géron, 2019] og [Aggarwal, 2015] og her introduceres teorien bag neurale netværk.

Grundstene til de neurale netværk vi kender i dag blev lagt af neurofysiolog Warren McCulloch og matematiker Walter Pitts i 1943. De opstillede en simplificeret model for, hvordan biologiske neuroner i hjernen arbejder sammen for at udføre komplekse beregninger. Siden da har der været flere perioder, hvor interessen for neurale netværk har været stor, og modellerne er blevet ændret og udvidet mange gange igennem tiden. I dag er vi i en periode, hvor interessen for neurale netværk igen er ekstremt høj, og hvor interessen er kommet for at blive. Dette skyldes blandt andet, at vi i dag har en så stor mængde data tilgængelig til at træne vores netværk, samt at der er sket en ekstremt høj stigning i computerkræfter siden 1990'erne. Det har medført, at det i dag tager væsentligt kortere tid at træne store neurale netværk, og generelt præsterer neurale netværk bedre end andre machine learning metoder.

Ideen bag neurale netværk bygger som tidligere nævnt på det menneskelige nervesystem, hvilket består af et netværk af milliarder af neuroner, der hver især er forbundet til tusinde andre neuroner. Strukturen af biologiske neurale netværk er et emne, som der stadig aktivt bliver forsket i, men man har dog kunne fastlægge, at neuronerne er organiseret i forskellige lag.

4.3.1 Single-layer netværk

Den mest simple type af neurale netværk kaldes single-layer neurale netværk, og et eksempel på dette kan ses på Figur 4.4.



Figur 4.4. Illustration af et single-layer netværk.

Et sådan netværk består af to lag af neuroner, et inputlag og et outputlag. Dog kaldes det et single-layer netværk, eftersom inputlaget også kaldes lag 0. Inputlaget består af neuroner i form af numeriske inputværdier x_1,x_2,$\dots x_d$, hvor hver neuron har en tilhørende vægt, som forbinder neuronerne i inputlaget med neuronen i outputlaget. Inputs til det neurale netværk kommer fra $n \times d$ datamatricen X, der består af observationer, x_{i1},x_{i2},$\dots x_{id}$, for $i = 1,2,\dots n$. Generelt indgår der altid en bias neuron i inputlaget, som noteres 1, hvilket tilfører en konstant værdi til netværket. Ud fra inputneuronernes værdier og deres tilhørende vægte kan der for hver observation opskrives en linearkombination givet ved

$$z_i = w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_dx_{id} = w_0 + \sum_{j=1}^{d} w_jx_{ij}, \quad \text{for } i = 1,\dots,n.$$

Neuronen i outputlaget udfører en matematisk beregning på denne linearkombination ved hjælp af en aktiveringsfunktion. Der findes forskellige typer af aktiveringsfunktioner, som vil blive uddybet i Afsnit 4.6, men for single-layer netværk er den mest anvendte funktion den lineære fortegnsfunktion *sign*, defineret som

$$\text{sign}(z) = \begin{cases} -1 \text{ hvis } z < 0, \\ 0 \quad \text{hvis } z = 0, \\ 1 \quad \text{hvis } z > 0. \end{cases}$$

Det estimerede output af det neurale netværk vil derfor være givet ved

$$\hat{y}_i = f(z_i),$$

hvor f er den valgte aktiveringsfunktion.

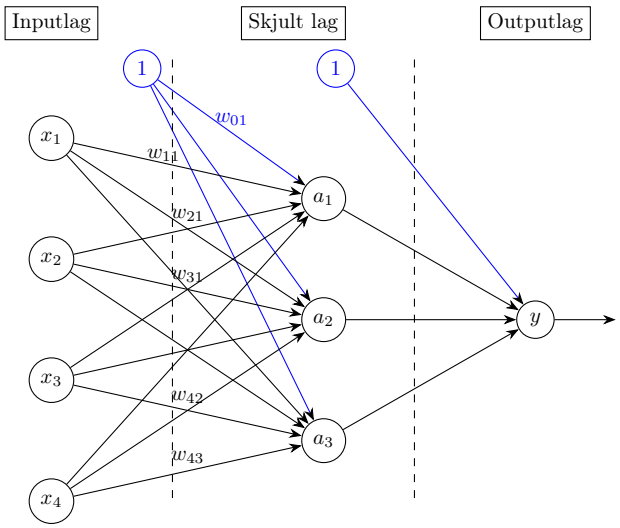
Estimeringen af et single-layer neuralt netværk består i at bestemme vægtene w_j, for $j = 0,1,2,\ldots,d$, således at der for så mange observationer i træningsmængden som muligt gælder, at $\hat{y}_i = y_i$. Algoritmen for estimering af vægtene starter med en tilfældig vægt vektor, som for hver iteration justeres. For hver iteration t sendes en tilfældig observation, x_{i1},$x_{i2},\ldots x_{in}$, igennem netværket for at beregne $\hat{y}_i$, og for hver iteration opdateres vægtene ved

$$w_j^{(t+1)} = w_j^{(t)} + \eta(y_i - \hat{y}_i)x_{ij},$$

hvilket viser, at opdateringen afhænger af fejlleddet, $(y-\hat{y})$, for den givne observation, der bliver sendt igennem netværket i den t'te iteration. Algoritmen kører gentagende gange træningsobservationerne igennem netværket og justerer vægtene, indtil der opnås konvergens. Parameteren η kaldes læringsraten, og den bestemmer, hvor meget vægtene justeres for hver iteration, og dermed hvor hurtigt der opnås konvergens. Vælges en høj værdi af η, vil det medføre, at konvergensen sker hurtigt, men dette vil ofte medføre at vægtene, og dermed den estimerede model, ikke er optimal. Hvis η vælges lavere, opnås mere optimale modeller, dog på bekostning af at konvergensen vil være langsommere. I praksis vil η starte med at have en høj værdi og herefter justeres løbende, så den bliver lavere for hver iteration.

4.3.2 Multi-layer neurale netværk

Eftersom single-layer modellen kun består af et inputlag og et outputlag, hvor der kun bliver udført én matematisk beregning, vil det ofte være en for simpel model. Mange problemer kræver mere komplekse modeller, og derfor kan mere komplekse neurale netværk opstilles, hvilke kaldes multi-layer neurale netværk. Disse netværk består af flere lag af neuroner kaldet skjulte lag. Et eksempel på et multi-layer netværk med ét skjult lag bestående af tre neuroner ses på Figur 4.5.



Figur 4.5. Illustration af et multi-layer netværk.

I et multi-layer netværk indeholder alle lag, på nær outputlaget, en bias neuron, og det er ikke længere kun i outputneuronen, at der udføres en matematisk beregning, det sker også i neuronerne i de skjulte lag. I et multi-layer feedforward netværk gælder, at informationen fra en neuron altid sendes videre i netværket, sådan at der ikke er cyklusser eller loops i netværket. Hvis det er et fuldt forbundet netværk, betyder det, at der er en forbindelse mellem alle neuroner i et lag til alle neuroner i det efterfølgende lag.

Hvis der tages udgangspunkt i netværket på Figur 4.5, vil outputtet i den k'te neuron i det skjulte lag være beregnet ved

$$a_k^{(1)} = f\Big(w_{0k}^{(1)} + \sum_{j=1}^{d^{(0)}} w_{jk}^{(1)} x_j\Big)$$

hvor f er aktiveringsfunktionen, og $d^{(0)} = 4$ angiver antallet af neuroner i lag 0 og dermed antallet af inputvariable i netværket. Superscriptet på neuronerne henviser til, hvilket lag neuronerne er i, mens superscriptet på vægtene henviser til, hvilket lag af neuroner, de er på vej ind i.

Resultaterne af beregningerne i neuronerne i det skjulte lag benyttes i beregningen af outputtet i neuronen i outputlaget. Outputtet fra Figur 4.5 beregnes derfor ved

$$\hat{y} = f\Big(w_0^{(2)} + \sum_{j=1}^{d^{(1)}} w_j^{(2)} a_j^{(1)}\Big), \tag{4.12}$$

hvor $d^{(1)} = 3$ er antallet af neuroner i det skjulte lag. For et generelt neuralt netværk med $m-1$ skjulte lag, beregnes outputtene i neuronerne i de skjulte lag og outputlaget ved

$$a_k^{(s)} = f\Big(w_{0k}^{(s)} + \sum_{j=1}^{d^{(s-1)}} w_{jk}^{(s)} a_j^{(s-1)}\Big) \qquad \text{for } s = 1, \ldots, m-1,$$

$$\hat{y} = a_1^{(m)} = f\Big(w_0^{(m)} + \sum_{j=1}^{d^{(m-1)}} w_j^{(m)} a_j^{(m-1)}\Big).$$

For at estimere netværkets vægte anvendes en algoritme kaldet Backpropagation, og denne vil blive beskrevet i følgende afsnit.

4.3.3 Backpropagation

Backpropagation er en algoritme til at estimere de optimale vægte i et multi-layer neuralt netværk. Algoritmen anvender en gradient descent metode, hvor det bestemmes, hvordan vægtene skal ændres, sådan at fejlene fra netværket reduceres. Generelt er der to faser i backpropagation, hvor den første fase kaldes forward fasen. Her beregnes en prædiktion af outputtet for observationerne i træningsmængden. I den næste fase, backward fasen, måles fejlen ud fra prædiktionen ved at betragte loss-funktionen, som sammenligner prædiktionerne med de sande værdier. Derefter gennemgås hvert lag i baglæns rækkefølge, altså fra outputlaget og tilbage mod inputlaget, for at bestemme hvor meget hver vægt i netværket har bidraget til fejlen. Dette gøres ved at anvende kædereglen på loss-funktionen for at få et udtryk for ændringen i denne. Til sidst ændres vægtene for at reducere fejlen, hvilket kaldes gradient descent steppet og udføres ved

$$\Delta w_{jk} = -\eta \frac{\partial \mathcal{L}}{\partial w_{jk}},$$

hvor $\mathcal{L}$ betegner loss-funktionen. Derefter gennemføres faserne igen med de opdaterede vægte. Valget af loss-funktion vil være forskellig i forhold til om det er et regressions- eller klassifikationsproblem. For regressionsproblemet vil den for eksempel være givet ved

$$\mathcal{L} = \frac{1}{2n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$

Når et neuralt netværk estimeres, kan der som tidligere nævnt anvendes forskellige aktiveringsfunktioner, hvor nogle af disse vil blive introduceret i næste afsnit.

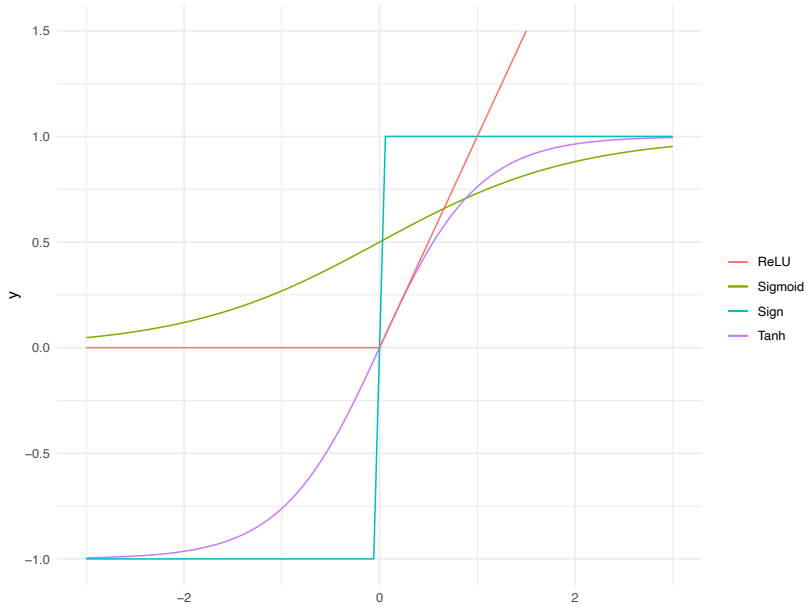
4.3.4 Aktiveringsfunktioner

Aktiveringsfunktionerne der anvendes i et multi-layer feedforward netværk er ikke restringeret til at være lineære fortegnsfunktioner, som det var tilfældet ved single-layer netværk. I stedet kan andre aktiveringsfunktioner anvendes, og der kan være forskellige funktioner i neuronerne i de skjulte lag og i outpulaget Nogle af de mest anvendte aktiveringsfunktioner i multi-layer neurale netværk er Sigmoid, noteret σ, Hyperbolsk Tangens (TanH) og Rectified Linear Unit function (ReLU). Sigmoid og TanH funktionerne tager reelle tal som input, og har outputs som henholdsvis ligger i intervallet (0,1) og (−1,1), og de er defineret som følgende

$$\sigma(x) = \frac{1}{(1+\exp(-x))}, \qquad \text{TanH}(x) = 2\sigma(2x) - 1 \tag{4.13}$$

ReLU funktionen tager et reelt tal som input og har et output, der enten er 0 eller lig inputværdien, afhængig af størrelsen på inputværdien:

$$\text{ReLU}(x) = \begin{cases} x \text{ hvis } x \geq 0 \\ 0 \text{ hvis } x < 0. \end{cases}$$



Figur 4.6. Aktiveringsfunktionerne Sign, Sigmoid, TanH og ReLU.

Figur 4.6 illustrerer opførslen af de fire aktiveringsfunktioner Sign, Sigmoid, TanH og ReLU.

Det er en fordel beregningsmæssigt, hvis aktiveringsfunktionerne er differentiable, da det er nødvendigt i backpropagation algoritmen. Både Sigmoid og Tahn er differentiable funktioner, men det er Sign og ReLU ikke, og det er derfor nødvendigt for disse to funktioner at opstille antagelser for de afledte.

Modellering af VIX indekset 5

Vi ønsker at undersøge om klimaforandringerne påvirker volatiliteten af S&P500 aktieindekset, hvilket gøres ved at opstille modeller for henholdsvis VIX indekset og den realiserede varians for S&P500. Modellerne opstilles på baggrund af teorien fra Kapitel 3 og 4. I dette kapitel vises resultaterne for modelleringen af VIX indekset, mens de tilsvarende resultater for modelleringen af den realiserede varians kan findes i Kapitel 6. Hvert kapitel afsluttes med en opsummering af resultaterne for alle de opstillede modeller.

5.1 Databehandling

Til at opstille modeller, der undersøger om klimaforandringer har en effekt på VIX indekset, anvendes variable i form af Google Trends data hentet fra [Google-Trends]. Klimaforandringerne påvirker blandt andet jorden i form af hyppigere forekomst af naturkatastrofer, og vi vil derfor som variable anvende antallet af søgninger på ord, der relateres til den globale opvarmning samt forskellige typer af naturkatastrofer. Vi har udvalgt de følgende 22 søgeord som variable:

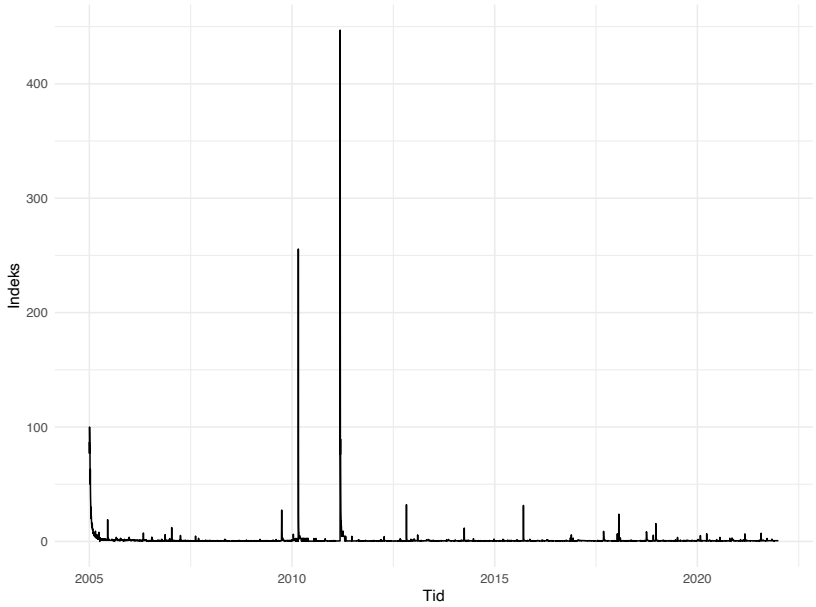
- Avalanche
- Blizzard
- Climate Change
- Drought
- Dust Storm
- Earthquake
- Flash Flood
- Flood
- Global Warming
- Hailstorm
- Heat Wave
- Landslide
- Mudslide
- Pandemic
- Sea Level Rise
- Sinkhole
- Thunderstorm
- Tornado
- Tropical Cyclone
- Tsunami
- Volcanic Eruption
- Wildfire

Vi har udelukkende anvendt søgninger i USA, da vi benytter volatilitetsindekser for S&P500, som består af de 500 største virksomheder i USA. Søgningerne er baseret på daglige observationer fra den 1. januar 2005 til den 31. december 2021. Det skal bemærkes, at vi har medtaget variablen *Pandemic*, selvom det ikke er en direkte naturkatastrofe. Den er dog medtaget, fordi mange af klimaforandringernes konsekvenser øger risikoen for pandemier, [Chan].

Google Trends datasættet er i `R` hentet dagligt med pakken `gtrendsR`, [Massicotte og Eddelbuettel, 2022], hvilket kun er muligt over en periode på otte måneder, inden Google automatiske ændrer observationerne til at være givet ugentligt eller månedligt. Derfor har vi for hver type naturkatastrofe hentet data for syv måneder af gangen, hvor den sidste måned i hvert interval overlapper det næste interval. Dette giver 34 tidsintervaller over perioden, som hver har forskellig standardisering. Dette tages højde for ved at sammenligne observationerne i de overlappende intervaller for at kunne beregne en faktor, som beskriver størrelsesforskellen mellem de to intervaller. I dette overlappende interval

er antal søgninger hver dag i første interval derfor divideret med antallet af søgninger i det næste, hvorefter vi har taget gennemsnittet for at opnå den faktor, som hele det næste interval ganges igennem med. På denne måde opnås et samlet dagligt datasæt målt i samme skala. Denne metode kræver, at det overlappende interval ikke udelukkende indeholder observationer med 0 søgninger, da det dermed ikke vil være muligt at finde en faktor at skalere intervallet med. Vi har taget højde for dette ved at vælge det overlappende interval til at være en måned, således det er et stort nok interval til altid at indeholde observationer, som ikke udelukkende er 0. I Google Trends datasættet er der indgange, som er angivet < 1, og vi har derfor været nødsaget til at indsætte en værdi, hvor vi har valgt værdien 0,1.

Figur 5.1 viser indekset for de daglige Google søgninger på *Tsunami* i perioden, hvor Figur 2.2 på Side 5 viste indekset for de månedlige observationer. Den daglige tidsrække er opnået ved at sammensætte 34 daglige intervaller ved brug af metoden, som blev beskrevet ovenfor. Vi har brugt tilsvarende metode for at opnå tidsrækkerne for de 21 resterende Google Trends variable, og plots af disse kan ses i Appendiks A på Side 85.



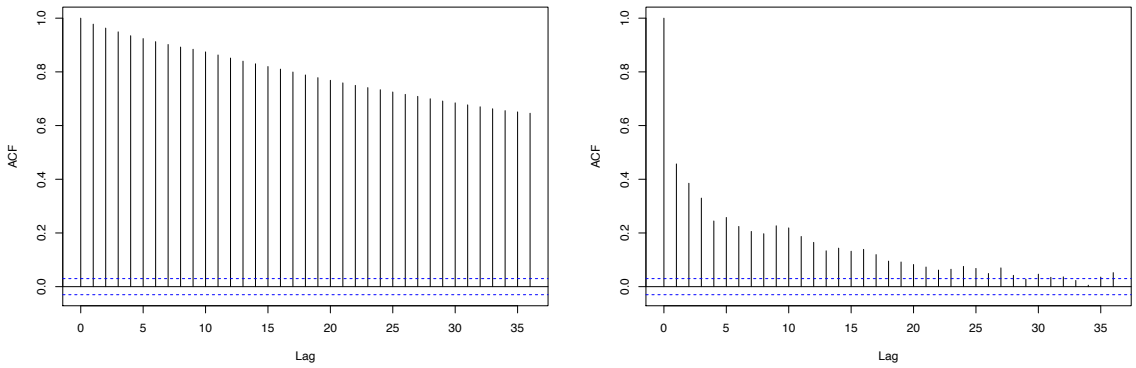
Figur 5.1. Daglig Google Trends data for *Tsunami*.

Det tjekkes, om alle 22 Google Trends tidsrækker er stationære ved at undersøge, om vi kan forkaste nulhypotesen om en enhedsrod. Dette gøres eftersom, at det kan give misvisende og spuriøse resultater, hvis tidsrækkerne ikke er stationære. Det testes i `R` med pakken `urca` [Pfaff et al., 2016] samt kommandoen `ur.df`, som anvender augmented Dickey-Fuller til at teste for en enhedsrod i tidsrækken. Resultatet er, at ingen af tidsrækkerne inderholder en enhedsrod ifølge testen, hvormed de alle er stationære.

VIX indekset er hentet fra Yahoo-finance, [Yahoo-Finance, 2022], over samme periode som de 22 Google Trends variable, og indekset kan ses på Figur 2.3 på Side 7 På samme vis som ved Google Trends datasættet undersøges, om VIX indekset er stationært ved at teste hypotesen om en enhedsrod. Denne hypotese forkastes, hvormed denne tidsrække også er stationær.

Figur 5.2a viser autokorrelationsfunktionenen for VIX indekset. Her ses, at tidsrækken har en tendens til long-memory, som blev beskrevet i Afsnit 3.1 eftersom autokorrelationensfunktionen aftager med en hyperbolsk rate. Ved at beregne Hurst koefficienten fra (3.2) på Side 12, opnås en værdi på 0,82, hvilket også indikerer, at tidsrækkens autokorrelation er persistent og udviser evidens for long memory. Da der er evidens for long memory, fraktionelt differenses tidsrækken, og for at gøre dette, skal d estimeres. Dette gøres i `R` med funktionen `arfima` fra pakken `arfima`, [Veenstra, 2022], hvilken bruger optimeringsalgoritmen BFGS til at estimere d. Værdien af d estimeres til 0,499, og ACF'en for den fraktionelt differenset tidsrække ses på Figur 5.2b. Her ses, at autokorrelationen er mindsket markant, så tidsrækken ikke længere udviser long memory. Det er derfor denne fraktionelt differensede tidsrække

for VIX indekset, som vil blive benyttet fremadrettet i analysen, hvilket vil sige, at når der fremover refereres til VIX indekset, menes der denne fraktionelt differensede tidsrække.



(a) ACF for VIX indeks. **(b)** ACF for fraktionelt differenset VIX indeks.

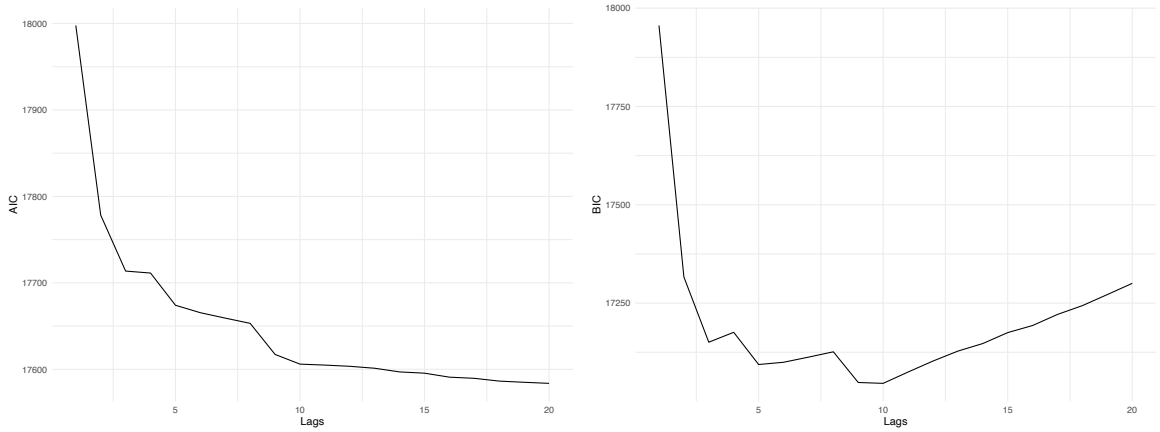
Figur 5.2

Eftersom datasættet med VIX indekset ikke indeholder observationer i weekender og på helligdage, så fjernes de tilsvarende observationer fra Google Trends datasættet, hvormed det består af 4280 observationer.

5.2 Autoregressive modeller for VIX

I dette afsnit opstilles autoregressive modeller for VIX indekset, hvilke sammenlignes for at finde den mest optimale model.

Med kommandoen `dynlm` fra pakken `dynlm`, [Zeileis, 2019] i `R` opstilles AR-modeller med forskellig orden for VIX indekset, hvorefter informationskriterierne AIC og BIC benyttes til at finde den optimale lagorden p. Informationskriterierne anvender den negative loglikelihoodfunktion og tillægger en straf for antal parametre i modellen, hvormed en lav værdi indikerer en optimal model. Forskellen i AIC og BIC ligger i måden, hvorpå de straffer for antallet af parametre, hvor BIC straffer hårdest. AIC er vist på Figur 5.3a og BIC på Figur 5.3b.



(a) AIC mod antal lags for en AR-model for VIX. **(b)** BIC mod antal lags for en AR-model for VIX.

Figur 5.3

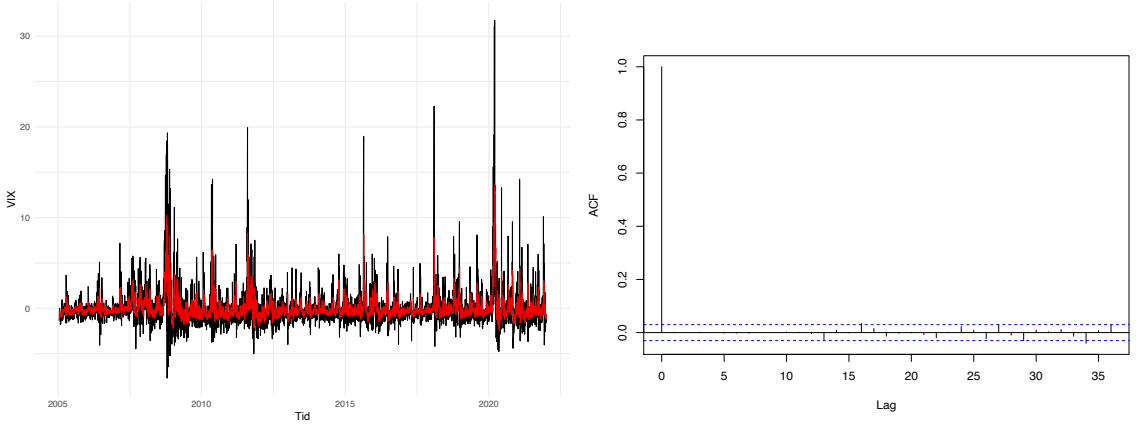
Fra Figur 5.3a ses, at AIC-værdien falder med antallet lags, der tilføjes til modellen, og derfor ville man fra AIC-værdierne vælge $p = 20$. Figur 5.3b viser, at BIC-værdierne er lavest ved lag 10, hvilket

svarer til, at modellen anvender laggede værdier fra de to foregående uger, eftersom weekender ikke indgår i datasættet. Af denne grund vælges den optimale orden til $p = 10$, da dette i forhold til BIC er det mest optimale valg. Vi opnår derfor en AR(10)-model, og koefficienterne for denne kan ses i Tabel 5.4.

	Koefficienter	Standardafvigelse	t-værdi	p-værdi
Skæring	0,002	0,029	0,081	0,9351
AR 1	0,310	0,015	20,302	2,2e-16*
AR 2	0,161	0,016	10,050	2,2e-16*
AR 3	0,091	0,016	5,636	1,8e-08*
AR 4	-0,031	0,016	-1,882	0,0599
AR 5	0,068	0,016	4,163	3,2e-05*
AR 6	0,018	0,016	1,119	0,2634
AR 7	0,005	0,016	0,278	0,7814
AR 8	-0,003	0,016	-0,184	0,8544
AR 9	0,075	0,016	4,672	3,1e-06*
AR 10	0,049	0,015	3,173	0,0015*

Tabel 5.4. Koefficienter samt p-værdier for AR(10)-modellen. P-værdier under 0,05 er markeret med *.

Den justerede R^2-værdi for AR(10)-modellen er 0,274, hvilket vil sige, at 27,4% af variationen i VIX indekset kan forklares ved modellen. På Figur 5.5a er de fittede værdier for VIX indekset vist som den røde kurve, mens den sorte kurve viser de sande værdier. Her ses, at modellen opfanger en del af udsvingene i VIX indekset, dog ikke størrelsen af disse, hvilket også afspejles i den justerede R^2-værdi for modellen. Figur 5.5b viser ACF'en for AR(10)-modellens residualer, hvilken viser, at de som ønsket er tilnærmelsesvis hvid støj, eftersom der ingen signifikant korrelation er.



(a) Fittede værdier for AR(10)-modellen for VIX indekset.

(b) ACF for residualerne fra AR(10)-modellen for VIX indekset.

Figur 5.5

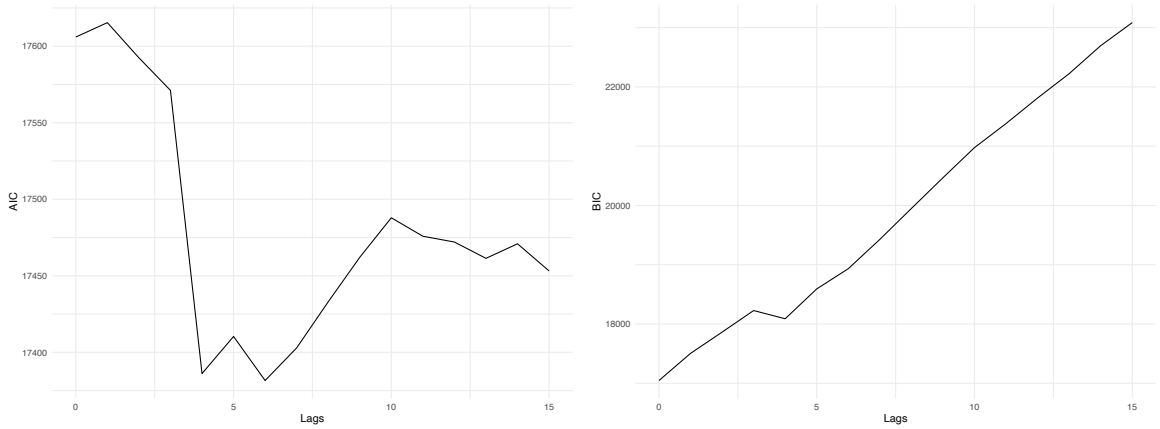
Opstillingen af AR-modellen er foretaget i to steps, hvor d først blev estimeret for derefter at estimere AR-modellen. Som alternativ kan d og koefficienterne estimeres i ét step, hvilket kan gøres i `R` med kommandoen `arfima` fra pakken `forecast`, [Hyndman, 2022], som anvender MLE til estimeringen, hvilket blev beskrevet i Afsnit 3.1.1. Med denne fremgangsmåde sammenligner vi AR-modeller med forskellig orden ved brug af AIC, og vi får resultatet, at værdien af d stiger med antal lags, og værdien af AIC falder med antal lags. Den mest optimale model er derfor en AR(10), som har $d = 0{,}4432$

og AIC=17651. Denne AIC-værdi er dog højere end den for AR(10)-modellen, som blev fittet i to steps, hvilken er 17605, og derfor fortsættes analysen med den hidtil gennemgåede model, som har koefficienterne angivet i Tabel 5.4.

5.2.1 Autoregressiv distributed lag model for VIX

For at undersøge om Google Trends variablene og dermed klimaforandringerne kan forbedre forklaringsgraden udvides AR(10)-modellen til en ADL-model, som blev beskrevet i Definition 3.1, således disse variable indgår i AR-modellen som eksterne regressorer. Inden modelleringen er Google Trends variablene skaleret, sådan at de ligger i intervallet [0,1].

Parameteren q i ADL-modellen findes på baggrund af informationskriterierne AIC og BIC som tidligere, hvor Figur 5.6a og 5.6b viser henholdsvis AIC- og BIC-værdierne plottet mod antallet af lags af de eksterne regressorer. Af AIC-værdierne ses, at de laveste værdier er for $q = 4$ samt $q = 6$. Af BIC-værdierne ses, at værdierne stiger i takt med antal lags, hvormed $q = 0$ er at foretrække. Dog ses et fald i BIC-værdien ved $q = 4$, som også har en lav AIC-værdi, hvorfor $q = 4$ vælges, og dermed inddrages Google Trends variablene i modellen op til lag 4.



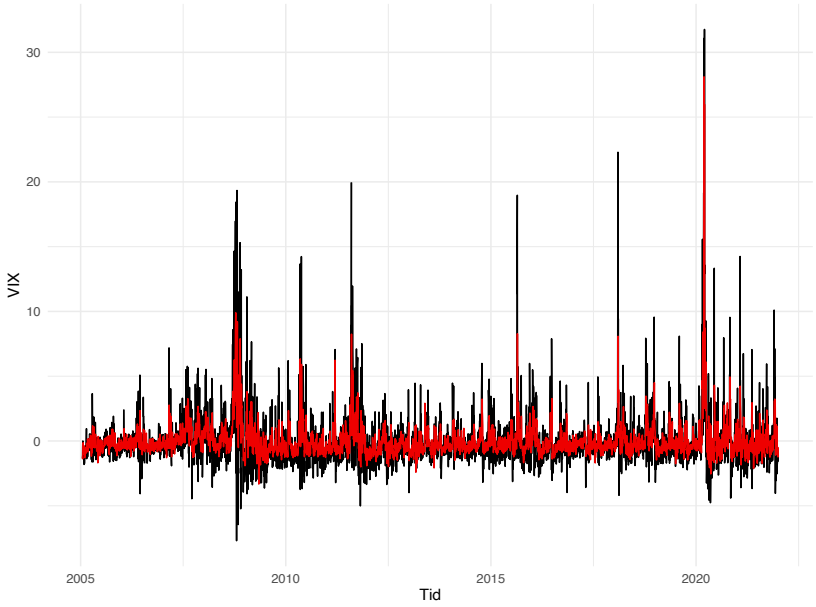
(a) AIC mod antal lags af Google Trends variablene for en ADL-model.

(b) BIC mod antal lags af Google Trends variablene for en ADL-model.

Figur 5.6

Tabel B.1 samt Tabel B.2 i Appendiks B viser koefficienterne, standardafvigelserne, t-værdierne samt p-værdierne for ADL-modellen, hvor tallet efter Google Trends variablenes navne refererer til lagget, altså svarer Wildfire 1 til variablen til tiden $t - 1$. Af p-værdierne i tabellen ses, at størstedelen af parametrene er insignifikante. De eneste af de 22 Google Trends variable, som er signifikante, er *Pandemic* til alle fire lags, *Sinkhole* til lag 1 og 4, *Tornado* til lag 2, *Tsunami* til lag 3 og 4, *Landslide* til lag 2 og *Heat Wave* til lag 4. At *Pandemic* er signifikant til alle lags, tyder på, at denne parameter har en særlig påvirkning på VIX indekset. Dette kan også ses af tabellen, eftersom koefficienterne for *Pandemic* har væsentlig højere absolutværdi end de andre signifikante parametre.

På Figur 5.7 ses et plot af de fittede værdier for ADL(10,4)-modellen samt de sande værdier for VIX. Her ses en smule mere variation i de fittede værdier i forhold til dem fra AR-modellen på Figur 5.5a. Desuden ses på Figur 5.7, at modellen er bedre til at opfange udsvinget i VIX indekset i starten af 2020, som skyldes coronapandemien. Dette udsving er formentlig opfanget bedre af modellen, fordi parameteren *Pandemic* indgår i modellen og er signifikant til alle fire lags, og desuden har *Pandemic* som før nævnt relativt store koefficienter i forhold til de andre variable, hvormed den har større indflydelse. Den justerede R^2-værdi for ADL-modellen er 0,324, hvilket er en forbedring i forhold til AR-modellen. Det betyder, at ADL-modellen forklarer en større andel af variationen i VIX indekset i forhold til AR-modellen, hvormed Google Trends variablene har bidraget positivt til fittet.

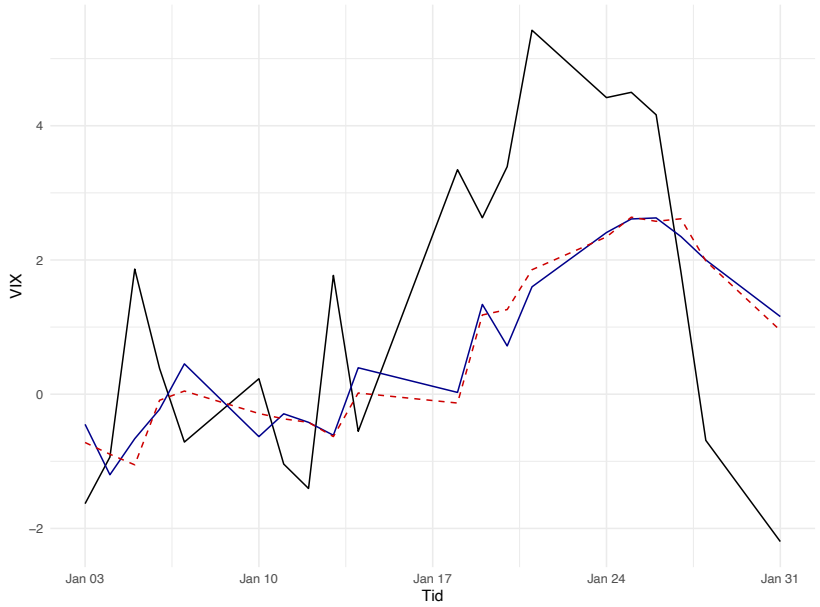


Figur 5.7. Fittede værdier for ADL-modellen for VIX indekset.

Vi undersøger, om det vil bidrage til modellen også at inddrage de nutidige værdier af de 22 Google Trends variable. Dette gøres særligt for at undersøge sammenhængen mellem VIX og Google Trends variablene, og om de nutidige værdier kan bidrage til at forklare indekset. Modellen, som indeholder de nutidige værdier af hver Google Trends variabel, samt værdier fra lag 1 til 4, har en justeret R^2-værdi på 0,325, hvilket er minimalt højere end modellen uden nutidige værdier. Derudover er AIC-værdien højere for modellen med den nutidige værdi end ADL-modellen uden, hvormed det vurderes, at det ikke bidrager signifikant at tilføje de nutidige værdier til modellen. Selvom modellen med de nutidige værdier er marginalt bedre in-sample, er den ikke overordnet at foretrække. Dette skyldes, at modellen vil indebære større usikkerhed out-of-sample, eftersom at de nutidige værdier af Google Trends variablene er ukendte, og dermed først skal prædikteres. På denne måde vil prædiktionerne af modellen afhænge af Google Trends prædiktioner, hvilket med stor sandsynlighed vil påvirke præcisionen. Af denne grund fortsættes analysen med ADL-modellen uden de nutidige værdier, og i næste afsnit anvendes modellen til at prædiktere VIX indekset.

5.2.2 Prædiktioner

I dette afsnit vurderes modellernes præcision out-of-sample ved at prædiktere nye værdier og sammenligne størrelsen af prædiktionsfejlene. På Figur 5.8 ses one-day-ahead forecasts af VIX indekset med AR(10)-modellen samt ADL(10,4)-modellen, hvilke er vist som henholdsvis den røde og blå kurve. Den sorte kurve viser de sande værdier af VIX indekset. Prædiktionerne er lavet for alle hverdage i januar 2022, og er beregnet som rullende forecast, hvilket vil sige, at hvert forecast er baseret på de foregående sande værdier, som løbende inddrages.



Figur 5.8. Sande værdier for VIX (sort), prædikterede værdier for VIX med AR(10) (rød) samt ADL(10,4) (blå).

Baseret på de prædikterede værdier kan *mean square error* (MSE) og *mean absolute error* (MAE) beregnes for at sammenligne præcisionen af de to modellers prædiktioner, og de beregnes ved

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n},$$
$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}.$$

Forskellen i de to statistikker ligger i, at enkelte store fejl kan have en stor indvirkning på MSE, fordi de kvadreres, mens de ikke vil have så stor en påvirkning på MAE. Desuden er det en fordel med MAE, at fejlene angives i samme enhed som prædiktionerne, så de kan sammenlignes direkte. For AR(10)-modellen er MSE-værdien 3,867 mens den for ADL(10,4) er 4,078. Desuden er MAE-værdien for AR(10)-modellen lig 1,651, mens den er 1,739 for ADL(10,4)-modellen. Det betyder, at AR-modellen er mest præcis i beskrivelsen af VIX indekset out-of-sample på trods af, at der ikke er stor forskel. Det tyder på, at det ikke har bidraget til ADL-modellens prædiktionsevne at tilføje Google Trends variablene til modellen.

Som før nævnt viste Tabel B.1 samt Tabel B.2 i Appendiks B, at størstedelen af Google Trends variablene er insignifikante, og dermed ikke bidrager til beskrivelsen af volatiliteten i markedet. Derfor er det interessant at undersøge, om variablene kan bidrage ved at anvende shrinkage eller dimensions reduktion metoder, som blev beskrevet i Kapitel 4.2.2. Det er interessant at undersøge, fordi nogle variable muligvis fejlagtigt bliver ekskluderet fra modellerne, fordi der er korrelation mellem flere af variablene. I næste afsnit vil korrelationen i Google Trends datasættet derfor blive undersøgt.

5.2.3 Korrelation mellem Google Trends variable

Variansen af koefficienterne ved en OLS regression, som benyttes under estimeringen af de autoregressive modeller, er givet ved

$$\text{Var}(\beta) = \sigma^2(X^T X)^{-1}. \tag{5.1}$$

Denne varians stiger, hvis flere af variablene er korrelerede. For at vise dette, så antag, at vi er interesserede i en bestemt koefficient, β_1. Så hvis β_2 betegner $(d-1)$-vektoren af de resterende koefficienter, kan OLS regressionsmodellen opskrive som

$$y = x_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{5.2}$$

hvor designmatricen X er opdelt i vektoren x_1 og matricen X_2 i overensstemmelse med opdelingen af β-vektoren. Som resultat af Frisch-Waugh-Lovell sætningen, som kan ses i Appendiks C, giver regressionen i (5.2) det samme estimat af β_1 som FWL regressionen, hvilken er

$$M_2 y = M_2 x_1 \beta_1, \tag{5.3}$$

hvor $M_2 \equiv I - X_2(X_2^T X_2)^{-1} X_2^T$. Estimatet for β_1 er givet som

$$\hat{\beta}_1 = \frac{x_1^T M_2 y}{x_1^T M_2 x_1},$$

med tilhørende varians

$$\text{Var}(\hat{\beta}_1) = \sigma_0^2 (x_1^T M_2 x_1)^{-1} = \frac{\sigma_0^2}{x_1^T M_2 x_1}. \tag{5.4}$$

Her er σ_0^2 variansen af fejlleddene, og desuden er $x_1^T M_2 x_1 = ||M_2 x_1||$, eftersom M_2 er en ortogonal projektionsmatrix og derfor idempotent. På denne måde kan det ses, at mængden af information datasættet giver om β_1 er proportional med den kvadrerede Euklidiske længde af vektoren $M_2 x_1$. Den kvadrerede Euklidiske længde af vektoren $M_2 x_1$ er summen af de kvadrerede residualer fra regressionen

$$x_1 = X_2 c + \varepsilon. \tag{5.5}$$

Det betyder, at når $||M_2 x_1||$ er lille, så er x_1 godt forklaret givet de resterende søjler i X, og dermed er SSR lille. Det gør, at variansen af $\hat{\beta}_1$ er tilsvarende stor. Modsat, når x_1 ikke forklares godt ved brug af de resterende søjler i X, så vil SSR være stor og variansen af $\hat{\beta}_1$ tilsvarende lille. På denne måde er variansen af $\hat{\beta}_1$ i (5.4) proportional til den inverse af summen af de kvadrerede residualer fra regressionen i (5.5). Det vil sige at præcisionen af $\hat{\beta}_1$ og dermed variansen afhænger ligeså meget af X_2 som af x_1. Det vil derfor kunne forstyrre et ellers præcist estimat af β_1 at tilføje ydereligere regressorer. Grunden til dette er, at hvis de ydereligere regressorer forklarer x_1 i (5.5) bedre end udgangspunktet, så bliver SSR mindre og variansen tilsvarende større. Denne situation kaldes kolinaritet, og betyder at regressorerne er lineært afhængige, [Davison og MacKinnon., 2009]. Der kan også være kolinaritet mellem tre eller flere variable, hvilket kaldes multikolinaritet, og dette kan ikke ses direkte i kovariansmatricen. Risikoen for multikollinearietet er større, jo flere variable der inddrages. En større varians vil have indflydelse på t-værdien og dermed indflydelse på, om koefficienten er signifikant. Ved brug af den j'te koefficient og dens varians, kan t-værdien findes ved

$$t_{\beta_j} = \frac{\beta_j}{(\text{Var}(\beta_j))^{1/2}}.$$

Jo større variansen af koefficienten bliver, jo mindre bliver t-værdien. En høj varians medfører derfor, at der er større sandsynlighed for, at værdien ikke overskrider den kritiske værdi og at koefficienten derfor betragtes som insignifikant. Dette fremgår også af Tabel B.1 samt Tabel B.2 i Appendiks B, hvor det kan ses, at alle insignifikante koefficienter er dem, hvor absolutværdien af t-værdien er mindre end den kritiske værdi på omkring 2.

En måde hvorpå det kan bestemmes, om der er multikollinaritet er ved at beregne varians inflation fakoren (VIF). Den kan for hver variabel beregnes som

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

hvor $R^2_{X_j|X_{-j}}$ er R^2-værdien fra en regression af X_j på alle andre prædiktorer. Den mindst mulige værdi af VIF er 1, hvilket indikerer, at der ikke er kollineraitet mellem variablene. En VIF værdi på over 5 indikerer høj kollinearitet, og hvis værdien overstiger 10 indikerer det, at der er en problematisk mængde korrelation mellem variablene, [James et al., 2017]. I Tabel 5.9 ses VIF-værdierne for de 22 Google Trends variable, hvoraf det ses, at alle værdierne er mindre end 5. Det betyder, at Google Trends datasættet baseret på VIF-værdierne ikke indeholder en problematisk korrelation mellem variablene, hvilket betyder at der ikke er stor sammenhæng mellem søgningerne. Det vil sige, at hvis befolkningen har en stor bekymring for eksempelvis *Global Warming*, hvilket kan ses som et stort udsving i denne variabel, kæder de det ikke sammen med andre naturkatastrofer og søger derfor ikke også på for eksempel *Heat Wave*. Korrelationsmatricen for Google Trends variablene er derudover undersøgt, og den viser i overensstemmelse med VIF resultaterne, at korrelationen generelt ligger omkring 0 for de fleste variable, med undtagelse af enkelte variable hvor korrelationsværdien ligger på omkring 30. Dog ses to variable med høj korrelation, hvilke er *Climate Change* og *Global Warming*. Disse to variable har en korrelation på 0,68, hvilket betyder, at søgninger på *Climate Change* ofte foretages i forbindelse med søgninger på *Global Warming* og omvendt.

	VIF		VIF		VIF
Avalanche	1,084	Global Warming	1,205	Sinkhole	1,031
Blizzard	1,008	Hailstorm	1,025	Thunderstorm	1,210
Climate Change	1,634	Heat Wave	1,134	Tornado	1,045
Drought	1,101	Landslide	1,249	Tropical Cyclone	1,112
Dust Storm	1,010	Mudslide	1,219	Tsunami	1,042
Earthquake	1,093	Pandemic	1,095	Volcanic Eruption	1,090
Flash Flood	1,334	Sea Level Rise	1,538	Wildfire	1,032
Flood	1,293				

Tabel 5.9. VIF-værdier for de 22 Google Trends variable.

På trods af at alle VIF-værdierne ikke indikerer en problematisk korrelation mellem variablene, vil vi stadig anvende shrinkage og dimensions reduktion metoderne fra Kapitel 4 på ADL-modellen for at undersøge, om resultaterne forbedres.

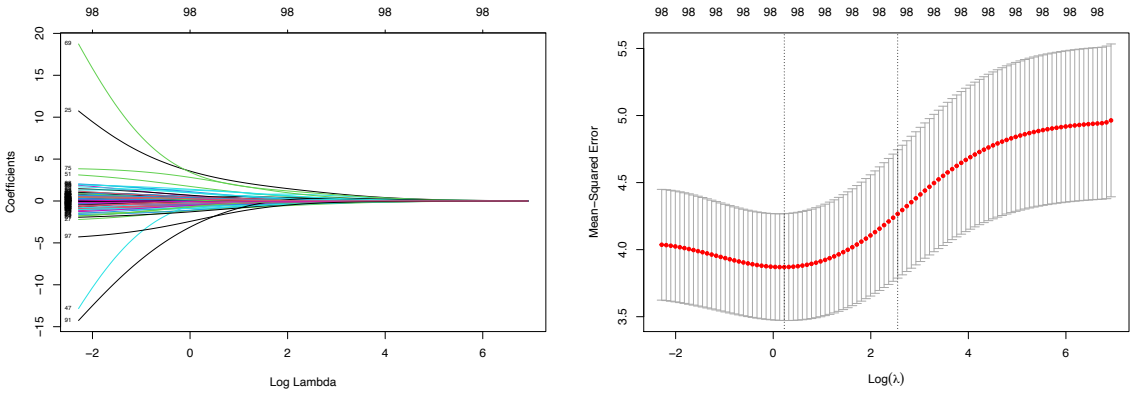
5.3 Shrinkage

På baggrund af teorien i Afsnit 4.1 udføres nu shrinkage på ADL-modellen fra Afsnit 5.2 for at mindske påvirkningen fra irrelevante variable. Dette gøres ved at krympe koefficienterne mod 0 med henblik på at mindske variansen af koefficienterne i modellerne.

5.3.1 Ridge regression for VIX

I dette afsnit udføres Ridge regression på ADL-modellen for VIX indekset fra Afsnit 5.2 ved brug af `R`-pakken `glmnet`, [Friedman et al., 2021]. Figur 5.10a viser et plot af koefficienterne mod størrelsen af logaritmen til λ, som er tuningparameteren, der bestemmer størrelsen af straffen. De fleste af koefficienterne er allerede små fra start, med undtagelse af de fire koefficienter for *Pandemic*, som har absolutværdier over 5. Når størrelsen af straffen stiger, ved at λ øges, ses det, at alle koefficienterne tvinges mod 0. Figur 5.10b viser MSE-værdier mod logaritmen til λ for de tilhørende modeller. Her markerer den første vertikale stiplede linje det λ, som giver den mindste fejl, hvilket i dette tilfælde er $\lambda = 1{,}254$.

Med dette valg af tuningparameter opnås en Ridge regression model med en justeret R^2-værdi på 0,264, hvilket er lavere end værdierne for de tidligere modeller. Dette er også forventeligt, eftersom vi har tilladt højere bias for at reducere variansen.



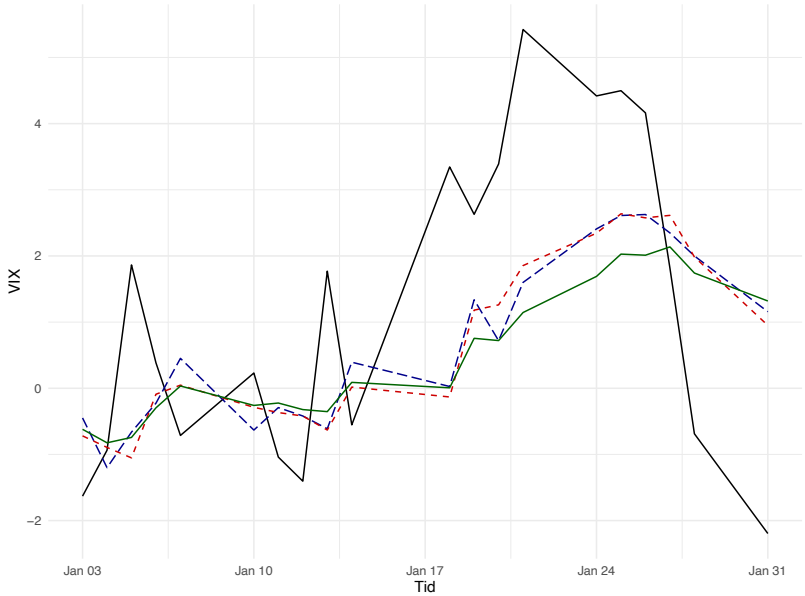
(a) Ridge regressions koefficienter mod log(λ). **(b)** MSE mod logaritmen af λ.

Figur 5.10

Koefficienterne, standardafvigelser, t-værdier og p-værdier for denne model kan ses i Tabel B.3 i Appendiks B. Her er standardafvigelserne af koefficienterne beregnet på baggrund af (4.3). Af tabellen ses, at de eneste Google Trends variable med signifikante koefficienterne er *Pandemic* til lag 1 og 3 samt *Tsunami* til lag 3. Disse tre koefficienter var også signifikante i forbindelse med ADL-modellen, hvis koefficienter er vist i Tabel B.1 samt B.2 i Appendiks B. Det betyder, at det også i Rigde regressionen er disse parametre, som har indflydelse på VIX indekset.

Med F-test undersøges, om de ekstra parametre i Ridge regressionen bidrager til forklaringen af VIX indekset, eller om AR(10)-modellen uden eksterne regressorer er bedre. Heraf fås en F-værdi på 0,283, hvormed denne er lavere end den kritiske værdi, så nulhypotesen ikke forkastes. Det vil sige, at de ekstra parametre i form af Google Trends variable i Ridge regressionen ikke bidrager til fittet in-sample af VIX indekset. Dette kan skyldes, at der i Ridge regression introduceres bias i koefficienterne for at forbedre prædiktionerne.

For at vurdere modellens præcisionen out-of-sample anvendes den nu til at prædiktere VIX indekset one-day-ahead for hverdage i januar 2022, hvilket er vist på Figur 5.11. Her viser den grønne kurve prædiktionerne med Ridge regressionen, mens den sorte kurve viser de sande værdier. Den røde og blå kurve er de tidligere forecastede værdier med henholdsvis AR(10) og ADL(10,4)-modellen.

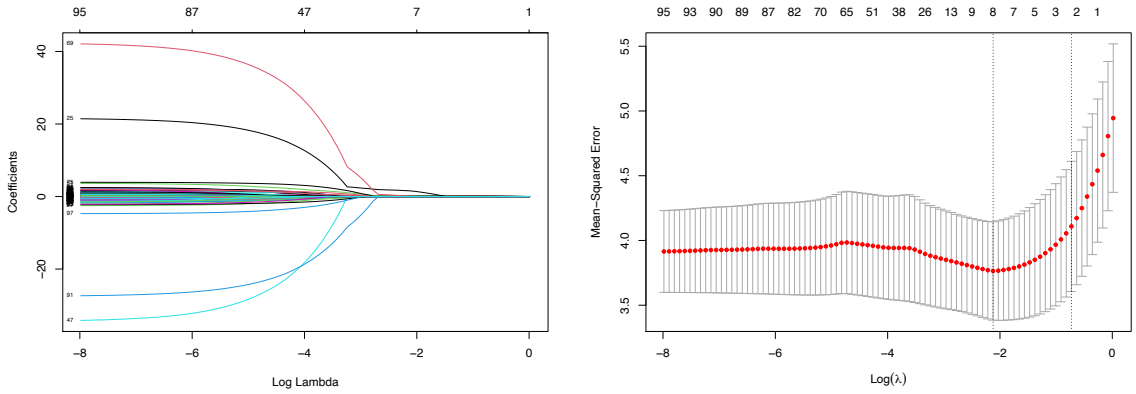


Figur 5.11. Sande værdier for VIX (sort), prædikterede værdier for VIX med AR(10) (rød), med ADL(10,4) (blå), samt med Ridge regression på ADL(10,4) (grøn).

Prædiktionerne vurderes ved fejlen, hvor MSE-værdien er 4,623 og MAE-værdien er 1,804, hvilke begge er højere end de tilsvarende værdier for de tidligere modeller. Det betyder, at Rigde regression hverken præsterer bedre in-sample eller out-of-sample end de tidligere autoregressive modeller.

5.3.2 Lasso for VIX

I dette afsnit udføres Lasso på ADL-modellen for VIX indekset fra Afsnit 5.2. Tilsvarende Ridge regression viser Figur 5.12a koefficienterne for Lasso mod logaritmen til λ. Det ses, at koefficienterne også her bliver tvunget mod 0 jo større straffen bliver, og nogle bliver præcis 0. Dette kan ses af talrækken øverst på figuren, som angiver antallet af tilbageværende parametre i modellen givet størrelsen af λ.



(a) Lasso koefficienter mod logaritmen til λ.

(b) MSE mod logaritmen til λ.

Figur 5.12

Figur 5.12b viser MSE-værdierne plottet mod logaritmen til λ, og den værdi af λ, som giver den mindste fejl, er markeret ved den første stiplede linje. Denne værdi af λ er lig 0,119, og med dette valg af tuningparameter fås en Lasso model med en justeret R^2-værdi på 0,267. Af talrækken øverst på figuren ses, at Lasso ønsker at beholde otte parametre i modellen, og koefficienterne for disse er vist i Tabel 5.13. Det ses af tabellen, at modellen kun vil beholde én Google Trends variabel, hvilken er *Pandemic* til lag 1 og 3. Dette viser ligesom de tidligere modeller, at *Pandemic* bidrager i forhold til at beskrive VIX indekset, dog viser p-værdien, at *Pandemic 3* ikke er signifikant.

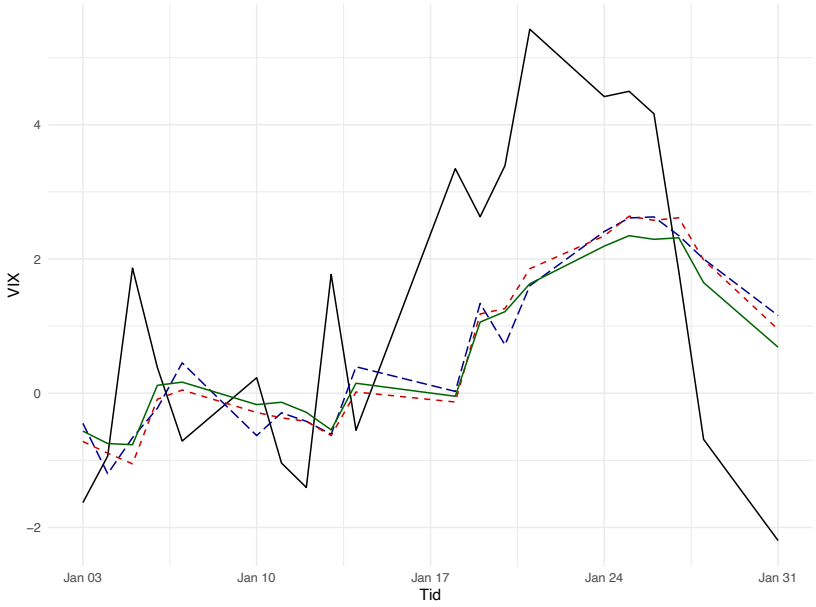
	Koefficienter	Standardafvigelse	t-værdi	p-værdi
Skæring	-0,014	0,002	-6,385	1,000
AR 1	0,284	0,002	115,814	5,91e-107*
AR 2	0,138	0,004	34,424	7,72e-57*
AR 3	0,067	0,002	27,630	2,37e-48*
AR 5	0,040	0,003	14,962	2,38e-27*
AR 9	0,050	0,003	16,930	3,71e-31*
AR 10	0,021	0,002	10,578	3,39e-18*
Pandemic 1	1,570	0,119	13,179	9,99e-24*
Pandemic 3	0,050	0,055	0,898	0,186

Tabel 5.13. Lasso koefficienter for VIX. P-værdier over 0,05 er markeret med *.

Standardafvigelserne for Lasso koefficienter kan ikke beregnes direkte, da koefficientestimaterne ikke kan udtrykkes på lukket form. Derfor har vi anvendt bootstrap metoden til at beregne standard-

afvigelserne. Dette har vi gjort ved først at opdele datasættet i 284 blokke, som hver består af 15 observationer. Herefter opstilles 284 modeller på datasættet, hvor der for hver model er undladt en blok, hvilket giver 284 koefficientestimater for hver variabel i modellen. På baggrund af disse koefficienter beregnes estimater for deres standardafvigelser, t-værdier samt p-værdier.

Figur 5.14 viser one-day-ahead prædiktionerne med Lasso modellen som den grønne kurve, mens den røde og blå kurve som tidligere viser prædiktionerne med AR- og ADL-modellen. Prædiktionerne er igen beregnet for alle hverdage i januar 2022. Prædiktionerne med Lasso har en MSE-værdi på 3,860 og en MAE-værdi på 1,668. Dermed er MSE-værdien for Lasso lavere end for alle de tidligere modeller, hvilket betyder, at Lasso ifølge MSE præsterer bedst out-of-sample. MAE-værdien er næstlavest i forhold til de tidligere modeller, hvor det kun er AR-modellens værdi, som er lavere. At MAE-værdien for Lasso ikke er lavere end AR-modellens kan indikere, at der kan være enkelte store fejl i AR-modellens prædiktioner, som gør, at den ikke præsterer bedst baseret på MSE.



Figur 5.14. Sande værdier for VIX (sort), prædikterede værdier for VIX med AR(10) (rød), med ADL(10,4) (blå), samt med Lasso på ADL(10,4) (grøn).

I forhold til Ridge regression er forecastet fra Lasso væsentlig bedre, hvormed denne form for shrinkage, hvor der straffes hårdere på størrelsen af koefficienterne, er at foretrække. Det kan tyde på, at der er enkelte variable, som har store koefficienter og derfor meget indflydelse, mens de andre er tæt på 0.

5.4 Dimensions reduktion

I dette afsnit anvendes dimension reduktions metoderne fra Afsnit 4.2 på Google Trends datasættet for at opnå et datasæt, der kan forklare en høj del af variationen i observationerne, men hvor antallet af variable er reduceret. Disse nye variable vil blive anvendt som eksterne variable i AR-modellen for VIX indekset fra Afsnit 5.2 for at undersøge, om de kan forbedre modelleringen. Derfor vil der i de efterfølgende afsnit blive opstillet ADL-modeller, hvor antallet af lags for VIX indekset fortsat vil være valgt til $p = 10$.

5.4.1 Principal components analyse for VIX

I dette afsnit udføres principal components analyse for Google Trends datasættet på baggrund af teorien fra Afsnit 4.2.1. Dette gøres for at reducere dimensionen af de eksterne variable ved at transformere Google Trends datasættet. Vi vil reducere de 22 Google Trends variable ved at opstille 22 principal komponenter, som er linearkombinationer af alle variablene og udvælge de komponenter, som forklarer mest af variationen i det oprindelige datasæt. Resultaterne af denne model vil blive

sammenlignet med alle de foregående modeller for at undersøge, om PCA komponenterne kan bidrage som eksterne regressorer i en ADL-model.

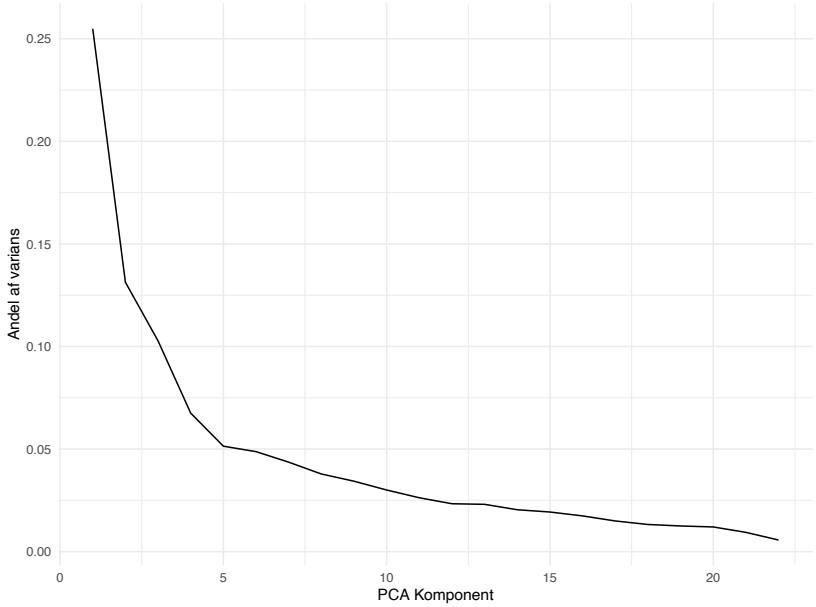
For at anvende PCA skaleres og centreres variablene, sådan at middelværdien fratrækkes, så observationerne centreres i 0. Herefter udføres PCA analysen ved at anvende kommandoen `prcomp` fra pakken `stats` i `R`. Når denne kommando anvendes på Google Trends datasættet, opnås samme antal principal komponenter som antallet af variable i det oprindelige datasæt. Resultatet af dette kan ses i Tabel 5.15, og heraf ses, at vi opnår 22 principal komponenter, som hver forklarer en del af variationen i datasættet. Kolonnen med *Andel af varians* angiver, hvor stor en del af variansen, hver komponent forklarer, hvormed det eksempelvis kan ses, at den første principal komponent, PC1, forklarer 25% af variansen i datasættet med Google Trends observationer. Det betyder, at 25% af informationen i datasættet kan opfanges med én principal komponent. PC2 forklarer 13% af variansen, og af den *Kumulative andel* kan det ses, at PC1 og PC2 tilsammen forklarer 39% af variansen i datasættet.

Komponent	Andel af varians	Kumulativ andel
PC1	0,255	0,255
PC2	0,131	0,386
PC3	0,103	0,489
PC4	0,068	0,557
PC5	0,051	0,608
PC6	0,049	0,657
PC7	0,044	0,700
PC8	0,038	0,738
PC9	0,034	0,772
PC10	0,030	0,802
PC11	0,026	0,829
PC12	0,023	0,852
PC13	0,023	0,875
PC14	0,020	0,896
PC15	0,019	0,915
PC16	0,017	0,932
PC17	0,015	0,947
PC18	0,013	0,960
PC19	0,013	0,973
PC20	0,012	0,985
PC21	0,009	0,994
PC22	0,006	1,000

Tabel 5.15. PCA analyse for VIX indekset.

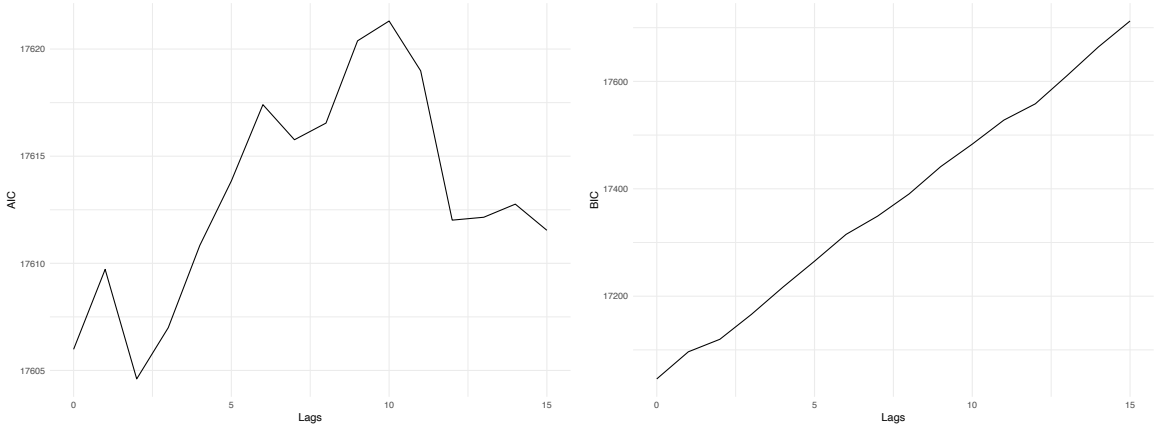
For at vælge antallet af principal komponenter, der skal indgå som eksterne variable i ADL-modellen, plottes andelen af variansen, som forklares af hver PCA komponent, hvilket er vist på Figur 5.16. Heraf ses igen, at PC1 forklarer 25% af variansen i datasættet, hvorefter andelen af forklaret varians falder for hver komponent. Der ses, at kurven efter fem komponenter flader ud, hvormed de efterfølgende komponenter ikke bidrager meget til beskrivelsen af variansen. For at undersøge hvor mange PCA komponenter der skal indgå i ADL-modellen, opstilles modeller med henholdsvis de første $1, 2, 3, 4$ og 5 komponenter som eksterne regressorer. Derefter betragter vi AIC-værdierne for hver af de fem modeller, når antal lags af de eksterne regressorer varieres. Heraf fås, at det udelukkende er modellen med to PCA komponenter, som ifølge AIC skal indeholde lags af PCA komponenterne. Ved de resterende modeller antyder AIC, at der ikke skal indgå PCA komponenter i ADL-modellen, og af denne grund

opstilles ADL-modellen, hvor der indgår to PCA komponenter som eksterne regressorer.



Figur 5.16. Andel af variansen forklaret af hver komponent.

For at vælge hvor mange lags af de to PCA komponenter der skal indgå i ADL-modellen, betragtes AIC- og BIC-værdierne for ADL-modellen med to PCA komponenter, hvor antal lags af komponenterne varieres. Dette er vist på Figur 5.17, hvor vi ønsker at finde minimumsværdien for informationskriterierne, da dette angiver den mest optimale model. Eftersom BIC-værdierne blot stiger med antallet af lags i modellen, vælges minimum for AIC-værdierne, hvilket er modellen, hvor $q = 2$. Dermed opstilles en ADL(10,2)-model med to principal komponenter som eksterne regressorer i stedet for de 22 Google Trends variable, som var tilfældet i Afsnit 5.2.1. Denne model har en justeret R^2-værdi på 0,2748, og forklarer dermed 27% af variationen i VIX indekset.



(a) AIC-værdier for valg af antal lags, q, af principal komponenterne i ADL-modellen for VIX.

(b) BIC-værdier for valg af antal lags, q, af principal komponenterne i ADL-modellen for VIX.

Figur 5.17

Koefficienterne, deres standardfejl, t-værdier og p-værdier for den opstillede ADL(10,2)-model med de to principal komponenterne som eksterne regressorer er vist i Tabel 5.18. Navnet "PC1 1" refererer til det første lag af PC1 komponenten, det vil sige værdien til tid $t-1$. Af p-værdierne ses, at det ikke er alle AR-komponenterne, som er signifikante, og at koefficienterne for begge lags af PC1 komponenten

er insignifikante. Derfor er det udelukkende den anden principal komponent, som er signifikant og som bidrager til fittet af VIX indekset.

PCA komponenterne konstrueres for at kunne forklare variationen i de eksterne regressorer, og vælges derfor ikke ud fra, hvad der forklarer variationen i VIX indekset bedst. Derfor kan det ses at nogle af koefficienterne, som estimeres for disse komponenter i ADL-modellen, er insignifikante. Dette skyldes, at koefficienterne i ADL-modellen beregnes ud fra at forklare variationen i VIX indekset, men PCA komponenterne er konstrueret til at forklare variationen i de eksterne regressorer. Derudover er der, som nævnt i Afsnit 5.2.3, lav korrelation i Google Trends datasættet, hvilket kan være skyld i, at PCA ikke bidrager meget.

	Koefficient	Standardfejl	t-værdi	p-værdi
Skæring	2,38e-03	2,90e-02	8,19e-02	0,935
AR 1	0,311	1,53e-02	20,347	7,54e-88*
AR 2	0,161	1,60e-02	10,104	1,03e-23*
AR 3	9,00e-02	1,62e-02	5,566	2,77e-08*
AR 4	-3,00e-02	1,62e-02	-1,849	6,46e-02
AR 5	6,61e-02	1,62e-02	4,069	4,80e-05*
AR 6	1,84e-02	1,62e-02	1,134	0,265
AR 7	5,08e-03	1,62e-02	0,313	0,754
AR 8	-3,43e-03	1,62e-02	-0,212	0,832
AR 9	7,47e-02	1,60e-02	4,672	3,08e-06*
AR 10	4,80e-02	1,53e-02	3,140	1,70e-03*
PC1 1	-1,317	1,175	-1,121	0,262
PC2 1	1,782	0,897	1,986	4,71e-02*
PC1 2	1,565	1,175	1,332	0,183
PC2 2	-2,050	0,896	-2,287	2,22e-02*

Tabel 5.18. Koefficienter ADL(10,2)-modellen for VIX med PCA komponenter som eksterne regressorer. P-værdier under 0,05 er markeret med *.

For at teste om de fire eksterne regressorer, PCA komponenterne, tilsammen bidrager til fittet af VIX indekset, anvendes F-test. Her testes PCA-modellen mod AR(10)-modellen, som kun indeholder lags af VIX indekset. Af dette fås en F-værdi på 2,340 samt en på p-værdi på 0,053, hvormed nulhypotesen ikke forkastes, og dermed bidrager de fire PCA komponenter ikke simultant til fittet. Dog er p-værdien kun lige over 0,05, og derfor er vi stadig interesserede i at fortsætte analysen.

Tabel 5.19 viser rotationsmatricen for de to PCA komponenter. Det vil sige, at tabellen viser andelen af hver Google Trends variabel, der skal indgå i de to PCA komponenter. Hvis vi betragter den første komponent, ses at *Global Warming* har en vægt på 0,96, hvilken er relativt stor i forhold til de andre, hvormed denne har en stor betydning for komponenten og dermed er vigtig for at forklare variationen i Google Trends observationerne. I den anden PCA komponent, har *Climate Change* den største absolutte værdi, og desuden er vægtene for *Drought* samt *Sea Level Rise* relativt store, hvormed disse variable også har en stor betydning i forhold til at beskrive variationen i Google Trends observationerne. At PCA komponenterne hovedsageligt afhænger af en enkelt Google Trends variabel, skyldes formentlig, at korrelationen i Google Trends datasættet er lav, hvilket sås i forbindelse med VIF-værdierne i Afsnit 5.2.3.

	PC1	PC2
Wildfire	0,010	-0,013
Avalanche	0,045	-0,060
Blizzard	-0,015	-0,005
Climate Change	-0,034	-0,682
Drought	0,228	-0,450
Dust Storm	-0,014	0,007
Earthquake	-0,055	-0,067
Flash Flood	-0,055	-0,070
Flood	-0,032	-0,073
Global Warming	0,955	0,049
Hailstorm	-0,040	-0,009
HeatWave	-0,078	-0,021
Landslide	-0,003	0,001
Mudslide	-0,008	0,001
Pandemic	-0,085	-0,036
Sea Level Rise	-0,016	-0,531
Sinkhole	-0,075	-0,091
Thunderstorm	-0,050	-0,061
Tornado	-0,025	-0,066
Tropical Cyclone	-0,019	-0,079
Tsunami	0,004	0,017
Volcanic Eruption	0,006	-0,073

Tabel 5.19. Rotationsmatrix for PCA-komponenter.

Som tidligere nævnt er hver PCA komponent en linearkombination af alle Google Trends variablene, hvilket vil sige, at de fra (4.6) på Side 18 har formen

$$Z_j = \phi_{1,j}X_1 + \cdots + \phi_{22,j}X_{22}, \quad \text{for } j = 1,2,$$

hvor $X_1,\ldots,X_{22}$ udgør de 22 Google Trends variable. Derudover er ADL(10,2)-modellen med to PCA komponenter til to lags på formen

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_{10} y_{t-10} + \beta_1 Z_{1,t-1} + \beta_2 Z_{2,t-1} + \beta_3 Z_{1,t-2} + \beta_4 Z_{2,t-2}. \tag{5.6}$$

Ved at indsætte udtrykket for hver PCA komponent i (5.6) kan vi gange den tilhørende β koefficient på hver enkelt koefficient i linearkombinationen, og derefter samle koefficienterne for hver variabel til samme lag. På den måde kan vi opnå et udtryk på samme form som ADL-modellen med de 22 Google Trends variable, sådan at koefficienterne for de to modeller kan sammenlignes.

Med denne metode kan koefficienterne for PCA komponenterne for ADL(10,2)-modellen fra Tabel 5.18 omskrives, således vi opnår koefficienter for hver af de laggede Google Trends variable. Disse omskrevne koefficienter er vist i Tabel 5.20. Her betyder *Wildfire 1* eksempelvis, at det er denne variabel til tid $t-1$. Disse koefficienter sammenlignes med de tilsvarende koefficienterne for ADL(10,4)-modellen, som er vist i Tabel B.1 og B.2 i Appendiks B, for at undersøge hvilke variable de to modeller hver lægger vægt på. Det skal dog bemærkes, at der i ADL(10,4)-modellen også indgår observationer af Google Trends variablene til lag 3 og 4, og derfor kan disse koefficienter ikke sammenlignes med. Af tabellerne ses, at alle AR koefficienterne er forholdsvis ens i de to modeller, og de er derfor ikke blevet påvirket af ændringen i de eksterne regressorer. Generelt i forhold til ADL(10,4)-modellen var koefficienterne for de laggede værdier af *Pandemic* markant højere end koefficienterne for de andre Google Trends variable, hvilket betød, at denne variabel generelt var vigtigere end de andre Google Trends Variable i forhold til at kunne beskrive VIX indekset. Af koefficienterne i Tabel 5.20 for ADL(10,2)-modellen med PCA

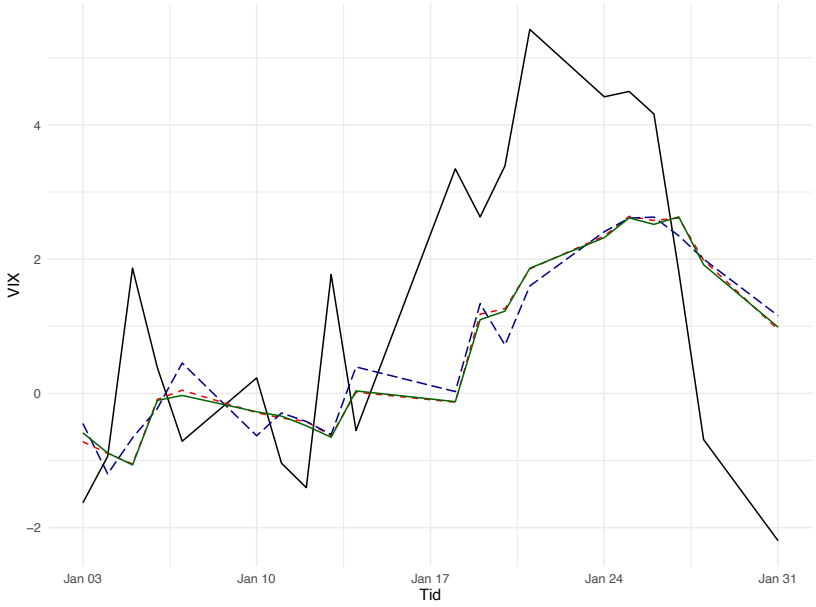
komponenter, kan det dog ses, at koefficienterne for *Pandemic* ikke er højere end koefficienterne for de andre variable, hvilket skyldes at *Pandemic* ikke har fået tildelt en stor vægt i forhold til beskrivelsen af variationen i Google Trends observationerne i PCA analysen. Af tabellen ses, at koefficienterne for *Drought*, *Climate Change* og *Global Warming* er relativt store i forhold til de andre, hvilket skyldes, at det i PCA analysen var disse, som der blev lagt vægt på i forhold til beskrivelsen af variationen i Google Trends observationerne.

	Koefficient		Koefficient		Koefficient
Skæring	0,002	Flood 1	-0,088	Drought 2	1,279
AR 1	0,311	Global Warming 1	-1,172	Dust Storm 2	-0,036
AR 2	0,161	Hailstorm 1	0,036	Earthquake 2	0,050
AR 3	0,090	Heat Wave 1	0,066	Flash Flood 2	0,058
AR 4	-0,030	Landslide 1	0,005	Flood 2	0,100
AR 5	0,066	Mudslide 1	0,011	Global Warming 2	1,395
AR 6	0,018	Pandemic 1	0,047	Hailstorm 2	-0,043
AR 7	0,005	Sea Level Rise 1	-0,925	Heat Wave 2	-0,080
AR 8	-0,003	Sinkhole 1	-0,063	Landslide 2	-0,006
AR 9	0,075	Thunderstorm 1	-0,043	Mudslide 2	-0,013
AR 10	0,048	Tornado 1	-0,084	Pandemic 2	-0,059
Wildfire 1	-0,036	Tropical Cyclone 1	-0,115	Sea Level Rise 2	1,063
Avalanche 1	-0,166	Tsunami 1	0,025	Sinkhole 2	0,069
Blizzard 1	0,012	Volcanic Eruption 1	-0,138	Thunderstorm 2	0,047
Climate Change 1	-1,170	Wildfire 2	0,042	Tornado 2	0,095
Drought 1	-1,102	Avalanche 2	0,193	Tropical Cyclone 2	0,131
Dust Storm 1	0,031	Blizzard 2	-0,015	Tsunami 2	-0,028
Earthquake 1	-0,046	Climate Change 2	1,345	Volcanic Eruption 2	0,159
Flash Flood 1	-0,053				

Tabel 5.20. Koefficienterne for Google Trends variablene til hvert lag opnået med PCA.

Modellen, som er opstillet, indeholder kun laggede værdier af PCA, og som alternativ kunne den nutidige værdi af PCA komponenterne inddrages i ADL-modellen, sådan at PCA komponenterne inddrages til tid t, $t-1$ og $t-2$. Dette ville give en indikation af, om Google Trends variablene til samme tidspunkt er med til at forklare variationen i VIX indekset. Dog fås for denne model en højere AIC-værdi og en lavere forklaringsgrad i form af den justerede R^2-værdi, som for denne model er 27,45, hvor den før var 27,48 Dette betyder, at modellen ikke er bedre in-sample, og eftersom den vil være vanskelig at anvende til prædiktioner, fordi de nutidige værdier af Google Trends variablene er ukendte, vurderes denne model overordnet til ikke at være bedre og den betragtes derfor ikke yderligere.

Når VIX-værdierne for alle hverdage i januar 2022 forecastes med samme metode som tidligere, fås værdierne, der er plottet som den grønne kurve på Figur 5.21. Heraf ses, at prædiktionerne med PCA metoden lægger sig meget op ad prædiktionerne fra den oprindelige AR(10)-model, som er vist ved den røde kurve. Den blå kurve viser som tidligere prædiktionerne med ADL(10,4)-modellen fra Afsnit 5.2.1.



Figur 5.21. Sande værdier for VIX (sort), prædikterede værdier for VIX med AR(10) (rød), med ADL(10,4) (blå), samt med PCA ADL(10,2) (grøn).

I forhold til prædiktionerne har ADL(10,2)-modellen med PCA komponenterne som eksterne regressorer en MSE-værdi på 3,848, hvilket er lavere end alle de foregående værdier. Desuden er MAE-værdien 1,651, hvilket er ens med AR(10)-modellens værdi, som er lavere end de andre modellers. Det vil sige, at PCA modellen præsterer bedst out-of-sample, eftersom den har lavest værdier i begge fejlestimater.

5.4.2 Partial least squares regression for VIX

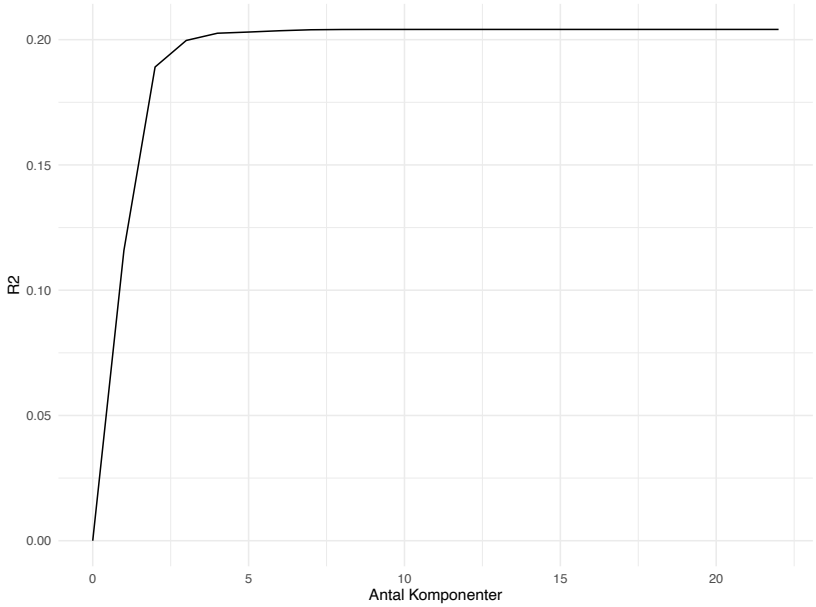
I dette afsnit udføres partial least squares regression for VIX indekset på baggrund af teorien fra Afsnit 4.2.2. Dette gøres, ligesom ved PCA, for at reducere dimensionen af de eksterne regressorer i ADL-modellen. PLS er en supervised dimensions reduktion metode, som til forskel fra PCA også anvender responsvariablen i konstruktionen af komponenterne. Resultaterne fra denne model vil blive sammenlignet med med resultaterne fra de tidligere modeller for at undersøge, hvilken model der forklarer VIX indekset bedst.

Først skaleres og centreres datasættet i 0, hvorefter kommandoen `plsr` fra `R` pakken `pls`, [Liland et al., 2021], benyttes til at opstille PLS komponenterne, og resultatet af denne analyse ses i Tabel 5.22. Denne tabel indeholder de 22 komponenter, som igen hver forklarer en del af variationen i henholdsvis de forklarende variable samt responsvariablen. Her ses eksempelvis, at den første komponent, *Comp 1*, forklarer 16% af variationen i de forklarende variable, X, samt 11% af variationen i responsvariablen VIX. Andelen af forklaret varians for VIX er angivet kumulativt, hvormed den stiger i starten for derefter at stagnere på 20,4%.

Komponent	Andel af varians X	Kumulativ andel af varians VIX
Comp 1	0,162	0,116
Comp 2	0,135	0,189
Comp 3	0,141	0,200
Comp 4	0,040	0,203
Comp 5	0,097	0,203
Comp 6	0,040	0,204
Comp 7	0,021	0,204
Comp 8	0,022	0,204
Comp 9	0,030	0,204
Comp 10	0,038	0,204
Comp 11	0,034	0,204
Comp 12	0,030	0,204
Comp 13	0,028	0,204
Comp 14	0,021	0,204
Comp 15	0,024	0,204
Comp 16	0,020	0,204
Comp 17	0,020	0,204
Comp 18	0,029	0,204
Comp 19	0,015	0,204
Comp 20	0,021	0,204
Comp 21	0,017	0,204
Comp 22	0,018	0,204

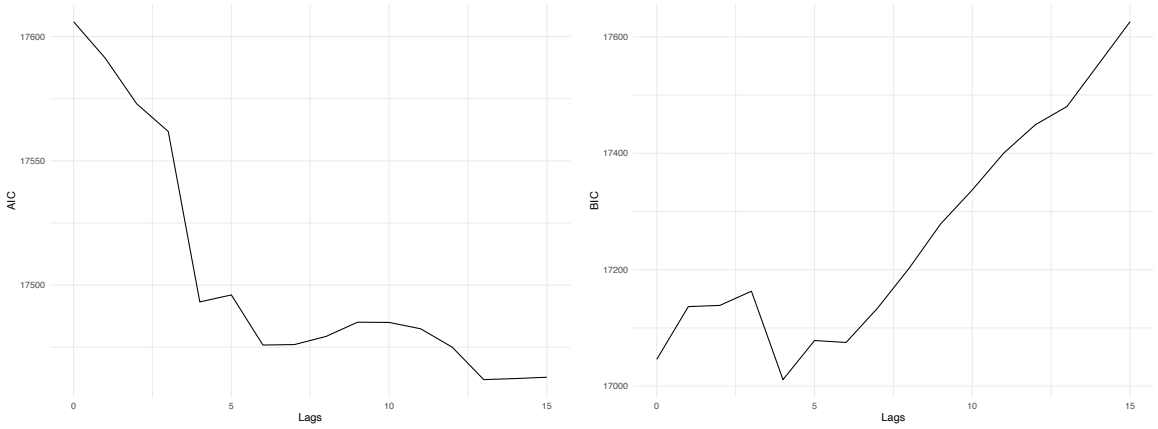
Tabel 5.22. PLS analyse for VIX.

For at afgøre antallet af komponenter der skal indgå i ADL-modellen, er R^2-værdierne for PLS regressionen af VIX plottet mod antallet af komponenter, der er anvendt i regressionen, hvilket ses på Figur 5.23. Af dette ses et knæk i kurven ved tre komponenter, hvorefter der ikke længere opnås en signifikant forbedring i R^2-værdien ved at øge antallet af komponenter. Derudover ses af Tabel 5.22 at der efter tre komponenter sker et stort fald i andelen af forklaret varians i X. Derfor udvælges de tre første PLS komponenter til at indgå som eksterne variable i ADL-modellen.



Figur 5.23. PLS R^2-værdier for VIX.

Herefter betragtes AIC- og BIC-værdier for at udvælge antallet af lags af de tre PLS komponenter, som skal indgå som eksterne regressorer i ADL-modellen for VIX. Af AIC-værdierne på Figur 5.24a ses, at den laveste værdi er ved $q = 15$, men der er også et væsentlig fald i AIC ved $q = 4$. Netop denne værdi er interessant, fordi BIC-værdierne er minimeret ved $q = 4$, hvilket ses på Figur 5.24b. Derfor vælges antal lags af PLS komponenterne, som skal indgå i ADL-modellen, til $q = 4$.



(a) AIC mod antal lags for en PLS-model for VIX. **(b)** BIC mod antal lags for en PLS-model for VIX.

Figur 5.24

Med dette antal lags har ADL(10,4)-modellen, med PLS komponenter som eksterne regressorer, en justeret R^2-værdi på 0,295. Dette er en højere forklaringsgrad sammenlignet med ADL-modellen med PCA komponenter, hvilket stemmer overens med, at PLS metoden anvender responsvariablen til at udvælge komponenterne. Modellens koefficienter, deres standardfejl, t-værdier samt p-værdier ses i Tabel 5.25. Her henviser "*Comp1 1*" for eksempel til første komponent til tid $t - 1$. Af tabellen ses blandt andet, at komponent 2 er signifikant til alle fire lags, hvormed denne særligt bidrager til ADL-modellens fit af VIX indekset. Derudover er de samme AR-komponenter som tidligere signifikante.

	Koefficient	Standardfejl	t-værdi	p-værdi
Skæring	2,10e-03	2,86e-02	7,34e-02	0,942
AR 1	0,313	1,52e-02	20,583	8,40e-90*
AR 2	0,154	1,59e-02	9,696	5,29e-22*
AR 3	7,16e-02	1,61e-02	4,464	8,24e-06*
AR 4	-1,75e-02	1,61e-02	-1,086	0,278
AR 5	5,77e-02	1,61e-02	3,592	3,32e-04*
AR 6	2,31e-02	1,61e-02	1,434	0,152
AR 7	2,35e-03	1,61e-02	0,146	0,884
AR 8	9,56e-03	1,60e-02	0,597	0,551
AR 9	6,94e-02	1,58e-02	4,394	1,14e-05*
AR 10	4,14e-02	1,52e-02	2,731	6,35e-03*
Comp1 1	2,209	1,639	1,348	0,178
Comp2 1	7,481	1,384	5,406	6,79e-08*
Comp3 1	1,634	0,904	1,808	7,06e-02
Comp1 2	-0,369	1,930	-0,191	0,848
Comp2 2	-10,581	1,726	-6,129	9,65e-10*
Comp3 2	-2,416	1,077	-2,244	2,49e-02*
Comp1 3	2,943	1,931	1,525	0,127
Comp2 3	10,982	1,734	6,335	2,62e-10*
Comp3 3	4,195	1,075	3,901	9,72e-05*
Comp1 4	-3,875	1,638	-2,366	1,80e-02*
Comp2 4	-7,282	1,385	-5,256	1,55e-07*
Comp3 4	-2,749	0,904	-3,040	2,38e-03*

Tabel 5.25. Koefficienter ADL(10,4)-modellen for VIX med PLS komponenter som eksterne regressorer. P-værdier under 0,05 er markeret med *.

Igen testes om de eksterne regressorer, her i form af PLS komponenter, tilsammen er signifikante i modelleringen af VIX indekset. Det gøres som tidligere med F-test, hvor modellen med PLS komponenter testet mod AR(10)-modellen. Heraf fås en F-værdi på 11,518 og en tilhørende p-værdi på 2,2e-16, hvilket betyder, af PLS komponenterne sammen er signifikante til beskrivelsen af VIX indekset.

Hver PLS komponent er som tidligere nævnt en linearkombination af de 22 Google Trends variable, og Tabel 5.26 viser rotationsmatricen, som indeholder de koefficienter, der indgår i linearkombinationerne. Dermed angiver tabellen størrelsen af vægten, der lægges på hver Google Trends variabel i de to PLS komponenter. Her ses blandt andet, at der for alle tre komponenter gælder, at *Pandemic* har den højeste vægt, mens der i de to første komponenter også lægges en forholdsvis stor vægt på *Global Warming*. Dette skyldes, at PLS som tidligere nævnt er en supervised metode, hvor responsvariablen også betragtes i beregningen af komponenterne. Komponenterne skal dermed både kunne forklare variationen i responsariablen, det vil sige VIX indekset, samt de eksterne regressorer som består af Google Trends variablene. I ADL(10,4)-modellen så vi, at *Pandemic* var signifikant i forhold til at forklare variationen i VIX indekset og i forbindelse med PCA i Afsnit 5.4.1 så vi, at *Global Warming* var signifikant i forhold til at forklare variationen i Google Trends datasættet. Det stemmer derfor overens med disse resultater, at PLS metoden lægger signifikant vægt på disse to variable.

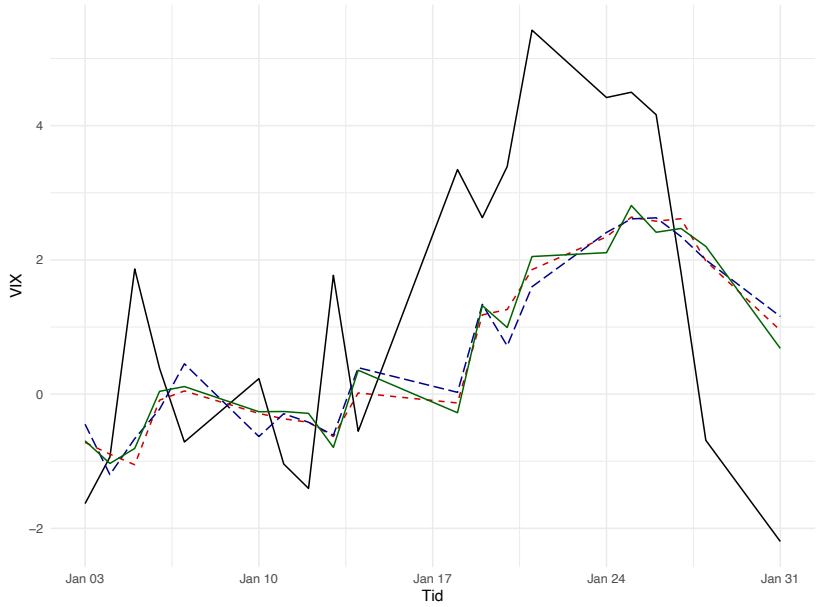
	Comp1	Comp2	Comp3
Wildfire	-0,028	-0,024	0,069
Avalanche	0,133	0,163	0,174
Blizzard	0,007	0,035	0,049
Climate Change	-0,078	0,042	0,199
Drought	-0,093	-0,117	0,822
Dust Storm	0,039	0,072	0,030
Earthquake	0,074	0,180	0,076
FlashFlood	-0,043	0,018	0,047
Flood	-0,131	-0,120	0,079
Global Warming	0,502	-0,355	-0,010
Hailstorm	-0,065	-0,046	-0,007
Heat Wave	-0,110	-0,072	-0,044
Landslide	-0,019	-0,026	-0,016
Mudslide	-0,063	-0,105	-0,145
Pandemic	0,744	1,223	0,668
Sea Level Rise	-0,237	-0,186	0,226
Sinkhole	-0,201	-0,161	0,076
Thunderstorm	-0,037	0,008	-0,018
Tornado	-0,051	-0,044	-0,049
Tropical Cyclone	-0,116	-0,134	-0,021
Tsunami	-0,013	-0,032	-0,057
Volcanic Eruption	-0,028	-0,050	-0,057

Tabel 5.26. Rotationsmatrix for PLS-komponenter.

På samme måde som i PCA kan vi gange koefficienterne for hver PLS komponent, som er vist i rotationsmatricen, med den respektive koefficient i ADL-modellen fra Tabel 5.25. Dermed opnås koefficientestimater, som kan sammenlignes direkte med ADL(10,4)-modellens koefficienter, idet der i begge modeller indgår fire lags af de eksterne regressorer. Disse koefficientestimater er vist i Tabel B.4 i Afsnit B.1.3 i Appendiks B, mens koefficienterne for ADL(10,4)-modellen, som der sammenlignes med, er vist i Tabel B.1 og B.2 i Appendiks B. Heraf ses, at vægtfordelingen mellem Google Trends variablene i de to modeller minder meget om hinanden, eftersom alle koefficienter for de laggede værdier af *Pandemic* er meget større end de andre variables koefficienter. Hvis disse koefficienter sammenlignes med PCA koefficienterne fra forrige afsnit ses, at der væsentlig forskel. Dette skyldes, at PLS som tidligere nævnt er en supervised metode, som forsøger at beskrive responsvariablen, VIX indekset, samtidig med også at beskrive variationen i Google Trends observationerne.

Som tidligere undersøges om de nutidige værdier af PLS komponenterne kan bidrage til et bedre fit af VIX indekset. Modellen, hvor de nutidige værdier tilføjes, har en justeret R^2-værdi på 0,299, hvilket er en smule højere end før, og desuden forbedres AIC-værdien også en smule. Dog er det ikke en stor nok forbedring i forklaringsgraden og AIC til, at modellen foretrækkes fremadrettet i forbindelse med forecasting, eftersom dette ville betyde, at vi skulle forecaste Google Trends variablene en dag frem for at kunne forecaste VIX indekset. Dette ville medfører stor usikkerhed i prædiktionerne, eftersom vi derved ville bygge vores forecastede værdier på andre forecastede værdier. Derfor anvendes modellen med de nutidige værdier ikke yderligere.

Med ADL(10,4)-modellen med PLS komponenter som eksterne regressorer prædikteres nu alle hverdage i januar 2022 på samme måde som tidligere. Prædiktionerne er vist som den grønne kurve på Figur 5.27, som også viser AR-modellen og den oprindelig ADL-models prædiktioner som henholdsvis den røde og blå kurve. Heraf ses, at PLS prædiktionerne er relativt ens med AR samt ADL-modellens prædiktioner, hvilket også ses i forbindelse med prædiktionsfejlene, hvor modellen har en MSE-værdi på 3,929 og en MAE-værdi på 1,680. Sammenlignet med PCA modellen, præsterer PLS bedre in-sample, idet de to justerede R^2-værdier er henholdsvis 0,275 og 0,295, hvilket formentlig skyldes, at der tages højde for VIX indekset under konstruktionen af PLS komponenterne. Dog er prædiktionsfejlene for PLS en smule højere end fejlene for PCA, og modellen præsterer derfor ikke bedre out-of-sample.



Figur 5.27. Sande værdier for VIX (sort), prædikterede værdier for VIX med AR(10) (rød), med ADL(10,4) (blå), samt med PLS ADL(10,4) (grøn).

I det følgende afsnit opstilles et neuralt netværk for VIX indekset for at undersøge, hvordan det kan bidrage til beskrivelsen af volatiliteten af S&P500 indekset, som udtrykkes ved VIX.

5.5 Neurale Netværk

I dette afsnit vil vi opstille neurale netværk for det fraktionelt differensede VIX indeks, på baggrund af teorien fra Afsnit 4.3. Dette gøres for at undersøge, om modeller opstillet med denne machine learning metode præsterer bedre end modeller opstillet med de hidtil anvendte metoder, og om klimaforandringerne kan bidrage til beskrivelsen af VIX indekset ved brug af neurale netværk. Inputtet i netværket vil bestå af laggede værdier af VIX indekset, hvorefter det senere vil blive udvidet til også at indeholde laggede værdier af Google Trends variablene. Alle inputvariablene standardiseres forud for modelleringen.

5.5.1 Neuralt netværk med VIX som input

For at udvælge et neuralt netværk opstilles en række forskellige netværk på en træningsmængde. Eftersom neurale netværk hurtigt bliver komplicerede, anvendes perioden juni 2015 - maj 2020 som træningsmængde, hvilket udgør 1258 observationer. Testmængden udgør de resterende 402 observationer, hvilket vil sige perioden fra juni 2020 til december 2021. Disse netværk vurderes in-sample på baggrund af AIC samt out-of-sample ved MSE og MAE, hvor de sidste to værdier er beregnet på testmængden. De neurale netværk opstilles i `R` med funktionen `neuralnet` fra pakken `neuralnet`, [Fritsch et al., 2019], hvor vi anvender aktiveringsfunktionen TanH fra (4.13) på Side 25.

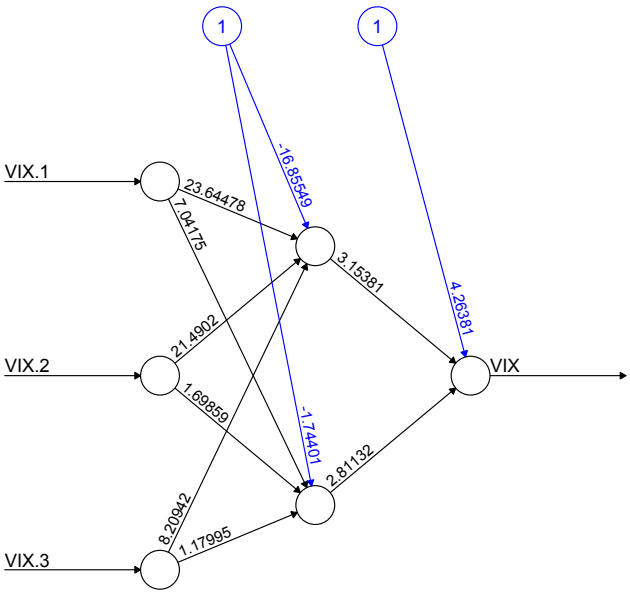
Tabel 5.28 viser resultaterne for en række neurale netværk med ét skjult lag, hvilke er opstillet på baggrund af laggede VIX værdier. Her angives antallet af lags, der indgår i netværket, som *VIX lags*.

Det betyder for eksempel, at "2 VIX lags" angiver, at vi har et neuralt netværk med to inputvariable, som er VIX indekset til tiden $t-1$ og til tiden $t-2$. Antal lags varieres gennem (1,2,3,5,10), og antallet af knuder i det skjulte lag betegnes d. I tabellen er de laveste værdier af henholdsvis AIC, MSE og MAE markeret med en *, og her har netværket med tre lags af VIX indekset og to knuder i det skjulte lag de laveste værdier out-of-sample i form af MSE- og MAE-værdierne. Derimod har netværket med 10 VIX lags og tre knuder den laveste AIC-værdi. Dog lægges mere vægt på out-of-sample præstationen, idet vi da undgår en model, der overfitter observationerne. Desuden udgør den første af disse to netværk det mest simple, hvormed det foretrækkes, og derfor vises resultaterne for netop dette neurale netværk, som er markeret med fed i tabellen.

	AIC	MSE	MAE
1 VIX lag			
d=1	5615	4,636	1,387
d=2	5441	4,689	1,384
d=3	5095	5,222	1,397
2 VIX lags			
d=1	5114	4,499	1,357
d=2	5122	4,500	1,357
d=3	4849	4,469	1,346
3 VIX lags			
d=1	5055	4,469	1,359
d=2	**4961**	**4,151***	**1,321***
d=3	3974	4,335	1,349
5 VIX lags			
d=1	5019	4,491	1,357
d=2	4383	4,288	1,338
d=3	4179	4,981	1,416
10 VIX lags			
d=1	5087	4,248	1,347
d=2	4499	4,188	1,371
d=3	3461*	5,939	1,424

Tabel 5.28. Resultater for en række neurale netværk af forskellig orden, hvor de laveste værdier er markeret med *. Her angiver d antal knuder i det skjulte lag, og den udvalgte model er markeret med fed.

Af tabellen ses, at det udvalgte neurale netværk har en AIC-værdi på 4961, og Figur 5.29 viser en illustration af netværket med de tilhørende beregnede vægte. På figuren angives første lag af VIX indekset som VIX.1, hvor tilsvarende gælder for de andre lags.



Figur 5.29. Neuralt netværk for VIX indekset med tre lags af VIX samt to knuder i det skjulte lag.

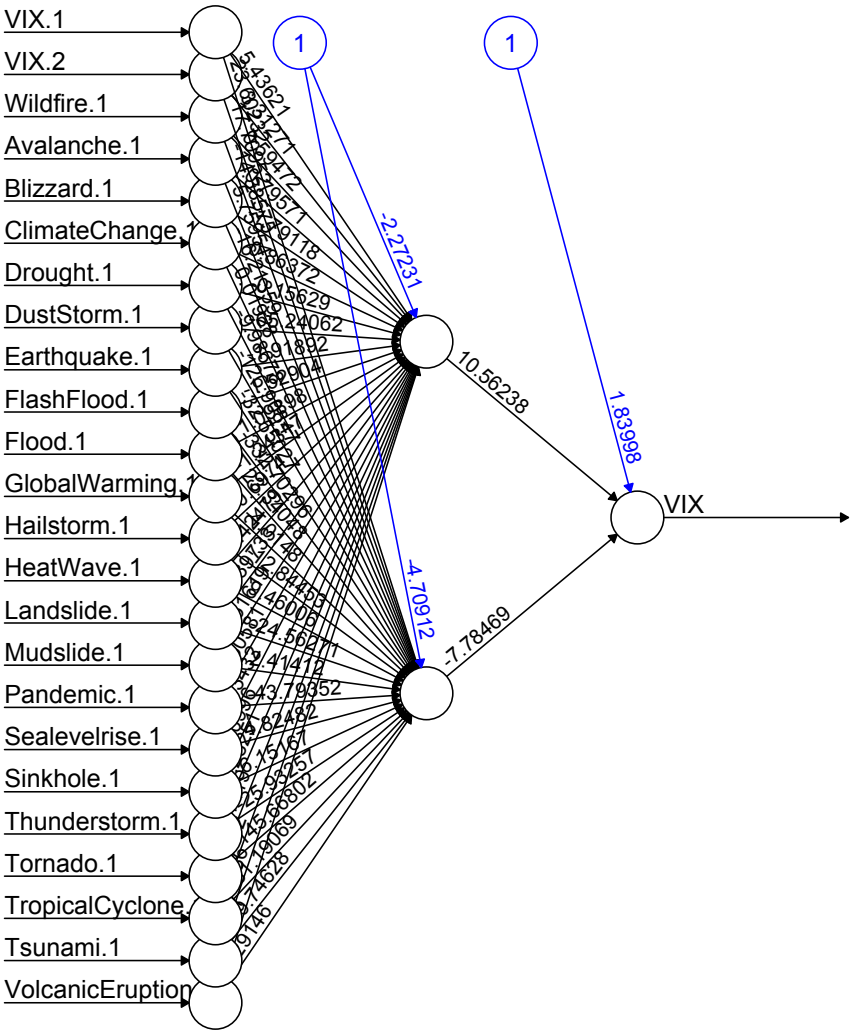
Dette neurale netværk har dermed kun haft tidligere værdier af VIX indekset som inputvariable, og i næste afsnit udvides netværket til også at inddrage Google Trends variablene.

5.5.2 Udvidet neuralt netværk

For at undersøge om Google Trends variablene kan bidrage til forklaringen af VIX indekset ved brug af neurale netværk, tilføjes disse som inputvariable i netværket.

Tabel B.5 i Afsnit B.1.4 i Appendiks B viser resultaterne for en række neurale netværk, som opstilles for at vurdere hvilket netværk, der præsterer bedst. De neurale netværk varieres i form af antal lags af VIX indekset, antal lags af Google Trends variablene samt antal knuder i det skjulte lag i det neurale netværk, hvor vi undersøger 45 forskellige kombinationer. Disse netværk vurderes igen både in-sample og out-of-sample. Tilsvarende netværkene i forrige afsnit opstilles de neurale netværk på en træningsmængde, som består af observationer over fem år, fra juni 2015 til maj 2020, og de testes på testmængden, som er den resterende periode frem til december 2021.

De laveste værdier i hver statistik er markeret ved *, og af dette ses ved MSE- og MAE-værdierne, at out-of-sample præsterer netværket med to VIX lags, ét lag af Google Trends variablene og to knuder i det skjulte lag bedst. Denne model er markeret med fed i tabellen. I forhold til AIC-værdierne har modellen med 10 VIX lags, tre Google Trends lags samt tre knuder i det skjulte lag den laveste værdi, dog har denne en meget høj MSE-værdi, hvilket kan tyde på, at den overfitter. Dermed udvælges de førstnævnte neurale netværk til at beskrive VIX indekset, da dette overordnet er bedst på alle tre parameter. Af tabellen ses, at det udvalgte neurale netværk har en AIC-værdi på 4085, som sammenlignet med AIC-værdien for det forrige netværk er lavere, hvormed dette netværk beskriver VIX indekset mere præcist in-sample. Det kunne derfor tyde på, at det har bidraget til modellens præcision i forklaringen af VIX indekset at inkludere Google Trends variablene. Figur 5.30 viser netværket med de tilhørende estimerede vægte, hvor VIX.1 som før angiver det første lag af VIX indekset. Derudover er alle Google Trends variablene også lagget, så de er til tiden $t-1$.



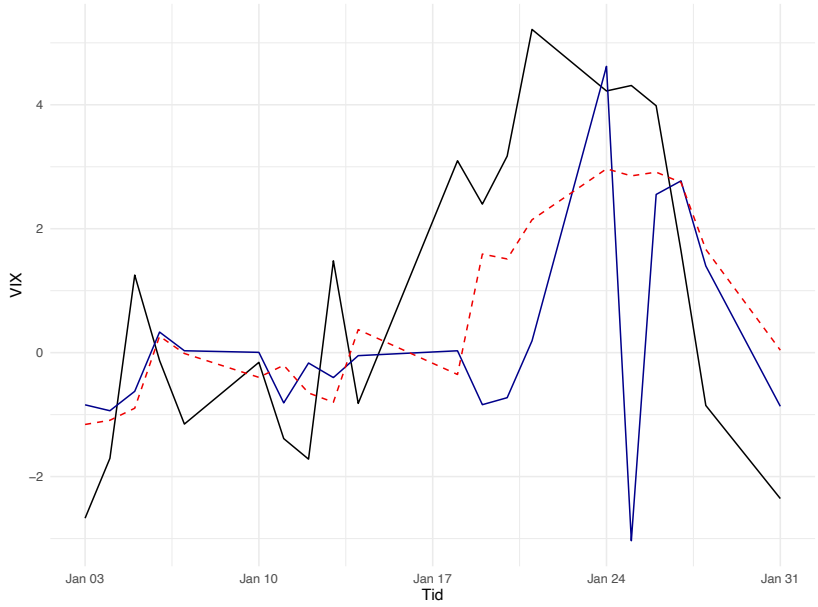
Figur 5.30. Neuralt netværk for VIX indekset med to lags af VIX, ét lag af Google Trends variablene samt to knuder i det skjulte lag.

Vægtene forbundet til de to knuder i det skjulte lag i det neurale netværk er vist i Tabel B.6 i Afsnit B.1.4 i Appendiks B. Eftersom beregningerne i de skjulte lag i et neuralt netværk er som en black box, er det svært at kommentere på vægtenes betydning, hvilket også nævnes i kilden [Ivanovs et al., 2021].

5.5.3 Prædiktioner af VIX med neurale netværk

I dette afsnit prædikteres VIX indekset med de to opstillede neurale netværk for at kunne sammenligne prædiktionsevner med de hidtil opstillede modeller for VIX indekset.

Figur 5.31 viser de prædikterede værdier for begge de neurale netværk, som er opstillet for VIX indekset. Det første netværk, der udelukkende havde lags af VIX som inputvariable, refereres til som 'Netværk 1', og det udvidede netværk med Google Trends variable som ekstra inputvariable refereres til som 'Netværk 2'. Prædiktionerne er som tidligere beregnet over perioden med alle hverdage i januar 2022. Prædiktionerne med Netværk 1 er vist som den røde kurve, mens den blå viser prædiktionerne med Netværk 2. Fra prædiktionerne med Netværk 1 fås en MSE-værdi på 3,042 samt en MAE-værdi på 1,528, mens Netværk 2 har værdier på henholdsvis 7,046 og 2,016. Dermed har det første netværk uden Google Trends variable de bedste prædiktioner, hvilket også ses på figuren, idet den blå kurve har et stort udsving, der ikke følger tendensen i de sande værdier, som er vist som den sorte kurve. Det har derfor ikke bidraget til netværkets prædiktionsevne at inkludere information om klimaforandringerne i form af Google Trends variablene.



Figur 5.31. Prædikterede værdier fra neurale netværk for VIX indekset. Prædiktionerne med Netværk 1 er vist som den røde kurve, mens prædiktionerne med Netværk 2 er vist som den blå.

I næste afsnit opsamles på resultaterne fra alle de opstillede modeller for VIX indekset.

5.6 Opsamling på resultaterne for VIX

For at sammenligne de otte opstillede modeller for VIX indekset har vi samlet resultaterne fra hver model i Tabel 5.32. Dette inkluderer MSE- og MAE-værdier til beskrivelse af modellernes præstation out-of-sample, samt justerede R^2-værdier for de autoregressive modeller og AIC-værdier for de neurale netværk for at vurdere modellerne in-sample. De bedste værdier er markeret med *, hvilket betyder, at af de autoregressive modeller er det ADL(10,4)-modellen, der har den højeste justerede R^2-værdi, hvormed den præsterer bedst in-sample og er bedst til at forklare variationen i VIX indekset. Den bedst mulige model in-sample fås dermed ved at inkludere alle 22 Google Trends variable som eksterne regressorer i modellen, og forbedres ikke af at udføre Shrinkage eller dimensions reduktion. Dette stemmer overens med resultaterne for F-test, hvor de eksterne regressorer tilsammen var insignifikante i forbindelse med Rigde og PCA. I PLS var de eksterne regressorer tilsammen signifikante, hvilket også passer med, at den justerede R^2-værdi er højere for denne model end den første AR(10), som blev opstillet. Dog skal det bemærkes, at der ingen justeret R^2-værdi er for de neurale netværk, hvormed disse modeller ikke er sammenlignelige med de autoregressive modeller in-sample.

Model	Justeret R^2	Forecast MSE	Forecast MAE
AR(10)	0,274	3,867	1,651
ADL(10,4)	0,324*	4,078	1,739
ADL Ridge	0,264	4,623	1,804
ADL Lasso	0,267	3,860	1,668
ADL PCA	0,275	3,848	1,651
ADL PLS	0,295	3,929	1,680
	AIC	Forecast MSE	Forecast MAE
Neuralt netværk 1	4961	3,042*	1,528*
Neuralt netværk 2	4085*	7,046	2,016

Tabel 5.32. Resultater for alle modellerne for VIX indekset, det bedste resultat i hver statistik er markeret med *.

Af resultaterne for de neurale netværk ses, at Netværk 2 har den laveste AIC-værdi, hvilket betyder, at dette netværk er bedst til at beskrive VIX indekset in-sample, og derfor har Google Trends variablene bidraget positivt. I forbindelse med prædiktionerne af VIX indekset for hverdage i januar 2022 har det neurale netværk uden Google Trends variable de laveste prædiktionsfejl, både i form af MSE og MAE, og dermed har denne model været mere præcis out-of-sample end alle de andre modeller, som er opstillet for VIX indekset.

Sammenlignet med den oprindelige AR-model, som ikke indeholdt Google Trends variablene, har der været en lille forbedring i nogle af de udvidede ADL-modeller, især i forbindelse med den justerede R^2-værdi, hvor ADL(10,4)-modellen forklarer 5 procentpoint mere af variationen i VIX indekset end AR(10)-modellen. Dermed har klimaforandringerne i form af Google Trends data bidraget til beskrivelsen af volatiliteten i det finansielle marked, når det er målt ved VIX indekset og modelleret ved ADL-modeller. Når de neurale netværk betragtes ses, at forecastet er væsentlig bedre, når netværket udelukkende er baseret på tidligere værdier af VIX indekset. Det vil sige, at når VIX indekset er blevet modelleret ved neurale netværk, har klimaforandringerne gjort prædiktionerne mere upræcise, hvilket tydeligt ses af forecastfejlene i tabellen, idet værdierne for dette netværk er de højeste.

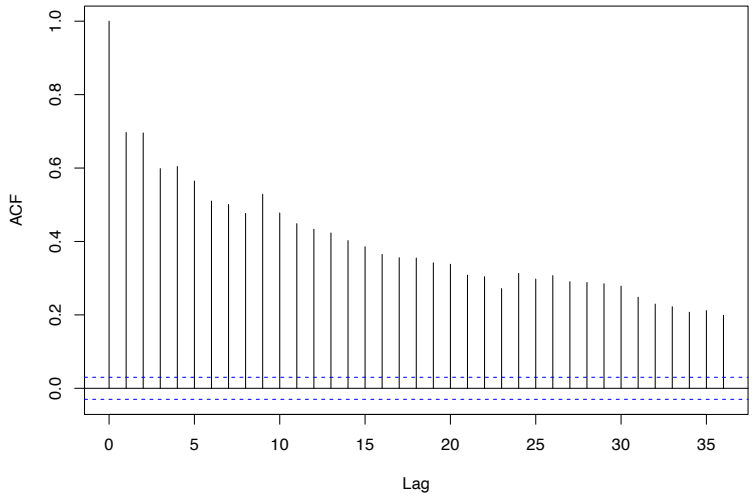
Volatilitet kan dog måles på flere måder, og derfor vil det i næste kapitel undersøges om Google Trends variablene kan bidrage til forklaringen af volatiliteten i markedet målt ved den realiserede varians.

Modellering af RV indekset 6

I dette kapitel vises resultaterne for modelleringen af realiseret varians for S&P500, som er tilsvarende de analyser, der blev foretaget i forrige kapitel for VIX indekset. Dette gøres, eftersom RV er en anden indikator for volatiliteten i S&P500 indekset, og det kunne derfor være interessant at undersøge, om der i de to volatilitetsindekser er samme tendenser i forhold til, om klimaforandringerne påvirker dem. Igen opsamles resultaterne for modelleringen af RV indekset til sidst i kapitlet.

Den realiserede varians for S&P500 er hentet fra kilden [Oxford-Man-Institute, 2022] over perioden fra den 1. januar 2005 til den 31. december 2021, hvilket er den samme periode, der blev anvendt i VIX modelleringen i forrige kapitel, og et plot af indekset kan ses på Figur 2.6 på Side 9. Tidsrækken for RV indekset består af 4269 observationer, hvortil det skal bemærkes, at dette er 11 observationer mindre end VIX indekset. Dette giver en lille variation i Google Trends datasættet, hvilket påvirker PCA analysen, som ellers ville være ens. På samme vis som ved VIX indekset undersøges, om RV indekset er stationært ved at teste hypotesen om en enhedsrod med funktion `ur.df` i `R`. Denne nulhypotese forkastes, hvormed tidsrækken for RV også er stationær.

Som tidligere nævnt er der evidens for, at volailitetsindekser udviser long memory, og derfor betragtes autokorrelationsfunktionen for RV indekset, hvilken er vist på Figur 6.1. Figuren viser, at der er en høj og persistent autokorrelation i RV indekset, og Hurst statistikken er beregnet til 0,742, hvilket indikerer, at der long memory i tidsrækken. Dog ses, at autokorrelationen er væsentlig mindre persistent i RV end den var i VIX , hvis ACF blev vist på Figur 5.2a på Side 29. Derfor estimeres parametren d, som angiver, hvor meget tidsrækken skal fraktionelt differenses, hvilket udføres på samme måde som tidligere ved brug af optimeringsalgortimen BFGS.



Figur 6.1. Autokorrelationsfunktion for RV indekset.

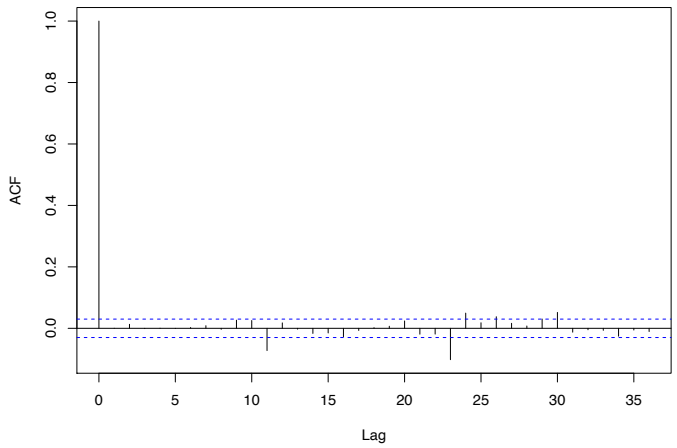
Den optimale værdi for parametren d i AR-modelleringen for RV indekset kan findes ved de to metoder, som blev anvendt for VIX i Afsnit 5.2. Det vil sige, at d kan findes først, hvorefter AR-modellerne opstilles på den fraktionelt differensede tidsrække, eller d og AR-modellens koefficienter kan estimeres simultant i ét step. Ved at opstille de optimale modeller med disse to metoder, fås med ét step en AR(9) og med to step en AR(10), og resultaterne i forbindelse med modellerne er vist i Tabel 6.2. Her ses, at d er væsentlig mindre ved modellen, som estimeres simultant, og at RSS værdierne for de to modeller er næsten ens.

Da parametren d næsten er 0 i et-steps modellen, antyder det, at der ikke er long memory i RV indekset. Dette skyldes højst sandsynligt, at når modellen estimeres i ét step, så tager AR parametrene højde for en så stor del af den høje autokorrelation i tidsrækken, at parameteren d ikke længere er så betydelig. Parameteren d har en p-værdi på 0,998, og dermed er d ikke signifikant, hvorfor det undersøges, om en AR-model på det udifferensede indeks er at foretrække. Resultaterne for denne model kan ligeledes ses i Tabel 6.2 som AR(9)-modellen "uden d". RSS værdierne for ét step modellen samt modellen på det udifferensede indeks er ens, hvilket kan tyde på, at der ikke er den store forskel i modellerne. Vi har desuden undersøgt koefficientestimaterne for begge modeller, hvilke var næsten identiske. Af AIC-værdierne ses, at modellen baseret på det udifferensede data har lavest værdi, hvorfor denne er den mest optimale. Det vil sige, at AR(9)-modellen, hvor RV indekset ikke er differenset er at foretrække, og denne model udvælges til at beskrive RV indekset. Når vi fremadrettet refererer til RV indekset, menes der derfor det ikke-differensede RV indeks.

Metode	Model	d	Standardafvigelse d	RSS	AIC
Ét step	AR(9)	4,58e-05	1,95e-02	1,39e-04	-61414
To steps	AR(10)	0,449	0,011	1,39e-04	-61318
Uden d	AR(9)	0	-	1,39e-04	-61458

Tabel 6.2. Resultater for de optimale modeller, som er estimeret i henholdsvis ét og to steps, samt en AR-model på udifferenset data.

For at undersøge om den udvalgte AR(9)-model fitter den realiserede varians tilstrækkeligt, plottes ACF'en for residualerne. Det ønskes, at disse er tilnærmelsesvis hvid støj, eftersom det vil betyde, at modellen opfanger variationen i RV indekset. ACF'en for residualerne er vist på Figur 6.3, hvor det ses, at der er meget begrænset autokorrelation, hvormed residualerne kan betragtes som hvid støj.



Figur 6.3. Autokorrelationsfunktion af residualerne for AR(9)-modellen.

I det følgende afsnit uddybes opstillingen af den valgte autoregressive model for RV indekset, og i de efterfølgende afsnit vil modellen blive udvidet ved at anvende Google Trends variablene for klimaforandringer som eksterne regressorer.

6.1 Autoregressive modeller for RV

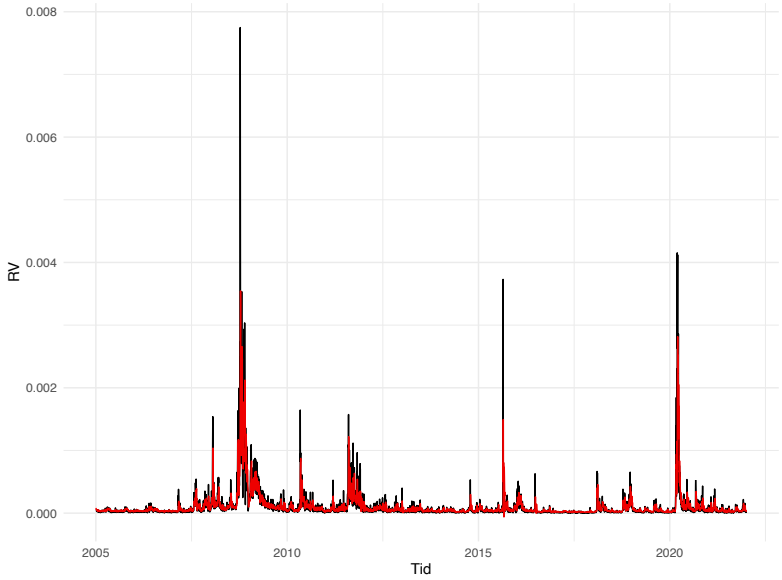
For at vælge den optimale lagorden, p, af en AR-model for den realiserede varians, betragtes AIC-værdierne for modeller med forskellig lagorden. Dette gøres automatisk med kommandoen `auto.arima` i `forecast`-pakken i `R`. Heraf er resultatet, at det antal lags som minimerer informationskriteriet er $p = 9$, hvormed der vælges en AR(9)-model med resultaterne, som blev vist i Tabel 6.2. Det vil sige, at der i beskrivelsen af den nuværende værdi anvendes RV-værdier fra de ni foregående dage, hvilket sammenlignet med modelleringen af VIX indekset er en dag mindre.

Koefficienterne fra AR(9)-modellen samt de tilhørende standardfejl, t-værdier og p-værdier ses i Tabel 6.4.

	Koefficienter	Standardafvigelse	t-værdi	p-værdi
Skæring	0,0001	3,1e-05	3,388	0,001*
AR 1	0,360	0,015	23,925	2,2e-16*
AR 2	0,321	0,016	20,067	2,2e-16*
AR 3	-0,018	0,017	-1,098	0,272
AR 4	0,110	0,017	6,586	4,5e-11*
AR 5	0,061	0,017	3,610	3,2e-04*
AR 6	-0,075	0,017	-4,481	7,4e-06*
AR 7	-0,021	0,017	-1,241	0,215
AR 8	-0,018	0,016	-1,095	0,273
AR 9	0,178	0,015	11,812	2,2e-16*

Tabel 6.4. Koefficienter for AR(9) for RV. P-værdierne under 0,05 er markeret med *.

Den justerede R^2-værdi for AR(9)-modellen er 0,601, hvilket betyder, at 60% af variationen i RV indekset forklares ud fra modellen. Sammenlignet med den tidligere AR-model for VIX indekset, er dette en betydeligt højere forklaringsgrad, eftersom AR-modellen kun forklarede 27,4% af variationen i VIX indekset. Figur 6.5 viser de fittede værdier for RV indekset som den røde kurve, samt de sande værdier som den sorte kurve.



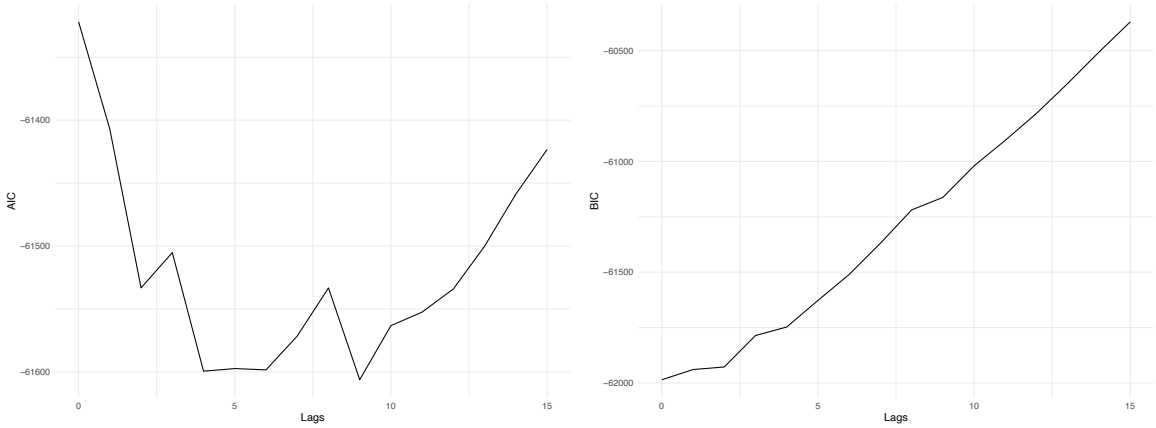
Figur 6.5. Fittede værdier for AR(9)-modellen for RV indekset.

Heraf ses, at modellen opfanger de fleste af udsvingene i RV indekset, hvilket også inkluderer udsvinget i 2020, som skyldes coronapandemien. Dog opfanger modellen ikke størrelsen af alle udsvingene, hvilket for eksempel ses i 2008 samt 2015. I næste afsnit udvides modellen, sådan at de 22 Google Trends variable indgår som eksterne regressorer.

6.1.1 Autoregressiv distributed lag model for RV

For at undersøge om Google Trends variablene kan forbedre forklaringsgraden, altså om klimaforandringerne kan bidrage til at forklare RV indekset, udvides AR-modellen til også at indeholde de 22 Google Trends variable.

På Figur 6.6 ses værdierne for AIC samt BIC plottet mod antallet af lags på de eksterne regressorer. Minimumsværdien for AIC er ved $q = 9$, mens BIC værdierne stiger med antallet af lags. Derfor vælges den optimale værdi for parametereren q i ADL-modellen til at være 9, hvormed en ADL(9,9)-model opstilles for RV indekset, så der medtages ni laggede værdier af Google Trends variablene i modellen.

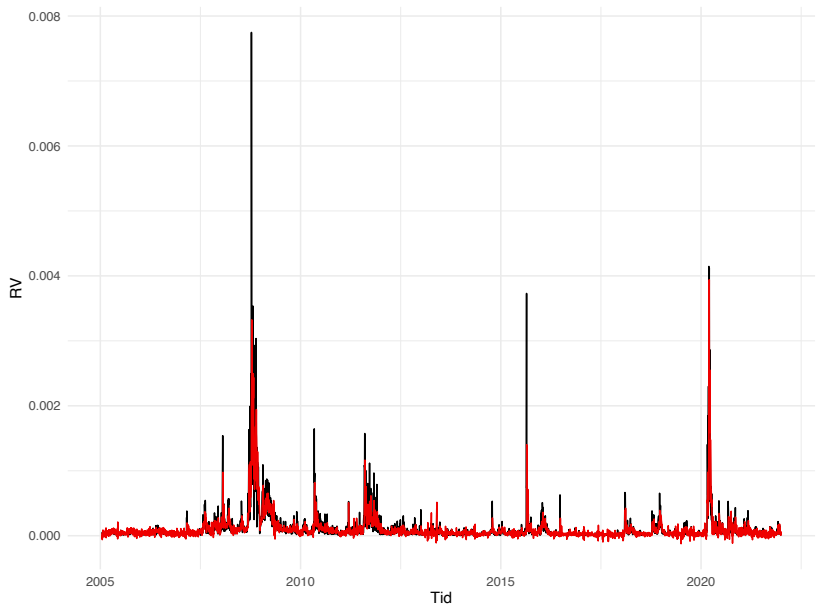


(a) AIC mod antal lags for en ADL-model for RV. **(b)** BIC mod antal lags for en ADL-model for RV.

Figur 6.6

Det blev i modelleringen af VIX indekset undersøgt, om skæringen bidrog til fittet, eftersom koefficienten for denne var insignifikant. Givet R^2-værdien for modellen, hvor denne var undladt, kunne det konkluderes, at skæringen bidrog. Dette blev dog ikke nævnt i afsnittet, hvor modellen blev opstillet for VIX, da det ikke førte til ændringer i modellen. Det undersøges nu, om skæringen i ADL(9,9)-modellen for RV indekset bidrager til fittet, eftersom koefficienten for skæringen i denne model også er insignifikant. De justerede R^2-værdier for modellerne med og uden skæring er henholdsvis 0,643 og 0,685, hvilket vil sige, at modellen uden skæring er bedre til at forklare VIX indekset. Af denne grund fortsættes analysen med modellen uden skæring, og koefficienterne, standardafvigelsen, t-værdierne samt p-værdierne for denne ADL-model kan ses i Tabel B.7 og B.8 i Appendiks B. Af p-værdierne ses, at størstedelen af parametrene for RV indekset er insignifikante. Nogle af de signifikante Google Trends variable er *Pandemic* til lag 1,2,3,7,8, 9, samt *Global Warming* til lag 1.

Figur 6.7 viser de fittede værdier for ADL(9,9)-modellen som den røde kurve samt det sande RV indeks som den sorte kurve. Sammenlignet med de fittede værdier for AR(9)-modellen på Figur 6.5 opfanger modellen nu størrelsen af udsvinget i RV indekset i 2020 tydeligere, men stadig ikke størrelsen af udsvingene i 2008 samt 2015. Det er formentlig tilfældet, fordi udsvinget i 2020 skyldes coronapandemien, som modellen nu har information om og tager højde for givet Google Trends variablene. Modellen har som tidligere nævnt en justeret R^2-værdi på 0,685, hvormed modellen forklarer 69% af variationen i RV indekset. Dette er også en klar forbedring i forhold til AR(9)-modellen som forklarede 60%, hvormed Google Trends variablene har bidraget til at forklare variationen i RV indekset.



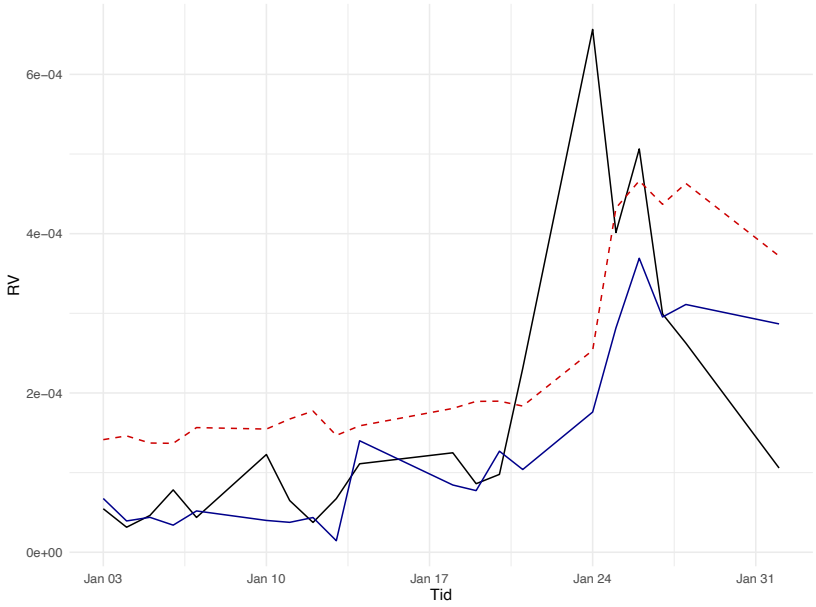
Figur 6.7. Fittede værdier for ADL-model for RV.

Det undersøges nu, om det vil forbedre modellen at inddrage de nutidige værdier af Google Trends variablene, det vil sige de eksterne regressorer til tid t. Den optimale model, som indeholder de nutidige værdier af Google Trends variablene, er en ADL(9,4), hvormed den indeholder laggede værdier fra lag 0 til 4, og denne model har en AIC-værdi på -61598, mens ADL(9,9)-modellen uden nutidige værdier har en AIC-værdi på -61608. Det vil sige, at AIC-værdien er lavest for ADL(9,9)-modellen uden de nutidige værdier. Derudover har modellen med de nutidige værdier en justeret R^2-værdi på 0,678 hvilket er lavere end modellen uden, hvormed den nutidige værdi ikke bidrager til forklaringen af RV indekset. Dette medfører også, at når vi i næste afsnit vil prædiktere RV indekset, kan dette gøres uden også at skulle prædiktere de 22 Google Trends variable.

6.1.2 Prædiktioner

I dette afsnit vurderes modellernes præcision out-of-sample ved at prædiktere nye værdier en måned frem, det vil sige alle hverdage i januar 2022. Prædiktionerne er beregnet på samme vis som for VIX indekset i Afsnit 5.2.2, og præcisionen af disse vurderes ved at sammenligne størrelsen af prædiktionsfejlene for de to modeller for RV indekset.

På Figur 6.8 er de sande værdier for RV indekset vist som den sorte kurve, og de prædikterede værdier med AR(9)- og ADL(9,9)-modellen er vist som henholdsvis den røde og blå kurve. På baggrund af disse forecasts beregnes MSE- samt MAE-værdier for de to modeller. AR(9)-modellen har en MSE-værdi på 2,021e-08 samt en MAE-værdi på 0,0001, mens ADL(9,9)-modellen har en MSE-værdi på 1,656e-08 og en MAE-værdi på 7,248e-05. Baseret på figuren samt prædiktionsfejlene er ADL(9,9)-modellen mest præcis i beskrivelsen af RV indekset out-of-sample, hvilket kan tyde på, at Google Trends variablene har bidraget til modellens prædiktionsevne. Det skal dog bemærkes, at AR-modellen i modsætning til ADL-modellen indeholder en skæring, hvilket gør, at denne models prædiktioner ligger højere, eftersom der tillægges en konstant på 0,0001, hvilket tilsyneladende forringer modellens prædiktioner.



Figur 6.8. Sande værdier for RV (sort), prædikterede værdier for RV med AR(10) (rød) samt udvidet ADL(9,9) (blå).

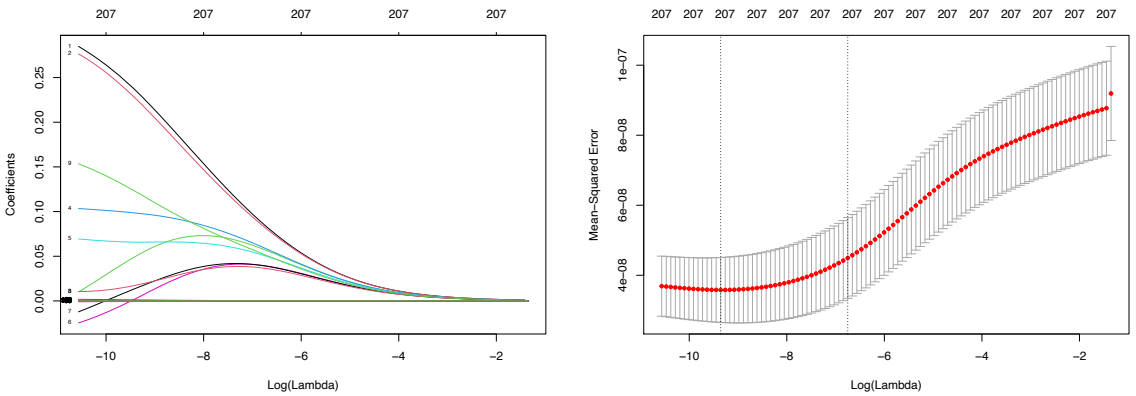
Prædiktionsfejlene kan sammenlignet med dem for VIX indekset virke små, men dette skyldes, at skalaen i de to volatilitetsindekser er meget forskellig. På samme måde som ved modelleringen af VIX indekset undersøges det nu, om shrinkage og dimensions reduktion metoderne fra Kapitel 4 kan forbedre resultaterne.

6.2 Shrinkage

I dette afsnit udføres shrinkage på ADL-modellen for realiseret varians, hvilket gøres for at mindske påvirkningen fra irrelevante variable. Analyserne er udført på baggrund af teorien i Afsnit 4.1 på samme vis som for VIX indekset i Afsnit 5.3.

6.2.1 Ridge regression for RV

I dette afsnit udføres Ridge regression på ADL(9,9)-modellen for RV indekset med `R`-pakken `glmnet`. Figur 6.9a viser koefficienterne for Ridge regression mod logaritmen til λ, hvor det ses, at koefficienterne krympes mod 0, når λ stiger.



(a) Ridge regression koefficienter for RV.

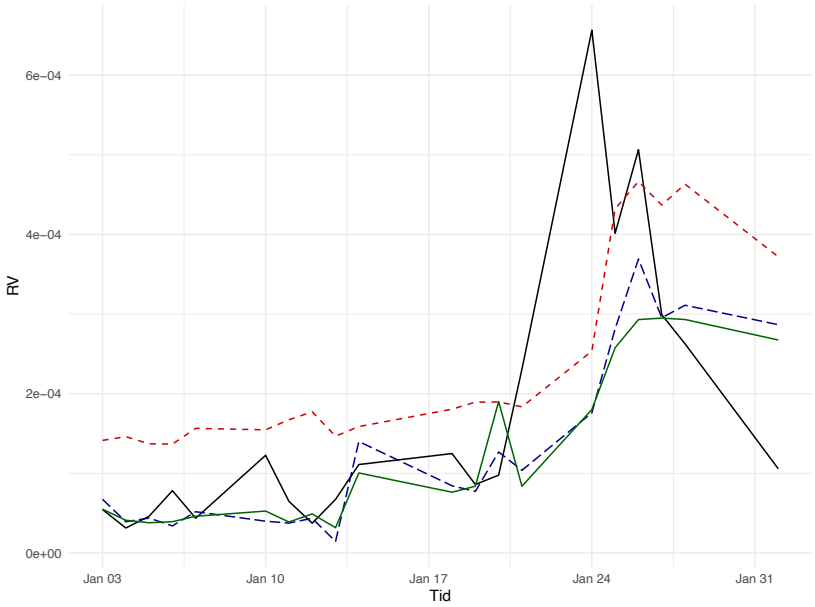
(b) Ridge regression MSE for RV.

Figur 6.9

På Figur 6.9b er MSE-værdierne for Rigde regressionen plottet mod logaritmen af den tilhørende λ værdi, og modellen med den mindste fejl opnås, når $\lambda = 8{,}656\text{e-}05$. Med dette valg af tuningparameter opnås en Ridge regression model, med koefficienter, standardafvigelser, t-værdier samt p-værdier vist i Tabel B.9 samt Tabel B.10 i Afsnit B.2.2 i Appendiks B. Heraf ses, at det udelukkende er AR parametrene til lag 1,2,3,4,5 og 9 samt *Pandemic* til lag 1, der er signifikante. Modellen har en justeret R^2-værdi på 0,579, hvilket betyder, at modellen forklarer 58% af variationen i RV indekset, hvilket er lavere end forklaringsgraden for AR(9) samt ADL(9,9)-modellen. Dette skyldes formentlig, at man går på kompromis med bias i forsøget på at reducere variansen ved brug af shrinkage metoderne. Desuden gælder der, at den justerede R^2-værdi straffer for antallet af prædiktorerer i modellen, og da Ridge regression ikke udfører variabel selektion, indgår alle 207 variable stadig i modellen på trods af, at mange af koefficienterne er insignifikante og dermed ikke bidrager særligt til fittet.

Med F-test testes om de eksterne regressorer i Ridge regressionen bidrager sammen til et bedre fit af RV indekset. Det gøres på tilsvarende måde som for VIX indekset med en F-test, hvoraf resultatet er en F-værdi på 0,121, hvilket er lavere end den kritiske værdi, hvormed nulhypotesen ikke forkastes. På baggrund af dette, vurderes det, at de ekstra parametre tilsammen ikke bidrager signifikant til fittet. Det samme gjorde sig gældende i forbindelse med VIX indekset, hvilket kan skyldes, at der i Ridge regression introduceres bias i koefficienterne for at kunne opnå bedre prædiktioner.

Figur 6.10 viser forecastet af RV indekset, hvor prædiktionerne med Ridge regression er vist som den grønne kurve.

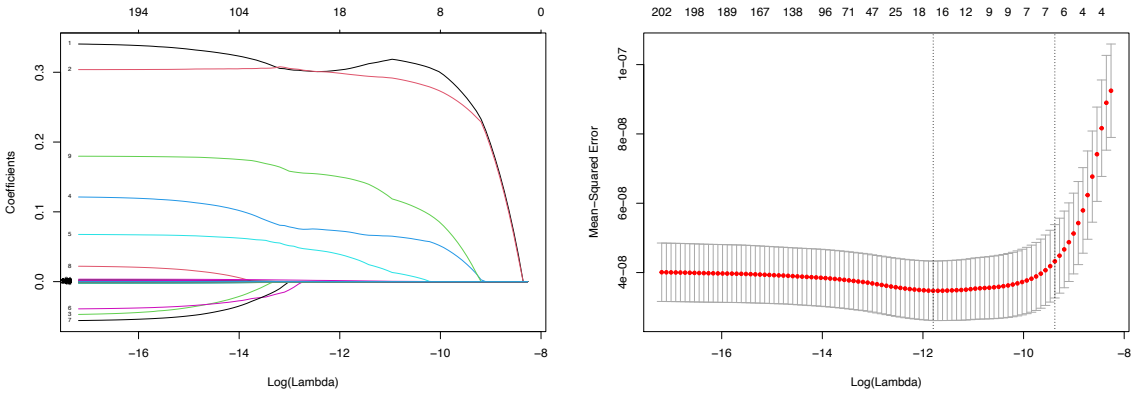


Figur 6.10. Sande værdier for RV (sort), prædikterede værdier for RV med AR(10) (rød), med ADL(9,9) (blå), samt med Ridge regression på ADL(9,9) (grøn).

Disse prædiktioner sammenlignes med de forecastede værdier for AR(9)-modellen, som er vist ved den røde kurve, samt ADL(9,9)-modellen, som er vist som den blå kurve. Prædiktionerne med Ridge regression har en MSE-værdi på 1,807e-08 og en MAE-værdi på 7,667e-08. Sammenlignet med prædiktionsfejlene fra AR(9)-modellen er denne model mere præcis out-of-sample, dog er ADL(9,9)-modellen stadig den mest præcise model out-of-sample.

6.2.2 Lasso for RV

I dette afsnit udføres Lasso på ADL-modellen for RV indekset. Figur 6.11a viser et plot af koefficienterne, når Lasso metoden anvendes på ADL(9,9)-modellen for den realiserede varians for S&P500. Her ses, at den del af koefficienterne, som ikke i forvejen er meget tæt på 0, langsomt nærmer sig 0, når λ stiger. Til forskel fra Ridge regression bliver koefficienterne præcis 0, når λ er stor nok.



(a) Lasso koefficienter for RV. **(b)** Lasso MSE for RV.

Figur 6.11

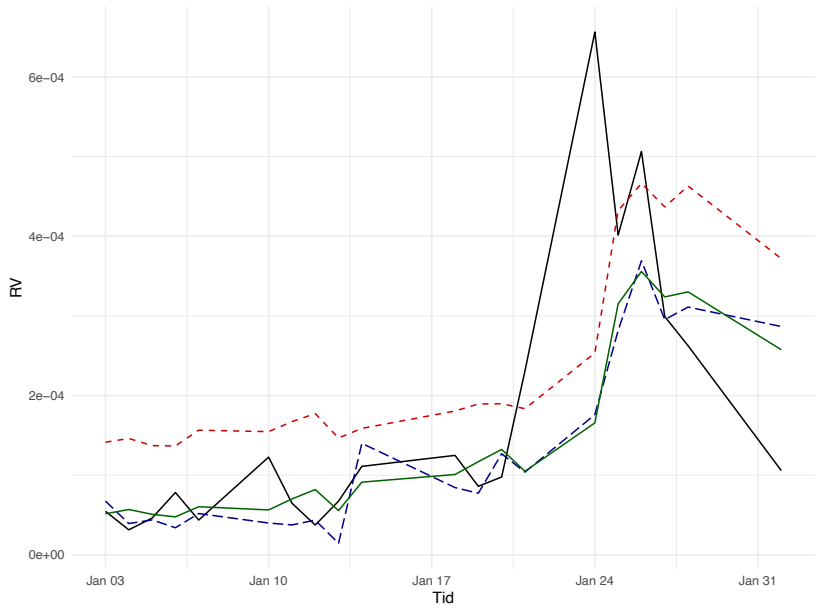
På Figur 6.11b ses MSE-værdierne plottet mod logaritmen til λ, hvor den laveste MSE værdi opnås ved λ = 7,529e-06. Til denne værdi af λ er der 16 koefficienter, som ikke krympes til at være præcis 0. Lasso modellen har en justeret R^2-værdi på 0,609, hvormed modellen forklarer 61% af variationen i RV indekset. Dermed opnås en højere forklaringsgrad ved at anvende Lasso metoden fremfor Ridge regression, eftersom denne havde en justeret R^2-værdi på 0,579. Dette skyldes højest sandsynligt, at den justerede R^2-værdi som tidligere nævnt straffer for antallet af prædiktorer, og eftersom Lasso udfører variabel selektion, er der kun 16 tilbageværende prædiktorer, og dermed er der ikke lige så mange insignifikante variable med i modellen, som der var i Ridge regression modellen. Sammenlignet med AR(9)-modellen, som havde en forklaringsgrad på 60%, har Lasso modellen altså en højere forklaringsgrad, men ADL(9,9)-modellen har stadig den højeste forklaringsgrad på 69%.

Koefficienter for de tilbageværende variable i modellen kan ses i Tabel 6.12. Tabellen viser desuden standardafvigelser, t-værdier og p-værdier, som er estimeret med bootstrap metoden på samme vis som for VIX indekset i Afsnit 5.3.2. Her opdeles datasættet i 284 forskellige datasæt, hvor der i hver er udeladt en blok bestående af 15 sammenhængende observationer, og der opstilles en model på hvert datasæt. Af tabellen ses som tidligere, at *Pandemic* er signifikant i forhold til at beskrive RV indekset, men også at Lasso vælger at beholde blandt andet *Avalanche* til tre lags i beskrivelsen af RV indekset.

	Koefficienter	Standardafvigelser	t-værdi	p-værdi
AR 1	0,306	6,68e-03	45,733	3,21e-68*
AR 2	0,297	5,91e-03	50,225	4,80e-72*
AR 4	7,09e-02	5,20e-03	13,637	1,13e-24*
AR 5	3,48e-02	7,78e-03	4,473	1,04e-05*
AR 9	0,147	6,76e-03	21,806	1,11e-39*
Global Warming 1	2,10e-05	5,40e-06	3,886	9,26e-05*
Pandemic 1	1,32e-03	2,32e-04	5,696	6,42e-08*
Avalanche 3	8,46e-06	4,89e-06	1,729	4,35e-02*
Avalanche 4	1,55e-05	4,03e-06	3,852	1,05e-04*
Tornado 4	9,29e-05	5,12e-05	1,817	3,62e-02*
Pandemic 5	-1,36e-04	6,23e-05	-2,187	0,984
Landslide 6	2,25e-05	1,40e-05	1,607	5,56e-02
Pandemic 6	-1,55e-04	7,85e-05	-1,975	0,975
Tornado 7	8,77e-05	1,54e-05	5,707	6,14e-08*
Avalanche 9	4,58e-06	3,64e-06	1,260	0,105
Pandemic 9	-7,55e-04	1,33e-04	-5,673	1,000

Tabel 6.12. Tilbageværende koefficienter for Lasso for RV samt deres standardafvigelser, t-værdier og p-værdier. P-værdier under 0,05 er markeret med *.

Figur 6.13 viser prædiktionerne med Lasso modellen som den grønne kurve. Disse prædiktioner sammenlignes med AR(9) samt ADL(9,9)-modellernes, som er vist som henholdvis den røde og den blå kurve. Prædiktionerne med Lasso modellen har en MSE-værdi på 1,633e-08 samt en MAE-værdi på 7,077e-05, hvilket gør Lasso til den mest præcise model out-of-sample. Generelt set har alle modellerne dog relativt ens prædiktionsfejl på den udvalgte out-of-sample periode.



Figur 6.13. Sande værdier for RV (sort), prædikterede værdier for RV med AR(9) (rød), med ADL(9,9) (blå), samt med Lasso på ADL(9,9) (grøn).

6.3 Dimensions reduktion

I dette afsnit udføres de tilsvarende dimension reduktions analyser for RV, som blev lavet for VIX indekset i Afsnit 5.4. Dette gøres for at undersøge, om dimensions reduktion af Google Trends variablene bidrager til en model med højere præcision in-sample og out-of-sample.

6.3.1 Principal components analyse for RV

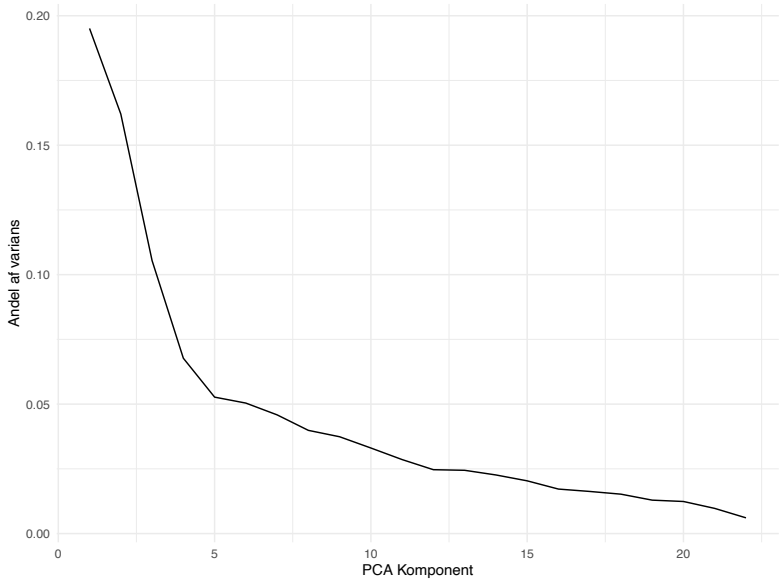
Ved brug af kommandoen `prcomp` i `R` pakken `stats` udføres PCA analyse på RV indekset, og resultatet heraf kan ses i Tabel 6.14. Af tabellen ses, at den første principal komponent, PC1, forklarer 19,5% af variansen i datasættet. Dermed kan 19,5% af informationen i datasættet opfanges med blot en enkelt principal komponent. Sammenlignet med PCA analysen for VIX indekset i Afsnit 5.4.1 er dette dog ikke lige så højt, eftersom den første principal komponent her forklarede 25% af variationen i datasættet. Af den kummulative andel ses, at PC1 og PC2 tilsammen forklarer 35,7% af variansen i datasættet. Grundet at der er 11 færre observationer i RV indekset end i VIX indekset opnås ikke de samme principal komponenter, eftersom datasættene med Google Trends variablene er forskellige.

Komponent	Andel af varians	Kumulativ andel
PC1	0,195	0,195
PC2	0,162	0,357
PC3	0,105	0,463
PC4	0,068	0,530
PC5	0,053	0,583
PC6	0,050	0,633
PC7	0,046	0,679
PC8	0,039	0,719
PC9	0,037	0,757
PC10	0,033	0,789
PC11	0,029	0,818
PC12	0,025	0,843
PC13	0,024	0,867
PC14	0,023	0,889
PC15	0,020	0,910
PC16	0,017	0,927
PC17	0,016	0,944
PC18	0,015	0,958
PC19	0,013	0,972
PC20	0,012	0,984
PC21	0,010	0,994
PC22	0,006	1,000

Tabel 6.14. PCA analyse for RV indekset.

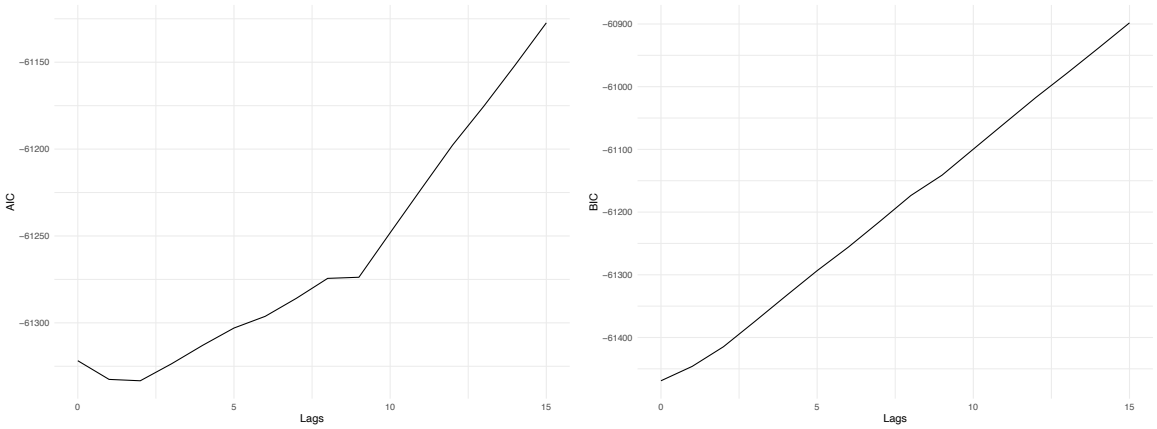
For at udvælge antallet af principal komponenter, der skal indgå som eksterne regressorer i ADL-modellen for RV indekset, plottes hver af komponenternes bidrag til beskrivelse af variansen i indekset, hvilket er vist på Figur 6.15. Det ses, at de første komponenter bidrager en del, hvorefter andelen falder, og det ønskes at anvende så få PCA komponenter som muligt til at beskrive mest mulig af variationen i Google Trends datasættet. Derfor vurderes forskellige modeller, der opstilles med stigende antal

PCA kompoenenter. Heraf fås resultatet, at modellen med seks PCA komponenter er den model med færrest komponenter, hvor AIC-værdierne angiver, at modellen bør inddrage PCA komponenter som eksterne regressorer. Modellerne med færre PCA komponenter har AIC-værdier som antyder, at AR(9)-modellen uden eksterne regressorer er at foretrække. Derfor vurderes det mest optimale antal af komponenter til seks, hvormed vi får forklaret 63% af variationen i Google Trends datasættet.



Figur 6.15. Andel af variansen forklaret af hver komponent.

Med de seks principal komponenter i ADL-modellen skal antallet af lags, q, vælges. Dette gøres som tidligere med informationskriterierne AIC og BIC, hvilke er vist på Figur 6.16, hvor værdierne er plottet mod antallet af lags af PCA komponenterne i modellen. Af Figuren ses, at BIC værdierne blot stiger med antallet af lags, men betragtes AIC-værdierne ses, at minimum findes ved at vælge to lags af PCA komponenterne. Derfor vælges $q = 2$, således at en ADL(9,2)-model opstilles med seks PCA komponenter som eksterne regressorer.



(a) AIC-værdier for valg af antal lags, q, af de seks principal komponenter i ADL-modellen for RV.

(b) BIC-værdier for valg af antal lags, q, af de seks principal komponenter i ADL-modellen for RV.

Figur 6.16

Koefficienterne, deres standardfejl, t-værdier samt p-værdier for ADL(9,2)-modellen med seks PCA komponenter er vist i Tabel 6.17. Modellen har en justeret R^2-værdi på 0,649. Af p-værdierne ses, at det udelukkende er PC5 komponenten til lag 1, samt PC6 komponenten til begge lags, som er signifikante. Dette kunne forklare, hvorfor det var nødvendigt med seks PCA komponenter i modellen, for at den

på baggrund af AIC-værdierne var at foretrække fremfor AR(9)-modellen. Derudover skal det igen bemærkes, at PCA komponenterne beregnes på baggrund af variationen i Google Trends datasættet, og vælges derfor ikke ud fra, hvad der forklarer variationen i RV indekset bedst. Det kan være derfor at de første PCA komponenter, som beskriver den største variation i Google Trends datasættet, ikke nødvendigvis er dem, som er signifikante i forhold til at forklare RV indekset. Derudover er der lav korrelation i datasættet for Google Trends variablene, hvilket gør, at PCA komponenterne er konstrueret sådan, at de har stor vægt på én af parametrene, og derfor ikke bidrager væsentligt til modelleringen af RV.

	Koefficient	Standardfejl	t-værdi	p-værdi
AR 1	0,361	1,51e-02	23,951	5,72e-119*
AR 2	0,326	1,61e-02	20,284	2,36e-87*
AR 3	-2,11e-02	1,68e-02	-1,259	0,212
AR 4	0,110	1,68e-02	6,584	5,13e-11*
AR 5	6,57e-02	1,68e-02	3,903	9,64e-05*
AR 6	-7,41e-02	1,68e-02	-4,420	1,01e-05*
AR 7	-2,27e-02	1,68e-02	-1,351	0,187
AR 8	-1,58e-02	1,61e-02	-0,983	0,326
AR 9	0,181	1,51e-02	12,000	1,24e-32*
PC1 1	5,59e-05	1,28e-04	0,436	0,662
PC2 1	-3,40e-05	7,06e-05	-0,482	0,630
PC3 1	-1,73e-04	1,02e-04	-1,698	8,95e-02
PC4 1	1,15e-04	1,01e-04	1,139	0,255
PC5 1	-2,16e-04	1,07e-04	-2,011	4,43e-02*
PC6 1	4,84e-04	1,07e-04	4,523	6,26e-06*
PC1 2	-2,14e-05	1,28e-04	-0,167	0,867
PC2 2	2,25e-05	7,06e-05	0,319	0,750
PC3 2	1,39e-04	1,02e-04	1,362	0,173
PC4 2	-6,78e-05	1,01e-04	-0,669	0,503
PC5 2	1,79e-04	1,07e-04	1,672	9,46e-02
PC6 2	-2,94e-04	1,07e-04	-2,744	6,10e-03*

Tabel 6.17. Koefficienter for ADL(9,2)-modellen for RV med PCA komponenter som eksterne regressorer. P-værdier under 0,05 er markeret med *.

Med F-test testes om PCA komponenterne simultant bidrager til fittet af RV indekset. Modellen testes som tidligere mod modellen uden de ekstra komponenter, hvilket vil sige AR(9)-modellen. Af testen fås en F-værdi 1,913 på samt en p-værdi på 0,033, hvilket betyder, at nulhypotesen forkastes, og det vurderes, at PCA komponenterne sammen er signifikante i beskrivelsen af RV indekset.

Rotationsmatricen for de seks PCA komponenter ses i Tabel 6.18, hvilken viser vægtene for hver Google Trends variabel, som indgår i linearkombinationerne i PCA komponenterne. Heraf ses eksempelvis, at den første PCA komponent lægger størst vægt på variablen *Global Warming*. Denne variabel har ikke tidligere vist sig at være betydelig til at forklare variationen i RV indekset, og dette kan derfor være årsagen til at PC1 komponenten ikke er signifikant i ADL(9,2) modellen. PC6 komponenten er den komponent, som indeholder den største vægt af *Pandemic*, som tidligere har vist sig at bidrage til beskrivelsen af volatilitetsindekserne, og det er muligvis derfor, at denne komponent til begge lags er signifikant i ADL(9,2)-modellen.

	PC1	PC2	PC3	PC4	PC5	PC6
Wildfire	0,019	-0,010	0,183	0,245	-0,813	-0,408
Avalanche	0,055	0,068	-0,111	-0,065	0,116	0,053
Blizzard	-0,019	0,007	-0,017	-0,023	0,006	-0,008
Climate Change	-0,058	0,645	-0,220	0,050	-0,064	0,189
Drought	0,333	0,278	0,846	-0,228	0,067	0,171
Dust Storm	-0,016	-0,007	0,007	0,002	-0,015	0,025
Earthquake	-0,068	0,057	0,006	0,061	-0,017	0,094
Flash Flood	-0,067	0,039	0,138	0,545	0,174	0,082
Flood	-0,039	0,044	0,131	0,574	0,264	-0,066
Global Warming	0,914	-0,011	-0,285	0,190	0,021	-0,064
Hailstorm	-0,051	0,008	0,018	-0,041	0,011	-0,024
Heat Wave	-0,086	-0,002	0,165	0,205	-0,075	-0,008
Landslide	-0,005	0,003	-0,014	-0,012	0,023	-0,016
Mudslide	-0,009	-0,001	0,001	-0,007	0,026	-0,002
Pandemic	-0,109	0,033	-0,056	0,077	-0,155	0,404
Sea Level Rise	-0,037	0,689	-0,148	-0,046	-0,073	-0,215
Sinkhole	-0,089	0,072	0,078	-0,124	0,392	-0,723
Thunderstorm	-0,055	0,038	0,093	0,219	0,024	0,065
Tornado	-0,031	0,057	0,015	0,075	0,046	0,031
Tropical Cyclone	-0,026	0,062	0,016	0,306	0,139	-0,045
Tsunami	0,003	-0,012	-0,015	0,009	0,016	0,002
Volcanic Eruption	0,004	0,079	-0,064	-0,028	0,059	0,032

Tabel 6.18. Rotationsmatrix for PCA-komponenter for RV.

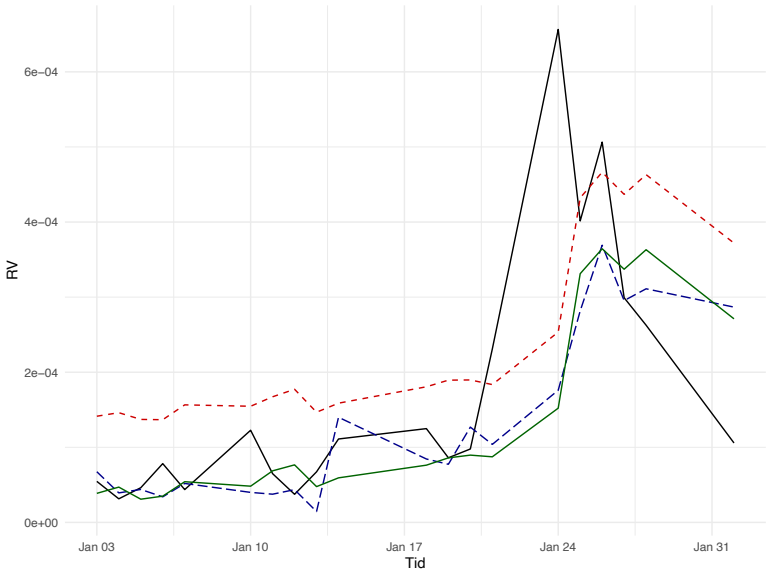
På samme vis som for VIX indekset kan vi gange koefficienterne i rotationsmatricen for hver PCA komponent med den respektive koefficient i ADL-modellen, sådan at der opnås en koefficient for hvert lag af Google Trends variablene fremfor lags af PCA komponenterne. På den måde opnås et udtryk på samme form som den oprindelige ADL-model med alle 22 Google Trends variable som eksterne regressorer. Koefficienterne opnået med PCA metoden ses i Tabel 6.19, mens de kan sammenlignes med koefficienterne i B.7 og B.8 i Appendiks B. I tabellen angiver tallet bag hver Google Trend variable antal lags, så *Wildfire 2* angiver denne variabel til tid $t-2$. Det ses, at AR-koefficienterne og koefficienterne for *Pandemic* er relativt store, hvilket også var tilfældet med den oprindelige ADL-model. Desuden er koefficienterne for *Sinkhole* også relativt store i forhold til de resterende koefficienter, hvilket ikke gjorde sig gældende i forbindelse med den oprindelige ADL-model. Det betyder, at *Sinkhole* formentlig er betydningsfuld i beskrivelsen af variationen i Google Trends datasættet, men ikke så relevant i forbindelse med RV indekset.

	Koefficient		Koefficient		Koefficient
AR 1	0,361	Global Warming 1	8,69e-05	Dust Storm 2	-8,94e-06
AR 2	0,326	Hailstorm 1	-2,47e-05	Earthquake 2	-3,10e-05
AR 3	-2,12e-02	Heat Wave 1	2,81e-06	Flash Flood 2	-8,43e-06
AR 4	0,110	Landslide 1	-1,23e-05	Flood 2	4,79e-05
AR 5	6,57e-02	Mudslide 1	-7,95e-06	Global Warming 2	-4,95e-05
AR 6	-7,41e-02	Pandemic 1	2,40e-04	Hailstorm 2	1,54e-05
AR 7	-2,27e-02	Sea Level Rise 1	-9,35e-05	Heat Wave 2	-3,48e-07
AR 8	-1,58e-02	Sinkhole 1	-4,69e-04	Landslide 2	8,00e-06
AR 9	0,181	Thunderstorm 1	3,11e-05	Mudslide 2	5,97e-06
Wildfire 1	-2,42e-05	Tornado 1	7,23e-06	Pandemic 2	-1,56e-04
Avalanche 1	1,27e-05	Tropical Cyclone 1	-2,28e-05	Sea Level Rise 2	4,89e-05
Blizzard 1	-6,08e-06	Tsunami 1	1,56e-06	Sinkhole 2	3,06e-04
Climate Change 1	1,24e-04	Volcanic Eruption 1	8,11e-06	Thunderstorm 2	-1,47e-05
Drought 1	-9,50e-05	Wildfire 2	-1,76e-05	Tornado 2	-1,79e-06
Dust Storm 1	1,36e-05	Avalanche 2	-5,15e-06	Tropical Cyclone 2	2,15e-05
Earthquake 1	4,92e-05	Blizzard 2	3,13e-06	Tsunami 2	-6,55e-07
Flash Flood 1	3,62e-05	Climate Change 2	-8,51e-05	Volcanic Eruption 2	-4,05e-06
Flood 1	-4,92e-05	Drought 2	9,38e-05		

Tabel 6.19. Koefficienterne for Google Trends variablene til hvert lag opnået med PCA for RV.

I modellen med PCA komponenter er der udelukkende anvendt lags af PCA komponenterne, og derfor er det interessant at undersøge om PCA komponenterne til samme tidspunkt, tid t, kan bidrage til at beskrive RV indekset. Hvis denne eksterne regressor også tilføjes til modellen med PCA komponenter, fås en anelse højere AIC-værdi og en anelse lavere forklaringsgrad i form af den justerede R^2-værdi, hvormed modellen med den nutidige værdi af PCA komponenterne ikke er bedre in-sample. Derudover ville modellen være vanskelig at anvende til forecast, eftersom det ville kræve, at Google Trends variablene bliver prædikteret inden prædiktionen af RV indekset. Af disse grunde anvendes modellen, hvor de nutidige værdier af PCA komponenterne er inkluderet, ikke yderligere i projektet.

Figur 6.20 viser PCA-modellens forecast af RV indekset henover alle hverdage i januar 2022 som den grønne linje, mens den røde er den oprindelige AR(9)-models prædiktioner og den blå er ADL(9,9)-modellens.



Figur 6.20. Sande værdier for RV (sort), prædikterede værdier for RV med AR(9) (rød), med ADL(9,9) (blå), samt med PCA ADL(9,2) (grøn).

Heraf ses, at forecastet minder om ADL(9,9)-modellens, og MSE prædiktionsfejlen er 1,769e-08, mens MAE-værdien er 7,554e-05, hvilke ikke er lavere end prædiktionsfejlene fra Lasso modellen i forrige afsnit, som derfor stadig er den bedste model out-of-sample.

I PCA tages kun højde for Google Trends variablene, når komponenterne konstrueres, og et alternativ til denne metode er som tidligere PLS, som netop tager højde for responsvariablen under konstruktionen. Denne metode anvendes i det følgende afsnit.

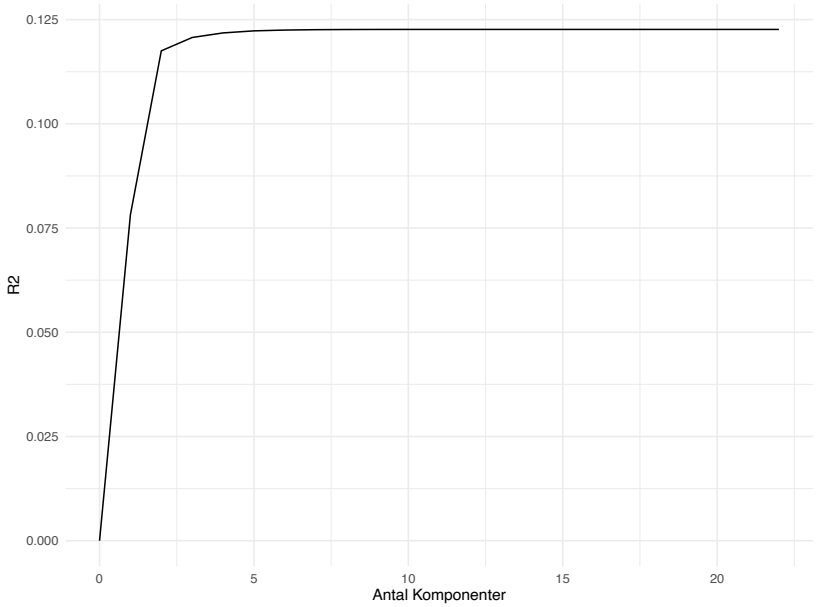
6.3.2 Partial least squares regression for RV

Tilsvarende Afsnit 5.4.2 udføres i dette afsnit PLS regression for RV indekset. Ved brug af kommandoen `plsr` i `R` udføres PLS regressionen på RV indekset, og resultatet af dette ses i Tabel 6.21. Tabellen viser for hver af de 22 komponenter, hvor stor en andel af variationen, de forklarer af både de forklarende variable samt responsvariablen, hvor andelen for responsvariablen er angivet kumulativt.

Komponent	Andel af varians X	Kumulativ andel af varians RV
Comp 1	0,144	0,078
Comp 2	0,113	0,118
Comp 3	0,118	0,121
Comp 4	0,079	0,122
Comp 5	0,077	0,122
Comp 6	0,038	0,123
Comp 7	0,031	0,123
Comp 8	0,038	0,123
Comp 9	0,042	0,123
Comp 10	0,028	0,123
Comp 11	0,035	0,123
Comp 12	0,031	0,123
Comp 13	0,028	0,123
Comp 14	0,017	0,123
Comp 15	0,023	0,123
Comp 16	0,031	0,123
Comp 17	0,023	0,123
Comp 18	0,023	0,123
Comp 19	0,023	0,123
Comp 20	0,018	0,123
Comp 21	0,022	0,123
Comp 22	0,019	0,123

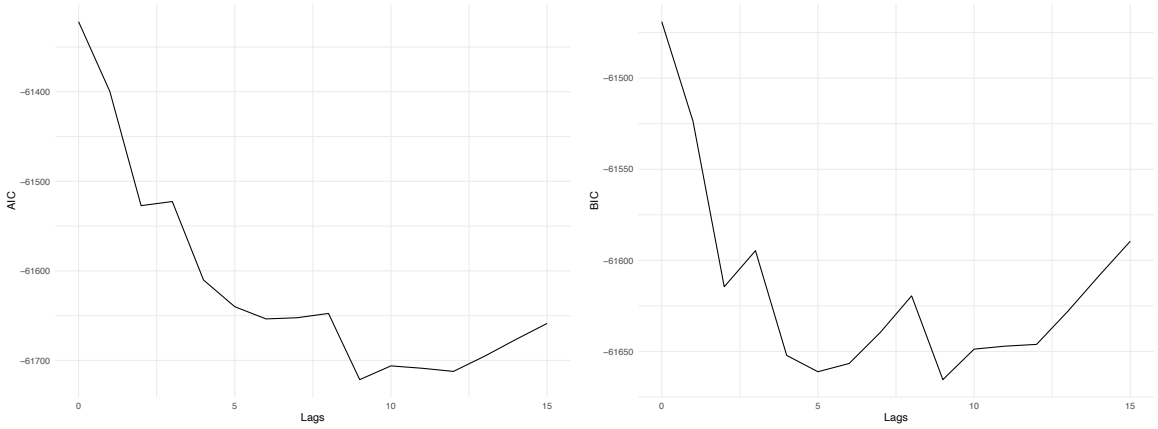
Tabel 6.21. PLS analyse for RV indekset.

For at afgøre antallet af PLS komponenter, der skal indgå som eksterne variable i ADL-modellen for RV, plottes R^2-værdierne for PLS mod antallet af komponenter. Dette er vist på Figur 6.22, hvor der ses et knæk i kurven for 2-3 komponenter. Af denne grund undersøges ADL-modeller med henholdsvis to og tre PLS komponenter som eksterne regressorer, og af dette fås, at AIC-værdien er lavest for modellen med tre komponenter. Eftersom der ikke ses en signifikant forbedring i R^2-værdien ved at øge antallet af komponenter yderligere, udvælges de tre første PLS komponenter til at indgå i ADL-modellen som eksterne regressorer.



Figur 6.22. PLS R^2-værdier for RV.

Antallet af lags af PLS komponenterne i ADL-modellerne vælges på baggrund af AIC- og BIC-værdierne. Figur 6.23 viser disse værdier plottet mod antallet af lags i modellen, hvoraf minimumsværdien angiver den optimale lagorden. Begge informationskriterier minimeres ved $q = 9$, hvormed der opnås en ADL(9,9)-model, hvor der indgår ni lags af de tre PLS komponenter. Den justerede R^2-værdi for denne model er 0,682.



(a) AIC mod antal lags for en PLS-model for RV. **(b)** BIC mod antal lags for en PLS-model for RV.

Figur 6.23

Koefficienterne, deres standardfejl, t-værdier samt p-værdier for ADL(9,9) modellen med tre PLS komponenter ses i Tabel 6.24. Det ses, at en relativ stor del af koefficienterne for PLS komponenterne er signifikante sammenlignet med koefficienterne for PCA komponenterne i ADL(9,2)-modellen. Dette kan skyldes, at PLS metoden også betragter responsvariablen, RV indekset, i konstruktionen af komponenterne. Disse komponenter beskriver RV indekset bedre end PCA komponenterne gjorde, eftersom PCA komponenterne udelukkende blev konstrueret til at forklare variationen i Google Trends datasættet. Modellen har som tidligere nævnt en justeret R^2-værdi på 0,682, hvilket gør den til den anden mest præcise model i forklaringen af RV indekset.

	Koefficient	Standardfejl	t-værdi	p-værdi
AR 1	0,337	1,50e-02	22,414	3,44e-105*
AR 2	0,315	1,59e-02	19,786	2,37e-83*
AR 3	-4,00e-02	1,68e-02	-2,386	1,71e-02*
AR 4	0,121	1,68e-02	7,238	5,38e-13*
AR 5	7,21e-02	1,69e-02	4,269	2,00e-05*
AR 6	-4,41e-02	1,68e-02	-2,627	8,64e-03*
AR 7	-4,58e-02	1,68e-02	-2,726	6,44e-03*
AR 8	1,26e-02	1,59e-02	0,789	0,430
AR 9	0,189	1,48e-02	12,762	1,22e-36*
Comp1 1	1,25e-03	1,67e-04	7,479	9,05e-14*
Comp2 1	1,55e-03	1,51e-04	10,265	2,02e-24*
Comp3 1	4,78e-04	9,70e-05	4,930	8,55e-07*
Comp1 2	-7,84e-04	1,98e-04	-3,955	7,78e-05*
Comp2 2	-1,14e-03	1,94e-04	-5,858	5,04e-09*
Comp3 2	-2,92e-04	1,15e-04	-2,541	1,11e-02*
Comp1 3	5,09e-04	2,00e-04	2,543	1,10e-02*
Comp2 3	7,98e-04	1,98e-04	4,027	5,74e-05*
Comp3 3	3,67e-04	1,15e-04	3,181	1,48e-03*
Comp1 4	-3,96e-04	2,00e-04	-1,983	4,74e-02*
Comp2 4	-4,71e-04	2,00e-04	-2,380	1,74e-02*
Comp3 4	-1,75e-04	1,16e-04	-1,514	0,130
Comp1 5	-2,48e-04	2,00e-04	-1,238	0,216
Comp2 5	1,85e-05	1,98e-04	9,34e-02	0,926
Comp3 5	-6,29e-05	1,15e-04	-0,546	0,585
Comp1 6	1,51e-04	1,99e-04	0,757	0,449
Comp2 6	-3,04e-04	1,98e-04	-1,539	0,124
Comp3 6	-2,08e-04	1,15e-04	-1,797	7,23e-02
Comp1 7	-3,06e-04	1,99e-04	-1,534	0,125
Comp2 7	-3,07e-04	1,98e-04	-1,549	0,121
Comp3 7	-1,85e-05	1,15e-04	-0,161	0,873
Comp1 8	4,24e-04	1,98e-04	2,147	3,19 e-02*
Comp2 8	7,40e-04	1,95e-04	3,804	1,44e-04*
Comp3 8	1,80e-04	1,15e-04	1,572	0,116
Comp1 9	-5,43e-04	1,68e-04	-3,239	1,21e-03*
Comp2 9	-8,76e-04	1,52e-04	-5,754	9,32e-09*
Comp3 9	-2,65e-04	9,73e-05	-2,721	6,55e-03*

Tabel 6.24. Koefficienter for ADL(9,9)-modellen for RV med PLS komponenter som eksterne regressorer. P-værdier under 0,05 er markeret med *.

Selvom det ikke er alle PLS komponenter, der individuelt er signifikante, undersøges det, om de tilsammen er signifikante og samlet bidrager til beskrivelsen af RV indekset. Det gøres som tidligere med F-test, hvoraf resultatet er en F-værdi på 18,172 og en p-værdi på 2,2e-16, hvilket betyder, at PLS komponenterne sammen bidrager til fittet af RV indekset.

Tabel 6.25 viser rotationsmatricen for PLS komponenterne. Det vil sige, at tabellen viser vægtene der anvendes til at konstruere hver PLS komponent. Her ses, at i den første komponent har *Pandemic* størst vægt, og desuden er vægten for *Global Warming* relativt stor. Dette gør sig også gældende i den anden komponent, hvor vægten for *Pandemic* er endnu større end for komponent 1, og i den tredje komponent er *Drought* den variabel med størst vægt. At vægten er fordelt sådan i de tre komponenter stemmer overens med tidligere resultater om, at *Pandemic* har en indflydelse på volatilitetsindekset RV, og at de andre variable med relativt store vægte formenligt er bedst til at beskrive variationen i Google Trends datasættet. På den måde beskrives variationen i både respons- og de forklarende variable ved PLS komponenterne.

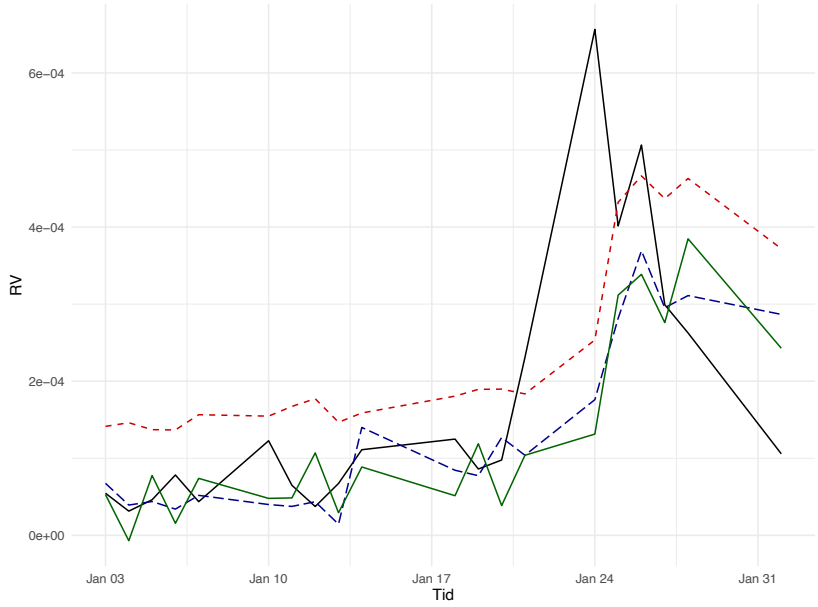
	Comp1	Comp2	Comp3
Wildfire	-0,047	-0,048	0,106
Avalanche	0,071	0,056	0,102
Blizzard	0,003	0,037	0,081
Climate Change	-0,101	0,020	0,049
Drought	-0,021	-0,114	0,792
Dust Storm	-0,003	-0,007	-0,138
Earthquake	0,047	0,141	-0,039
Flash Flood	-0,078	-0,005	0,028
Flood	-0,106	-0,045	0,242
Global Warming	0,560	-0,329	-0,028
Hailstorm	-0,117	-0,105	-0,059
Heat Wave	-0,129	-0,056	0,013
Landslide	-0,003	0,004	0,029
Mudslide	-0,041	-0,059	-0,102
Pandemic	0,743	1,180	0,460
Sea Level Rise	-0,171	-0,074	0,290
Sinkhole	-0,146	-0,035	0,217
Thunderstorm	-0,072	-0,035	-0,114
Tornado	0,003	0,060	0,069
Tropical Cyclone	-0,095	-0,081	0,031
Tsunami	-0,008	-0,025	-0,061
Volcanic Eruption	-0,009	-0,022	-0,069

Tabel 6.25. Rotationsmatrix for PLS-komponenter for RV.

For at sammenligne koefficienterne fra ADL(9,9)-modellen i Tabel B.7 og B.8 med koefficienterne fra PLS modellen, omskrives koefficienterne på samme vis som tidligere. Dette gøres ved at gange koefficienterne for hver PLS komponent, som er vist i rotationsmatricen, med den respektive koefficient i ADL-modellen med PLS komponenter fra Tabel 6.24. De omskrevne koefficienter kan ses i Tabel B.11 og B.12 i Afsnit B.2.3 i Appendiks B. Det kan af tabellerne ses, at koefficienterne for PLS ADL(9,9)-modellen og den oprindelige ADL(9,9)-model har forholdsvis ens størrelsesforhold, hvormed de to modeller ikke adskiller sig væsentligt fra hinanden.

Derudover er det ved at betragte den justerede R^2-værdi samt AIC-værdien undersøgt, om det ville bidrage til modellen at inkludere de nutidige værdier af PLS komponenterne. Dog forbedres ingen af disse værdier ved at inddrage de nutidige værdier i modellen, hvorfor denne model ikke betragtes yderligere.

På Figur 6.26 ses de forecastede værdier med ADL(9,9)-modellen med PLS komponenterne som den grønne kurve. Derudover viser den sorte kurve de sande værdier for RV indekset, mens den røde og den blå kurve viser de tidligere prædikterede værdier med henholdsvis AR(9)- og ADL(9,9)-modellen, og det ses, at prædiktionerne med PLS modellen lægger sig tæt op af prædiktionerne fra ADL(9,9)-modellen. Det skal igen bemærkes, at AR(9)-modellen i modsætning til de andre modeller indeholder en positiv skæring, som medfører, at prædiktionerne ligger højere. Prædiktionerne med ADL(9,9)-modellen med PLS komponenterne har en MSE-værdi på 1,961e-08 og en MAE-værdi på 8,714e-05. På trods af at modellen med PLS komponenterne er mere præcis end modellen med PCA komponenter in-sample, baseret på de justerede R^2-værdier, kan det ses af prædiktionsfejlene, at PCA modellen er mere præcis out-of-sample.



Figur 6.26. Sande værdier for RV (sort), prædikterede værdier for RV med AR(9) (rød), med ADL(9,9) (blå), samt med PLS ADL(9,9) (grøn).

I næste afsnit vil RV indekset blive modelleret med neurale netværk for at undersøge, om resultater fra disse modeller viser samme tendenser. Derudover ønskes det også at undersøge, om klimaforandringerne kan bidrage til forklaringen af RV indekset ved brug af neurale netværk.

6.4 Neurale Netværk

I dette afsnit opstilles to neurale netværk for RV indekset på samme vis som de blev opstillet for VIX indekset i Afsnit 5.5. Dette gøres i `R` ved brug af pakken `neuralnet`, hvor vi anvender TanH som aktiveringsfunktion. Netværkene opstilles på en tilsvarende træningsmængde som den for VIX indekset, der består af observationer fra juni 2015 til maj 2020, hvilket udgør 1253 observationer. Derudover anvendes testmængden, som består af de resterende observationer frem til december 2021.

For at kunne vurdere om Google Trends variablene bidrager til fittet og forecastet med neurale netværk opstilles først et neuralt netværk, der udelukkende er baseret på tidligere værdier af RV, hvorefter det udvides til også at have Google Trends variable som input.

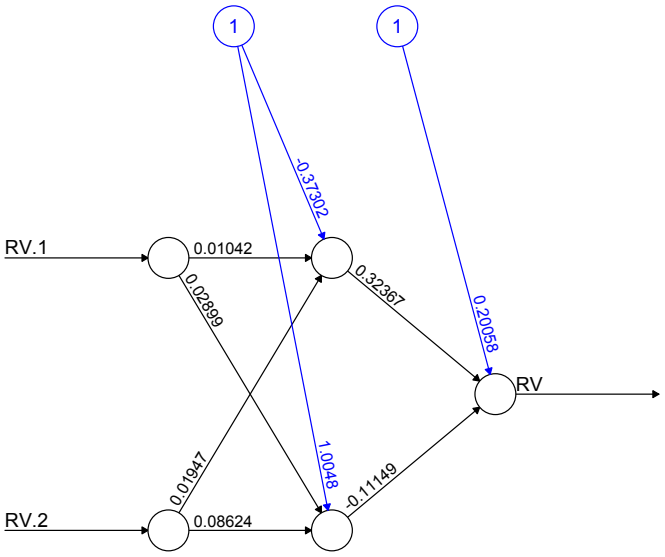
6.4.1 Neuralt netværk med RV som input

I dette afsnit opstilles et neuralt netværk for RV indekset, som udelukkende indeholder tidligere værdier af RV indekset som inputvariable. For at bestemme det optimale antal af laggede værdier af RV indekset, der skal indgå i netværket, opstilles en række forskellige netværk, som vurderes in-sample ved AIC og out-of-sample ved MSE og MAE på testmængden. Disse værdier kan ses i Tabel 6.27. I tabellen er de laveste værdier i hver statistik markeret med en *, og det ses heraf, at modellen med to RV lags samt to knuder i det skjulte lag har de laveste værdier out-of-sample, mens modellen med ét lag af RV og én knude i det skjulte lag har lavest AIC-værdi. Dog vægtes out-of-sample resultater højere, sådan at vi undgår en model, der overfitter observationerne, hvormed det førstnævnte netværk udvælges til at beskrive RV indekset, og denne model er markeret med fed i tabellen.

	AIC	MSE	MAE
1 RV lag			
d=1	8*	3,68e-09	3,87e-05
d=2	14	3,75e-09	3,84e-05
d=3	20	4,86e-09	4,21e-05
2 RV lags			
d=1	10	3,71e-09	3,65e-05
d=2	**18**	**3,58e-09***	**3,63e-05***
d=3	26	3,63e-09	3,64e-05
3 RV lags			
d=1	12	3,83e-09	3,70e-05
d=2	22	4,32e-09	4,10e-05
d=3	32	4,12e-09	3,83e-05
5 RV lags			
d=1	16	3,69e-09	3,67e-05
d=2	30	4,05e-09	3,82e-05
d=3	44	5,28e-09	4,37e-05
10 RV lags			
d=1	26	3,77e-09	3,72e-05
d=2	50	3,97e-09	4,33e-05
d=3	774	4,73e-09	4,11e-05

Tabel 6.27. Resultater for forskellige neurale netværk for RV, der kun er baseret på tidligere værdier af indekset. De laveste værdier er markeret med *, og den udvalgte model er markeret med fed.

Det udvalgte neurale netværk har en AIC-værdi på 18,00, og en illustration af netværket er vist på Figur 6.28.



Figur 6.28. Neuralt netværk for RV indekset med to lags af RV som inputvariable, samt to knuder i det skjulte lag.

For at vurderer om vi kan opnå et bedre fit af RV indekset udvides det neurale netværk nu til også at indeholde de 22 Google Trends variable som inputvariable.

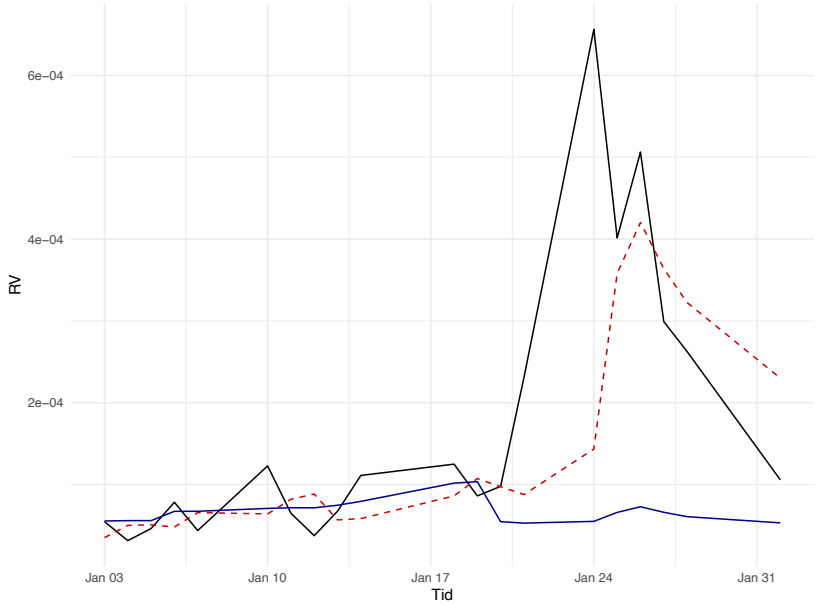
6.4.2 Udvidet neuralt netværk

I dette afsnit udvides det neurale netværk som før nævnt til også at indeholde Google Trends variable som input. Tabel B.13 i Afsnit B.2.4 i Appendiks B viser resultaterne for en række neurale netværk, hvor antal lags af RV, antal lags af Google Trends variablene samt antal knuder i det skjulte lag varieres. Her er der 45 forskellige modelkombinationer, som alle er opstillet for at bestemme den optimale orden. De mindste AIC-værdier og prædiktionsfejl er som tidligere markeret med *, og her ses, at der er tre forskellige modeller, der har bedste værdier. Derfor betragtes hvilken model, der generelt er bedst over alle tre parametre, og dette vurderes til at være modellen med fem RV lags, to Google Trends lags samt én knude i det skjulte lag. Denne model er markeret med fed i tabellen. Som det ses af Tabel B.13 har den valgte model en AIC-værdi på 148. Dette er en klar forværring sammenlignet med det neurale netværk uden Google Trends variablene, som havde en AIC-værdi på 18. Det har derfor forværret modellens præcision in-sample at inkludere information om klimaforandringerne i form af Google Trends variablene. Vægtene for den udvalgte model er vist i Tabel B.14 i Afsnit B.2.4 i Appendiks B, og det er som tidligere nævnt vanskeligt at analysere på vægtene, idet beregningerne i de skjulte lag er som en black box.

I næste afsnit anvendes de to neurale netværk til at prædiktere RV indekset i januar 2022, så prædiktionsevnerne for de to opstillede neurale netværk kan sammenlignes indbyrdes, men også sammenlignes med alle de foregående modeller for RV indekset.

6.4.3 Prædiktioner af RV med neurale netværk

Det neurale netværk opstillet i Afsnit 6.4.1 kaldes fremover Netværk 1, mens det udvidede netværk fra Afsnit 6.4.2 kaldes Netværk 2. Som tidligere prædikteres RV for alle hverdage i januar 2022, hvor værdierne kan ses på Figur 6.29. Prædiktionerne med Netværk 1 kan ses som den røde kurve, mens prædiktionerne fra Netværk 2 kan ses som den blå kurve. Netværk 1 har en MSE-værdi på 1,648e-08 samt en MAE-værdi 6,901e-05, mens Netværk 2 har værdier på henholdsvis 4,005e-08 og 1,161e-04. Det vil sige, at prædiktionerne med det neurale netværk, som også inddrager Google Trends variablene som inputvariable er mindre præcise, hvilket også ses tydeligt på figuren. Det vil sige, at klimaforandringerne i form af Google Trends variable ikke bidrager hverken in-sample eller out-of-sample til modelleringen af RV indekset, når neurale netværk anvendes.



Figur 6.29. Prædikterede værdier med Netværk 1 (rød) samt Netværk 2 (blå) for RV indekset.

I næste afsnit opsamles der på resultaterne fra de opstillede modeller for RV indekset ved at sammenligne modellernes MSE-, MAE- samt justerede R^2-værdier for de autoregressive modeller og AIC-værdier for de neurale netværk.

6.5 Opsamling på resultaterne for RV

Tabel 6.30 viser resultaterne fra de opstillede modeller for RV-indekset. Igen markerer * de bedste værdier, og i forbindelse med fittet af RV indekset har ADL(9,9)-modellen den højeste justerede R^2-værdi. Dette viste sig også at være tilfældet i forbindelse med modelleringen af VIX-indekset, hvor ADL(10,4)-modellen havde højest justeret R^2-værdi. Dog præsterer ADL-modellen med ni lags af de 22 Google Trends variable ikke bedst i forbindelse med prædiktionerne, hvor ADL Lasso-modellen har den laveste MSE-værdi, mens MAE-værdien er lavest for Netværk 1.

Model	Justeret R^2	Forecast MSE	Forecast MAE
AR(9)	0,601	2,021e-08	0,0001
ADL(9,9)	0,685*	1,656e-08	7,248e-05
ADL Ridge	0,579	1,807e-08	7,667e-05
ADL Lasso	0,609	1,633e-08*	7,077e-05
ADL PCA	0,649	1,769e-08	7,554e-05
ADL PLS	0,682	1,961e-08	8,714e-05
	AIC	Forecast MSE	Forecast MAE
Neuralt netværk 1	18*	1,648e-08	6,901e-05*
Neuralt netværk 2	148	4,005e-08	1,161e-04

Tabel 6.30. Resultater for alle modellerne for RV indekset, det bedste resultat i hver statistik er markeret med *.

Når de justerede R^2-værdier betragtes ses, at der sker en forbedring i forklaringsgraden ved at inddrage Google Trends observationerne, som indikatorer for klimaforandringer i de autoregressive modeller. Forklaringsgraden stiger fra 60% med AR-modellen til 69% med ADL-modellen, hvorved data for klimaforandringerne har bidraget til at beskrive den realiserede varians for S&P500 indekset. I forbindelse med F-tests blev det observeret ved modellerne, at i Ridge var de eksterne regressorer tilsammen insignifikante, mens det modsatte gjorde sig gældende i forbindelse med PCA og PLS. Dette understøttes af, at PCA og PLS modellerne har højere justerede R^2-værdier end AR(9)-modellen uden Google Trends variablene.

Ved brug af neurale netværk kan det ses, at det forringer modellernes præcision at indrage Google Trends variablene både in-sample og out-of-sample. Det vil sige, at i neurale netværk bidrager klimaforandringerne ved hjælp af Google Trends variable ikke til at beskrive den realiserede varians for S&P500 indekset.

Ved at sammenligne resultaterne for RV og VIX indekset ses, at in-sample har det i de autoregressive modeller hjulpet til forklaringen af indekserne at inddrage Google Trends variablene. Også out-of-sample har klimaforandringerne bidraget til at opnå mere præcise forecast, idet MSE- og MAE-værdierne for både RV og VIX indekset er lavest for en af de udvidede ADL-modeller. I forbindelse med de neurale netværk har det forværret resultaterne in-sample for RV indekset at inddrage klimaforandringerne i modelleringen, på trods af at det for VIX indekset gav bedre resultater. Out-of-sample har det i de neurale netværk for både RV og VIX indekset ikke bidraget positivt at inddrage klimaforandringer, idet prædiktionerne med de neurale netværk, som har haft Google Trends variable som input, blev forringet.

7 Diskussion

I dette kapitel diskuteres de anvendte metoder i projektet. Desuden vil der perspektiveres til hvilke andre valg af metoder, der kunne have været anvendt, så resultaterne i projektet kunne være blevet forbedret.

Til at beskrive klimaforandringerne har vi i dette projekt benyttet Google Trends data for blandt andet naturkatastrofer, hvilket hovedsageligt afspejler befolkningens interesse og bekymring for fænomenerne. Det kan diskuteres hvorvidt Google Trends datasættet også afspejler hyppigheden af forekomsten af naturkatastrofer, hvor der muligvis ville have været andet data eller indekser, som bedre ville kunne beskrive det. Dog er befolkningens usikkerhed omkring klimaforandringerne relevant, eftersom volatilitetindekserne også afspejler usikkerheden i det finansielle marked, hvorfor det giver mening at modellere dem med Google Trends data. I udvælgelsen af Google Trends variablene kan det diskuteres, om *Pandemic* kan betragtes som værende en naturkatastrofe, men eftersom mange af klimaforandringernes konsekvenser øger risikoen for pandemier, har vi valgt at inkludere denne.

For at kunne vurdere om klimaforandringerne har bidraget, skal det afgøres, hvorvidt målet med modellen er at forklare eller prædiktere volatilitetsindeksene. Generelt er tendensen, at Google Trends variablene bidrager til forklaringen, eftersom der generelt opnås en forbedring i de justerede R^2-værdier. Dette kræver dog ofte mere komplicerede modeller, som eksempelvis ADL(9,9)-modellen, hvor der var 198 koefficienter, der tilhørte Google Trends variablene, hvilket eksempelvis besværliggør modelfortolkningen. Forbedringen i prædiktionsfejlene er minimale, når modellerne udvides, og der opstår dermed et trade-off mellem en højere forklaringsgrad og en lavere kompleksitet. Google Trends variablene har dermed bidraget til at forklare allerede observerede udsving, men generelt er det vanskeligt at prædiktere volatiliteten.

I projektet er lagordenen af de eksterne regressorer valgt til at være ens for alle variable. Det kunne muligvis have bidraget positivt, hvis vi havde betragtet forskellig lagorden af de forskellige Google Trends variable. Så ville vi formentlig have fået simplere modelkonstruktioner, idet modellerne måske kun krævede mange lags af én af de 22 Google Trends variable, som for eksempel *Pandemic*, for at opnå det bedste fit. Det kunne også ses af koefficienttabellerne, at flere lags var insignifikante, og at modellen muligvis kun medtog dem, fordi et højere lag var signifikant. Det ville muligvis kunne forbedre resultaterne, hvis vi havde fjernet disse insignifikante variable, men af tidsmæssige årsager blev dette ikke foretaget. Derudover blev der af tidsmæssige årsager heller ikke udført F-test på Lasso modellerne, eftersom det ikke kunne udføres på samme vis som ved de andre modeller, da testen bør tage højde for, at vi allerede har udført variabel selektion.

Vi kunne have inddraget de nutidige værdier i modelleringen, hvilket også blev undersøgt, men det bidrog ikke signifikant in-sample. Dog kunne vi have undersøgt, om prædiktionerne af volatilitetsindekserne blev forbedret, hvis disse også blev inddraget. Vi ville dog ikke forvente, at prædiktionerne blev forbedret, eftersom de da ville blive baseret på forecasts af de 22 Google Trends variable, hvilket ville gøre prædiktionerne mere usikre.

I forbindelse med neurale netværk kunne det have været en fordel at opstille recurrent netværk, idet de er fordelagtige i analysen af tidsrækkedata, fordi korrelationen i observationerne da bevares bedre. Dog blev dette ikke gennemført af tidsmæssige årsager. I de neurale netværk, som blev opstillet, indgik

laggede værdier af det respektive volatilitetsindeks, dog var antallet af lags forskellige i netværkene med og uden Google Trends variable som input. Dette skyldes, at hvert netværk er valgt ud fra, hvilket der var det mest optimale. Det gør dog, at de er vanskelige at sammenligne direkte, når de ikke har samme antal lags af volatilitetsindekserne som input.

I modelleringen af begge volatilitetsindekser så vi en tendens til, at det oftest var variablen *Pandemic*, som var signifikant, og havde store koefficenter sammenlignet med de andre variable. Idet vi betragter volatiliteten for det amerikanske marked, så indikerer det, at kun globale naturkatastrofer har en indvirkning. Hvis vi i stedet for kun at betragte klimaforandringer også havde medtaget Google Trends data for andre store kriser i modelleringen, som eksempelvis finanskrisen i 2008, kunne vi formentlig have forklaret en større andel af variationen i indekserne. Hvis vi ikke havde lavet modelleringen over årene 2020-2021, så coronapandemien ikke indgik i datasættet, ville resultatet om, at Google Trend datasættet bidrager til forklaringen af volatiliteten muligvis have været anderledes. Såfremt tendensen med klimaforandringer fortsætter, og de får en endnu større indflydelse for befolkningen på globalt plan, vil Google Trends datasættet formentligt fremadrettet kunne være med til at forklare mere af volatiltieten i det finansielle marked.

Konklusion 8

På baggrund af teori fra tidsrækkeanalyse og machine learning metoder har vi i dette projekt opstillet modeller til at beskrive volatillitetsindekser for S&P500 indekset. Dette indebærer autoregressive modeller, som er udvidet til at indeholde eksterne regressorer, hvor shrinkage og dimensions reduktions metoder er anvendt på disse, samt neurale netværk. Dette har vi gjort for at undersøge, om volatiliteten i det finansielle marked bliver påvirket af klimaforandringerne. Et af resultaterne fra analysen er, at for både VIX og RV indekset opnås de højeste justerede R^2-værdier i forbindelse med modeller, hvor Google Trends variablene indgår som eksterne regressorer. Heraf kan det overordnet konkluderes, at Google Trends datasættet, som beskriver klimaforandringerne, bidrager til forklaringen af VIX og RV indekserne in-sample, og til beskrivelsen af de udsving, som der tidligere har været. Dette vil sige, at evidensen i projektet viser, at folks bekymringer om klimaforandringer, i form af de søgeord vi har udvalgt, påvirker det finansielle marked. Dog er de udvidede modeller, som indeholder Google Trends variablene, væsentlig mere komplekse end modellerne uden de eksterne variable, hvormed der opstår et trade-off i forhold til kompleksiteten, og hvor meget de ekstra variable bidrager til forklaringsgraden. Igennem projektet blev prædiktionerne, som blev lavet med de forskellige modeller for hvert volatilitetsindeks, overordnet set ikke forbedret væsentligt ved de udvidede modeller, og der kan dermed sættes tvivl om den forøgede modelkompleksitet er det værd, hvis man har til henseende at anvende modellerne til prædiktioner.

Succesen af de autoregressive modeller, som blev opstillet med machine learning metoder, afhænger af, at der er korrelation imellem variablene i datasættet, hvilket det i projektet blev konkluderet, at der generelt set ikke var. Derfor har modellerne opstillet med disse metoder ikke haft de rette betingelser til at kunne bidrage på optimal vis til modelleringen.

Af analyserne i projektet kan det konkluderes, at det ofte har været variablen *Pandemic*, som var signifikant i modelleringen af volatilitetsindeksene. Hertil skal det bemærkes, at vi har betragtet volatiliteten på det amerikanske marked i form af S&P500 indekset, og at det derudfra kan konkluderes, at det kræver globale naturkatastrofer før volatiliteten her påvirkes.

Litteratur

Aggarwal, **2015**. Charu C. Aggarwal. *Data Mining*. Springer, 2015. ISBN 978-3-319-14141-1.

Baillie, **1996**. Richard T. Baillie. *Long memory processes and fractional integration in econometrics*. Journal of econometrics, 1996.

Bobeica og Bojesteanu, **2008**. Gabriel Bobeica og Elena Bojesteanu. *Long Memory in Volatility. An Investigation on the Central and Eastern European Exchange Rates*. European Research Studies, 2008.

Care-Danmark. Care-Danmark. *Konsekvenser af klimaforandringer*. URL: https://care.dk/det-goer-vi/klima/konsekvenser-af-klimaforandringer. Sidst besøgt d. 29. maj 2022.

Cboe, **2022**. Cboe. *Cboe VIX FAQ*, 2022. URL: https://www.cboe.com/tradable_products/vix/faqs/. Sidst besøgt d. 3. marts 2022.

Chan. Harvard T.H. Chan. *Coronavirus and Climate Change*. URL: https://www.hsph.harvard.edu/c-change/subtopics/coronavirus-and-climate-change/. Sidst besøgt d. 17. maj 2022.

Choi og Varian, **2012**. Hyunyoung Choi og Hal Varian. *Predicting the Present with Google Trends*. The Economic Record, 2012.

Davison og MacKinnon., **2009**. Russell Davison og James G. MacKinnon. *Econometrics Theory and Methods*. Oxford University Press, 2009.

Fidelity-International. Fidelity-International. *Understanding stock market volatility and how it could help you*. URL: https://www.fidelity.com.sg/beginners/your-guide-to-stock-investing/understanding-stock-market-volatility-and-how-it-could-help-you. Sidst besøgt d. 18. maj 2022.

Friedman et al., **2021**. Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon og James Yang. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2021. URL: https://CRAN.R-project.org/package=glmnet. R package version 4.1-3, Sidst besøgt d. 1. maj 2022.

Fritsch et al., **2019**. Stefan Fritsch, Frauke Guenther, Marvin N. Wright, Marc Suling og Sebastian M. Mueller. *neuralnet: Training of Neural Networks*, 2019. URL: https://CRAN.R-project.org/package=neuralnet. R package version 1.44.2, Sidst besøgt d. 17. maj 2022.

Ghosh og Bouri, **2022**. Bikramaditya Ghosh og Elie Bouri. *Long Memory and Fractality in the Universe of Volatility Indices*. Wiley, 2022.

Google. Google. *FAQ about Google Trends data*. URL: https://support.google.com/trends/answer/4365533?hl=en. Sidst besøgt d. 16. februar 2022.

Google-Trends. Google-Trends. *Google Trends*. URL: https://trends.google.com/trends/?geo=DK. Sidst besøgt d. 28. februar 2022.

Géron, 2019. Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2019.

Haldrup og Valdés, 2015. Niels Haldrup og J. Eduardo Vera Valdés. *Long Memory, Fractional Integration, and Cross-Sectional Aggregation*. Aarhus University og CREATES, 2015.

Hyndman, 2022. Rob Hyndman. *forecast: Forecasting Functions for Time Series and Linear Models*, 2022. URL: https://CRAN.R-project.org/package=forecast. R package version 8.16, Sidst besøgt d. 10 maj 2022.

Ivanovs et al., 2021. Maksims Ivanovs, Roberts Kadikis og Kaspars Ozols. *Perturbation-based methods for explaining deep neural networks: A survey*. ELSEVIER, 2021. URL: https://www.sciencedirect.com/science/article/pii/S0167865521002440.

James et al., 2017. Gareth James, Daniela Witten, Trevor Hastie og Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017. ISBN 978-1-4614-7137-0.

Koopman et al., 2005. *Siem Jan Koopman, Borus Jungbacker og Eugenie Hol*. Journal of Empircal Finance, 2005.

Liland et al., 2021. Kristian Hovde Liland, Bjørn-Helge Mevik, Ron Wehrens og Paul Hiemstra. *pls: Partial Least Squares and Principal Component Regression*, 2021. URL: https://CRAN.R-project.org/package=pls. R package version 2.8-0, Sidst besøgt d. 28. april 2022.

Maitra og Yan, 2008. Saikat Maitra og Jun Yan. *Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression*. Casualty Actuarial Society, 2008. URL: https://www.casact.org/sites/default/files/database/dpp_dpp08_08dpp76.pdf.

Massicotte og Eddelbuettel, 2022. Philippe Massicotte og Dirk Eddelbuettel. *gtrendsR: Perform and Display Google Trends Queries*, 2022. URL: https://CRAN.R-project.org/package=gtrendsR. R package version 1.5.0, Sidst besøgt d. 3. marts 2022.

Nasdaq, 2022. Nasdaq. *SPX Historical Data*, 2022. URL: https://www.nasdaq.com/market-activity/index/spx/historical. Sidst besøgt d. 28. februar 2022.

Oxford-Man-Institute, 2022. Oxford-Man-Institute. *Oxford-Man Institute of Quantitative Finance, Realized Library*, 2022. URL: https://realized.oxford-man.ox.ac.uk/data/download. Sidst besøgt d. 25. marts 2022.

Pfaff et al., 2016. Bernhard Pfaff, Eric Zivot og Matthieu Stigler. *urca: Unit Root and Cointegration Tests for Time Series Data*, 2016. URL: https://CRAN.R-project.org/package=urca. R package version 1.3-0, Sidst besøgt d. 15. april 2022.

Rahimikia og Poon, 2021. *Eghbal Rahimikia og Ser-Huang Poon*. SSRN Electronic Journal, 2021.

Sankar, 2021. Aadhithya Sankar. *Principal Component Analysis Part 1: The Different Formulations*, 2021. URL: https://towardsdatascience.com/principal-component-analysis-part-1-the-different-formulations-6508f63a5553. Sidst besøgt d. 15. marts 2022.

Shumway og Stoffer, 2017. Robert H. Shumway og David S. Stoffer. *Time Series Analysis and Its Application with R Examples*. Springer, 2017. ISBN 978-3-319-52452-8.

Sowell, 1992. Fallaw Sowell. *Maximum likelihood estimation of stationary univariate fractionally integrated time series models*. Journal of Econometrics, 1992.

Torelli et al., 2013. *Nicola Torelli, Fortunato Pesarin og Avner Bar-Hen*. Springer, 2013. ISBN 978-3-642-35587-5.

Union, 2022. Den Europæriske Union. *Klimaforandringer*, 2022. URL: https://europa.eu/climate-pact/about/climate-change_da. Sidst besøgt d. 28. februar 2022.

Veenstra, 2022. JQ Veenstra. *arfima: Fractional ARIMA (and Other Long Memory) Time Series Modeling*, 2022. URL: https://CRAN.R-project.org/package=arfima. R package version 1.8-0, Sidst besøgt d. 14. april 2022.

Whaley, 2009. Robert E. Whaley. *Understanding the VIX*. 2009.

Xiong et al., 2016. Ruoxuan Xiong, Eric P. Nichols og Yuan Shen. *Deep Learning Stock Volatility with Google Domestic Trends*. The Economic Record, 2016.

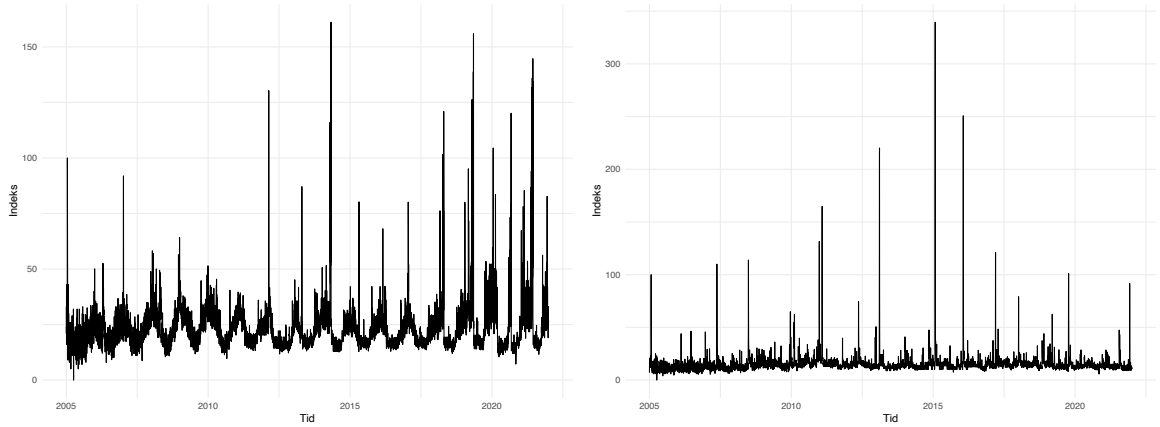
Yahoo-Finance, 2022. Yahoo-Finance. *CBOE Volatility Index (VIX)*, 2022. URL: https://finance.yahoo.com/quote/%5EVIX/. Sidst besøgt d. 28. februar 2022.

Zeileis, 2019. Achim Zeileis. *dynlm: Dynamic Linear Regression*, 2019. URL: https://CRAN.R-project.org/package=dynlm. R package version 0.3-6, Sidst besøgt d. 17. maj 2022.

Ørregaard Nielsen og Frederiksen, 2007. Morten Ørregaard Nielsen og Per Houmann Frederiksen. *Finite Sample Comparison of Parametric, Semiparametric, and Wavelet Estimators of Fractional Integration*. Econometric Reviews, 2007.

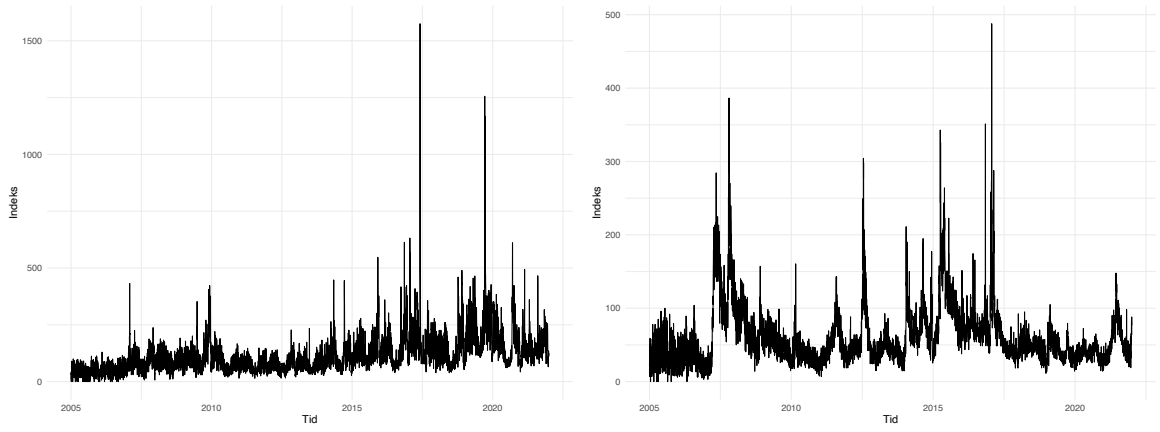
Google Trends variable A

I dette appendiks vises plots af de resterende 21 Google Trends variable fra Kapitel 5.



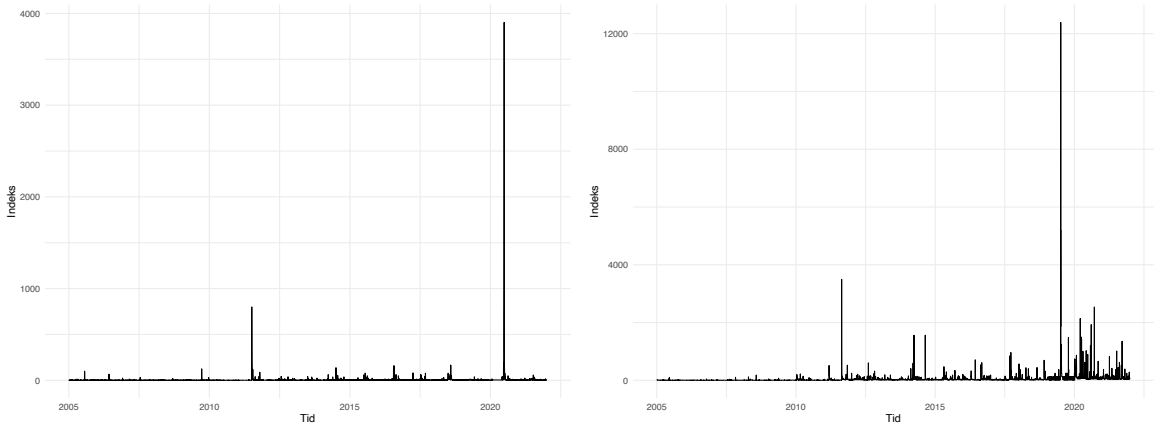
(a) Avalanche. **(b)** Blizzard.

Figur A.1



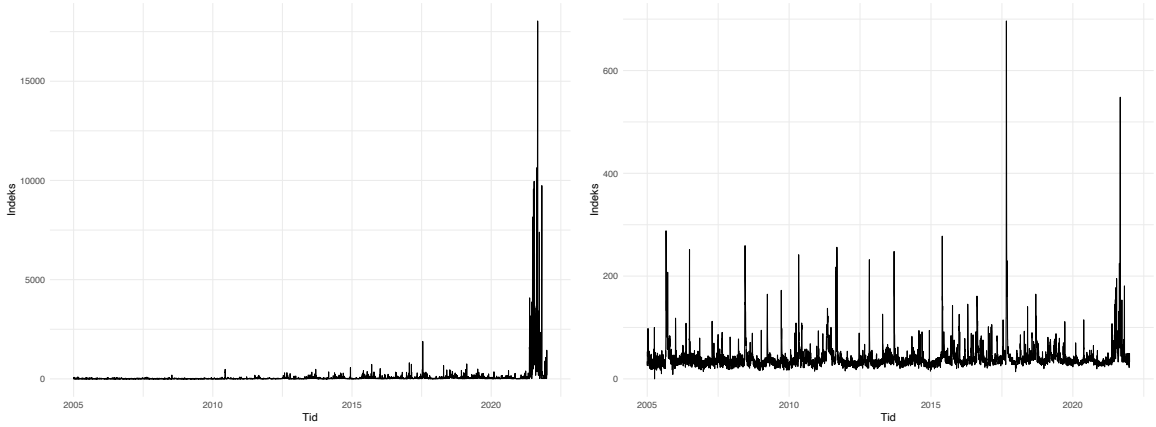
(a) Climate Change. **(b)** Drought.

Figur A.2



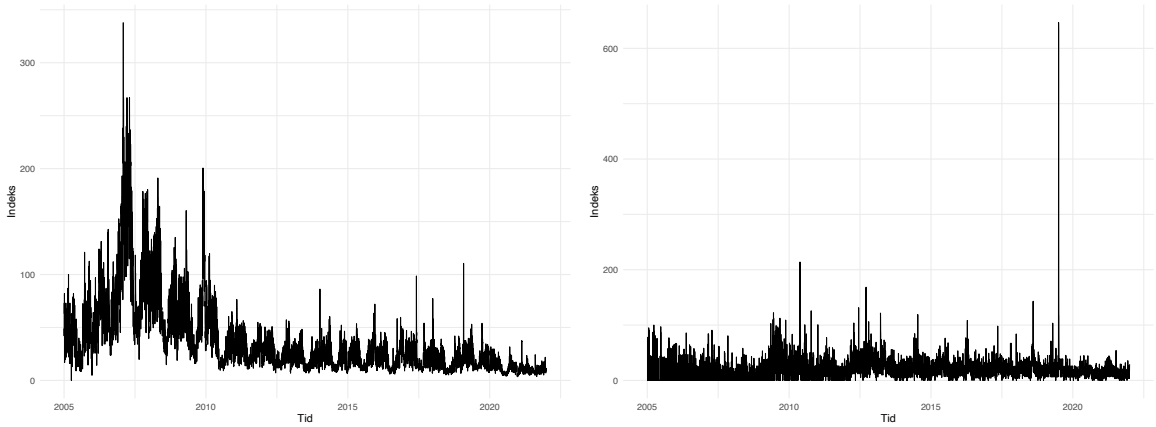
(a) Dust Storm. **(b)** Earthquake.

Figur A.3



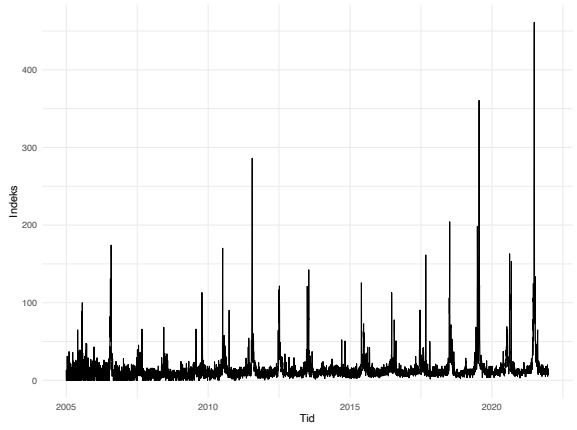
(a) Flash Flood. **(b)** Flood.

Figur A.4

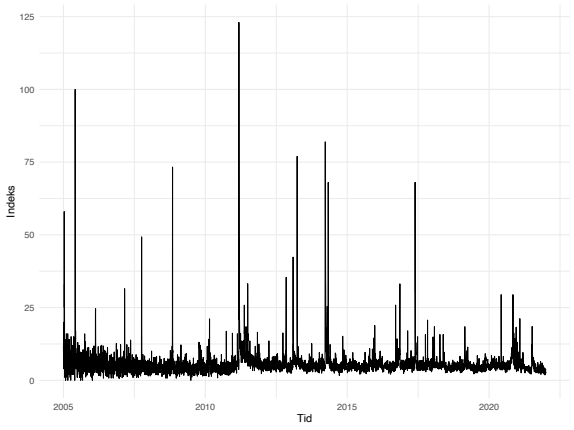


(a) Global Warming. **(b)** Hailstorm.

Figur A.5

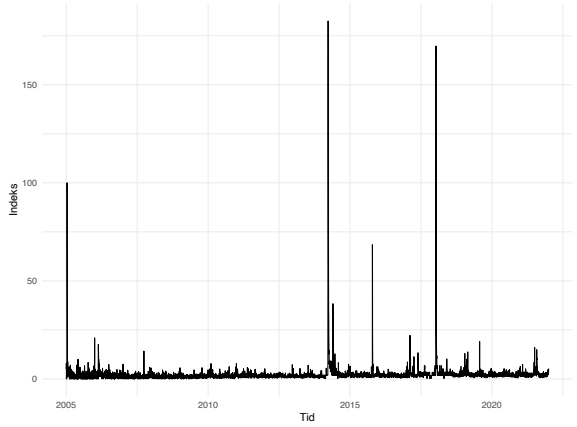


(a) Heat Wave.

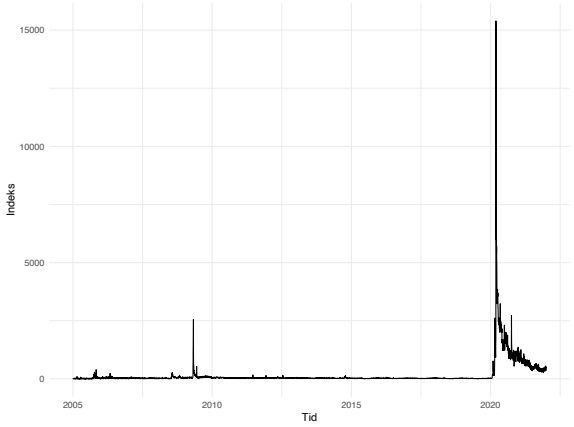


(b) Landslide.

Figur A.6

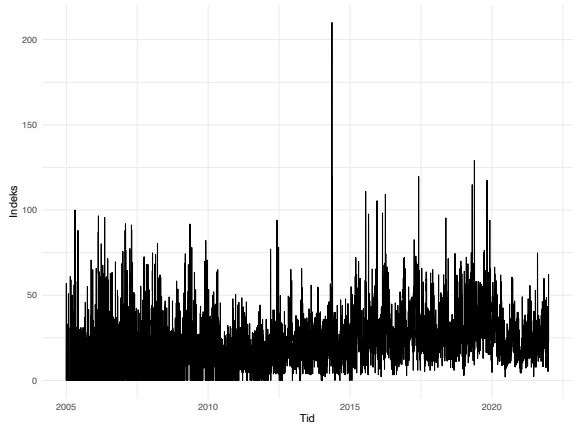


(a) Mudslide.

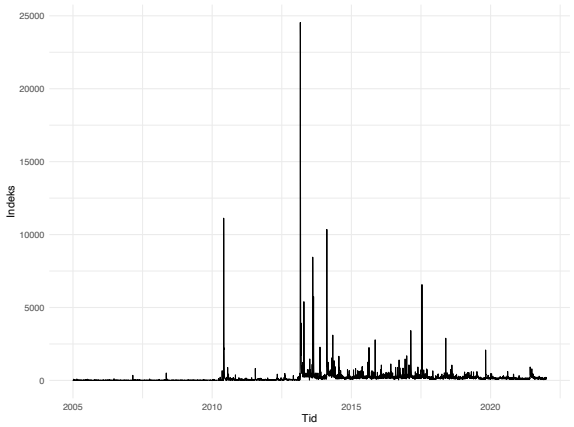


(b) Pandemic.

Figur A.7

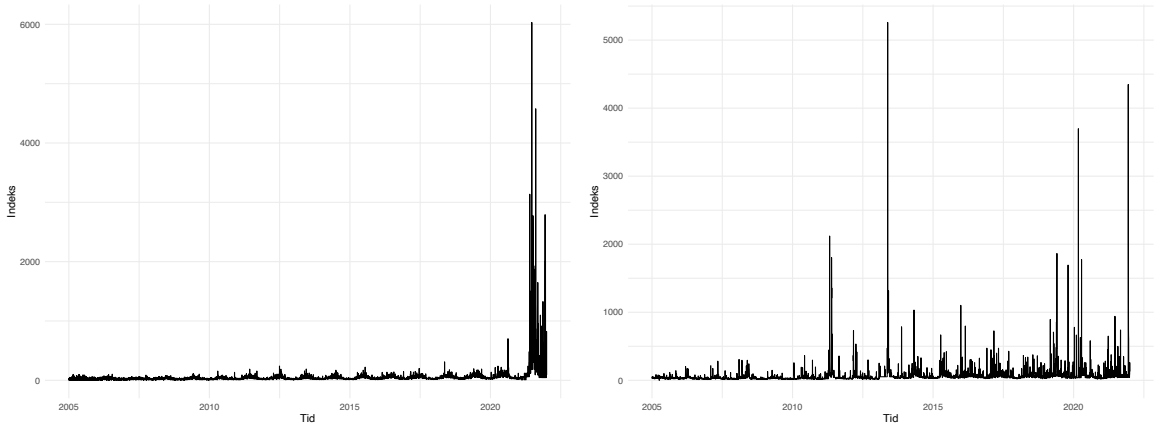


(a) Sea Level Rise.



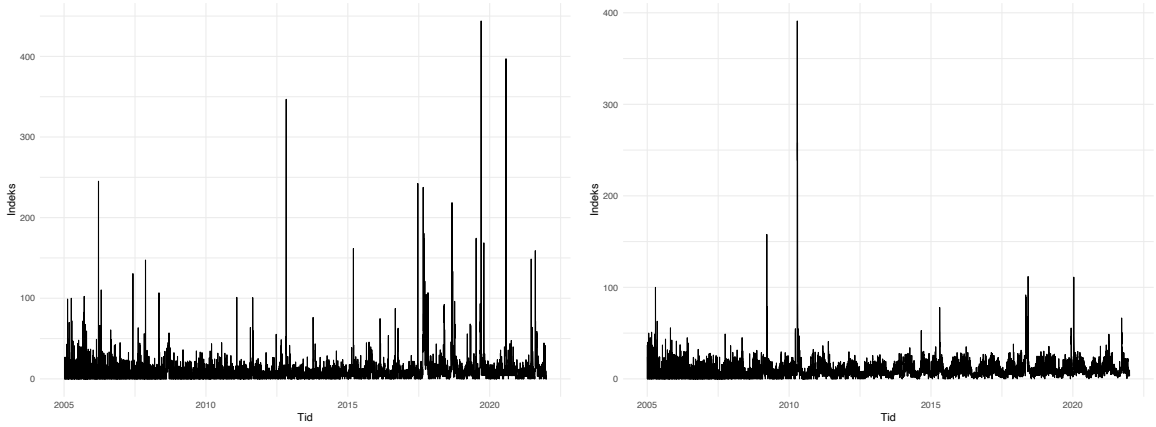
(b) Sinkhole.

Figur A.8



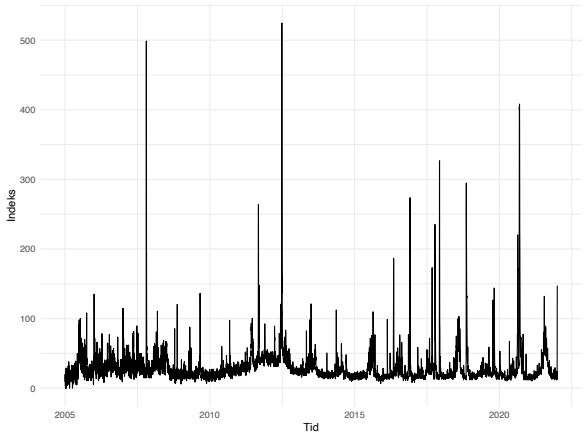
(a) Thunderstorm. **(b)** Tornado.

Figur A.9



(a) Tropical Cyclone. **(b)** Volcanic Eruption.

Figur A.10



Figur A.11. Wildfire.

Tabeller fra modellering af VIX og RV indekserne

B

B.1 VIX indekset

I dette afsnit vises resultater fra modelleringen af VIX indekset, som er placeret her af pladsmæssige årsager.

B.1.1 Koefficienter for ADL-modellen for VIX

Tabel B.1 samt Tabel B.2 viser koefficienter, standardfejl, t-værdier samt p-værdier for ADL-modellen for VIX indekset.

	Koefficient	Standardfejl	t-værdi	p-værdi
Skæring	-0,118	0,157	-0,756	0,450
AR 1	0,326	0,015	21,572	2e-16*
AR 2	0,150	0,016	9,543	2e-16*
AR 3	0,054	0,016	3,369	0,001*
AR 4	-0,011	0,016	-0,712	0,477
AR 5	0,061	0,016	3,819	0,0001*
AR 6	0,029	0,016	1,839	0,066
AR 7	-0,012	0,016	-0,748	0,454
AR 8	0,032	0,016	2,039	0,042*
AR 9	0,058	0,016	3,698	0,0002*
AR 10	0,040	0,015	2,690	0,007*
Wildfire 1	0,284	0,887	0,320	0,749
Avalanche 1	-0,666	0,649	-1,027	0,305
Blizzard 1	1,063	1,247	0,852	0,394
Climate Change 1	-1,485	1,122	-1,324	0,186
Drought 1	0,221	0,793	0,278	0,781
Dust Storm 1	0,267	1,914	0,140	0,889
Earthquake 1	-0,096	1,151	-0,084	0,933
Flash Flood 1	-1,476	1,444	-1,022	0,307
Flood 1	1,201	1,299	0,925	0,355
Global Warming 1	-0,967	0,984	-0,983	0,326
Hailstorm 1	-1,403	1,014	-1,384	0,167
Heat Wave 1	-1,851	1,139	-1,625	0,104
Landslide 1	1,355	0,934	1,451	0,147
Mudslide 1	-0,916	1,407	-0,651	0,515
Pandemic 1	21,712	1,937	11,210	2e-16*
Sea Level Rise 1	-0,774	0,453	-1,706	0,088
Sinkhole 1	-2,536	0,996	-2,546	0,011*
Thunderstorm 1	-1,583	1,202	-1,317	0,188
Tornado 1	-1,246	1,190	-1,048	0,295
Tropical Cyclone 1	-0,121	0,631	-0,192	0,848
Tsunami 1	1,136	2,076	0,547	0,584
Volcanic Eruption 1	-0,928	1,194	-0,777	0,437
Wildfire 2	-0,428	1,148	-0,372	0,710
Avalanche 2	0,680	0,671	1,013	0,311
Blizzard 2	-1,649	1,608	-1,026	0,305

	Koefficient	Standardfejl	t-værdi	p-værdi
Climate Change 2	2,471	1,338	1,847	0,065
Drought 2	0,252	0,915	0,275	0,783
Dust Storm 2	-1,306	2,503	-0,522	0,602
Earthquake 2	0,371	1,171	0,317	0,751
Flash Flood 2	0,789	1,928	0,409	0,683
Flood 2	-0,906	1,782	-0,509	0,611
Global Warming 2	1,764	1,159	1,522	0,128
Hailstorm 2	0,316	1,105	0,286	0,775
Heat Wave 2	1,392	1,535	0,907	0,365
Landslide 2	-2,154	1,061	-2,031	0,042*
Mudslide 2	1,266	1,979	0,640	0,523
Pandemic 2	-34,679	2,776	-12,494	2e-16*
Sea Level Rise 2	-0,252	0,464	-0,543	0,587
Sinkhole 2	1,719	1,303	1,320	0,187
Thunderstorm 2	0,045	1,274	0,035	0,972
Tornado 2	3,662	1,342	2,729	0,006*
Tropical Cyclone 2	0,717	0,697	1,028	0,304
Tsunami 2	1,309	2,046	0,640	0,523
Volcanic Eruption 2	0,068	1,396	0,048	0,961
Wildfire 3	0,161	1,145	0,141	0,888
Avalanche 3	0,499	0,670	0,744	0,457
Blizzard 3	0,331	1,608	0,206	0,837
Climate Change 3	-1,817	1,339	-1,358	0,175
Drought 3	-1,424	0,915	-1,557	0,120
Dust Storm 3	-0,477	2,498	-0,191	0,849
Earthquake 3	-2,277	1,166	-1,953	0,051
Flash Flood 3	-0,327	1,933	-0,169	0,866
Flood 3	0,956	1,784	0,536	0,592
Global Warming 3	0,960	1,160	0,828	0,408
Hailstorm 3	-1,181	1,110	-1,065	0,287
Heat Wave 3	-1,103	1,537	-0,717	0,473
Landslide 3	-0,001	1,168	-0,001	1,000
Mudslide 3	0,597	1,975	0,302	0,762
Pandemic 3	42,565	2,794	15,236	2e-16*
Sea Level Rise 3	0,690	0,463	1,489	0,137
Sinkhole 3	2,267	1,304	1,739	0,082
Thunderstorm 3	-0,015	1,259	-0,012	0,991

Tabel B.1. Koefficienter for ADL-modellen for VIX, del 1. P-værdier under 0,05 er markeret med *.

	Koefficient	Standardfejl	t-værdi	p-værdi
Tornado 3	-1,149	1,344	-0,855	0,393
Tropical Cyclone 3	-1,029	0,698	-1,476	0,140
Tsunami 3	3,934	1,888	2,084	0,037*
Volcanic Eruption 3	-0,009	1,396	-0,007	0,995
Wildfire 4	-0,470	0,882	-0,532	0,595
Avalanche 4	-0,169	0,645	-0,256	0,795
Blizzard 4	-0,113	1,247	-0,091	0,928
Climate Change 4	1,986	1,132	1,755	0,079
Drought 4	1,351	0,792	1,706	0,088
Dust Storm 4	0,023	1,907	0,012	0,990
Earthquake 4	-0,394	1,130	-0,348	0,728
Flash Flood 4	0,634	1,431	0,443	0,658
Flood 4	-0,740	1,296	-0,571	0,569
Global Warming 4	-1,670	0,984	-1,697	0,090
Hailstorm 4	0,514	1,026	0,501	0,617
Heat Wave 4	2,538	1,161	2,186	0,029*
Landslide 4	1,931	1,062	1,818	0,069
Mudslide 4	-1,554	1,366	-1,138	0,255
Pandemic 4	-27,585	1,958	-14,090	2e-16*
Sea Level Rise 4	-0,640	0,454	-1,409	0,159
Sinkhole 4	-2,149	0,995	-2,159	0,031*
Thunderstorm 4	0,696	1,176	0,592	0,554
Tornado 4	2,093	1,193	1,755	0,079
Tropical Cyclone 4	0,259	0,636	0,407	0,684
Tsunami 4	-4,744	1,819	-2,608	0,009*
Volcanic Eruption 4	0,591	1,194	0,495	0,620

Tabel B.2. Koefficienter for ADL-modellen for VIX, del 2. P-værdier under 0,05 er markeret med *.

B.1.2 Koefficienter for Ridge regression på ADL-modellen for VIX

Tabel B.3 viser koefficienterne, standardafvigelsen, t-værdier samt p-værdier for Ridge regression i forbindelse med VIX.

	Koefficient	Standardfejl	t-værdi	p-værdi
Skæring	-0,151	0,176	-0,861	0,786
AR 1	0,195	0,018	10,774	0,0001*
AR 2	0,122	0,019	6,438	0,0007*
AR 3	0,078	0,019	4,09	0,005*
AR 4	0,013	0,019	0,679	0,264
AR 5	0,048	0,019	2,478	0,028*
AR 6	0,026	0,019	1,356	0,117
AR 7	0,016	0,019	0,850	0,217
AR 8	0,017	0,019	0,888	0,208
AR 9	0,050	0,019	2,658	0,023*
AR 10	0,043	0,018	2,360	0,032*
Wildfire 1	0,078	0,750	0,104	0,461
Avalanche 1	-0,362	0,657	-0,552	0,698
Blizzard 1	0,342	0,866	0,395	0,355
Climate Change 1	-0,237	0,830	-0,286	0,607
Drought 1	0,099	0,727	0,136	0,448
Dust Storm 1	-0,138	0,904	-0,152	0,558
Earthquake 1	-0,796	0,914	-0,871	0,788
Flash Flood 1	-0,437	0,850	-0,514	0,686
Flood 1	0,243	0,822	0,296	0,390
Global Warming 1	-0,016	0,779	-0,021	0,508
Hailstorm 1	-1,070	0,859	-1,245	0,866
Heat Wave 1	-0,657	0,815	-0,805	0,771
Landslide 1	0,394	0,804	0,490	0,322
Mudslide 1	-0,154	0,841	-0,183	0,569
Pandemic 1	3,210	0,856	3,749	0,007*
Sea Level Rise 1	-0,508	0,502	-1,013	0,821
Sinkhole 1	-1,014	0,795	-1,276	0,871
Thunderstorm 1	-1,010	0,918	-1,100	0,839
Tornado 1	-0,420	0,900	-0,467	0,670
Tropical Cyclone 1	-0,010	0,646	-0,016	0,506
Tsunami 1	0,456	0,948	0,481	0,325
Volcanic Eruption 1	-0,429	0,883	-0,486	0,676
Wildfire 2	-0,137	0,826	-0,166	0,563
Avalanche 2	0,475	0,671	0,708	0,255
Blizzard 2	-0,449	0,855	-0,525	0,689
Climate Change 2	0,564	0,858	0,657	0,270
Drought 2	0,108	0,777	0,140	0,447
Dust Storm 2	-0,400	0,824	-0,485	0,676
Earthquake 2	0,061	0,914	0,067	0,475
Flash Flood 2	-0,091	0,830	-0,109	0,541
Flood 2	0,220	0,821	0,268	0,400
Global Warming 2	0,200	0,822	0,243	0,409
Hailstorm 2	-0,225	0,872	-0,258	0,597
Heat Wave 2	0,009	0,845	0,011	0,496
Landslide 2	-0,617	0,837	-0,737	0,753
Mudslide 2	0,159	0,818	0,195	0,427
Pandemic 2	-0,158	0,758	-0,209	0,579
Sea Level Rise 2	-0,183	0,513	-0,358	0,632
Sinkhole 2	0,598	0,832	0,718	0,252
Thunderstorm 2	-0,157	0,918	-0,171	0,565
Tornado 2	1,572	0,901	1,745	0,071
Tropical Cyclone 2	0,285	0,683	0,418	0,347
Tsunami 2	1,253	0,952	1,316	0,123
Volcanic Eruption 2	-0,250	0,890	-0,281	0,605
Wildfire 3	-0,054	0,825	-0,066	0,525
Avalanche 3	0,247	0,671	0,368	0,364
Blizzard 3	-0,193	0,855	-0,226	0,585
Climate Change 3	0,156	0,860	0,182	0,431
Drought 3	-0,108	0,778	-0,139	0,553
Dust Storm 3	-0,377	0,824	-0,458	0,667
Earthquake 3	-1,198	0,913	-1,312	0,877
Flash Flood 3	0,014	0,827	0,017	0,494
Flood 3	0,158	0,823	0,192	0,428
Global Warming 3	0,099	0,823	0,120	0,455
Hailstorm 3	-0,606	0,872	-0,695	0,741
Heat Wave 3	0,219	0,846	0,259	0,403
Landslide 3	0,161	0,831	0,193	0,427
Mudslide 3	0,075	0,814	0,092	0,465
Pandemic 3	2,913	0,755	3,856	0,006*
Sea Level Rise 3	0,247	0,512	0,483	0,325
Sinkhole 3	0,886	0,832	1,065	0,168
Thunderstorm 3	0,101	0,916	0,110	0,459
Tornado 3	0,389	0,901	0,431	0,342
Tropical Cyclone 3	-0,434	0,683	-0,635	0,724
Tsunami 3	2,614	0,972	2,689	0,022*
Volcanic Eruption 3	0,073	0,890	0,082	0,469
Wildfire 4	-0,240	0,749	-0,321	0,619
Avalanche 4	-0,019	0,657	-0,029	0,511
Blizzard 4	-0,040	0,866	-0,046	0,518
Climate Change 4	0,440	0,835	0,528	0,310
Drought 4	0,266	0,727	0,366	0,365
Dust Storm 4	-0,476	0,905	-0,526	0,689
Earthquake 4	-0,876	0,911	-0,961	0,810
Flash Flood 4	0,198	0,855	0,232	0,413
Flood 4	-0,024	0,814	-0,030	0,511
Global Warming 4	-0,027	0,778	-0,035	0,513
Hailstorm 4	0,087	0,861	0,101	0,462
Heat Wave 4	1,003	0,823	1,219	0,139
Landslide 4	0,899	0,800	1,124	0,156
Mudslide 4	-0,476	0,820	-0,580	0,707
Pandemic 4	-2,484	0,854	-2,907	0,983
Sea Level Rise 4	-0,209	0,501	-0,417	0,653
Sinkhole 4	-0,568	0,794	-0,715	0,747
Thunderstorm 4	0,352	0,913	0,385	0,358
Tornado 4	1,329	0,900	1,476	0,100
Tropical Cyclone 4	-0,089	0,649	-0,138	0,552
Tsunami 4	-2,158	0,980	-2,202	0,961
Volcanic Eruption 4	0,251	0,883	0,284	0,394

Tabel B.3. Koefficienter for Ridge ADL-modellen for VIX indekset. P-værdier under 0,05 er markeret med *.

B.1.3 Omskrevne PLS koefficienter for VIX

Tabel B.4 viser de omskrevne koefficienterne for ADL-modellen med PLS komponenter som eksterne regressorer.

	Koefficient		Koefficient		Koefficient
Skæring	0,002	Wildfire 2	0,096	Heat Wave 3	-1,299
AR 1	0,313	Avalanche 2	-2,193	Landslide 3	-0,409
AR 2	0,154	Blizzard 2	-0,494	Mudslide 3	-1,943
AR 3	0,072	Climate Change 2	-0,900	Pandemic 3	18,418
AR 4	-0,018	Drought 2	-0,709	Sea Level Rise 3	-1,797
AR 5	0,058	Dust Storm 2	-0,854	Sinkhole 3	-2,041
AR 6	0,023	Earthquake 2	-2,112	Thunderstorm 3	-0,096
AR 7	0,002	Flash Flood 2	-0,287	Tornado 3	-0,839
AR 8	0,010	Flood 2	1,130	Tropical Cyclone 3	-1,901
AR 9	0,070	Global Warming 2	3,600	Tsunami 3	-0,633
AR 10	0,041	Hailstorm 2	0,531	Volcanic Eruption 3	-0,868
Wildfire 1	-0,128	HeatWave 2	0,910	Wildfire 4	0,093
Avalanche 1	1,796	Landslide 2	0,322	Avalanche 4	-2,179
Blizzard 1	0,359	Mudslide 2	1,481	Blizzard4	-0,417
Climate Change 1	0,470	Pandemic 2	-14,826	Climate Change 4	-0,553
Drought 1	0,259	Sealevelrise 2	1,514	Drought 4	-1,043
Dust Storm 1	0,678	Sinkhole 2	1,594	Dust Storm 4	-0,763
Earthquake 1	1,633	Thunderstorm 2	-0,027	Earthquake 4	-1,805
Flash Flood 1	0,115	Tornado 2	0,602	Flash Flood 4	-0,092
Flood 1	-1,059	Tropical Cyclone 2	1,510	Flood 4	1,165
Global Warming 1	-1,566	Tsunami 2	0,486	Global Warming 4	0,670
Hailstorm 1	-0,501	Volcanic Eruption 2	0,674	Hailstorm 4	0,608
Heat Wave 1	-0,854	Wildfire 3	-0,055	Heat Wave 4	1,071
Landslide 1	-0,263	Avalanche 3	2,909	Landslide 4	0,307
Mudslide 1	-1,159	Blizzard 3	0,611	Mudslide 4	1,405
Pandemic 1	11,882	Climate Change 3	1,070	Pandemic 4	-13,620
Sea Level Rise 1	-1,549	Drought 3	1,884	Sea Level Rise 4	1,654
Sinkhole 1	-1,524	Dust Storm 3	1,038	Sinkhole 4	1,742
Thunderstorm 1	-0,051	Earthquake 3	2,511	Thunderstorm 4	0,133
Tornado 1	-0,522	Flash Flood 3	0,266	Tornado 4	0,653
Tropical Cyclone 1	-1,293	Flood 3	-1,374	Tropical Cyclone 4	1,484
Tsunami 1	-0,365	Global Warming 3	-2,467	Tsunami 4	0,443
Volcanic Eruption 1	-0,527	Hailstorm 3	-0,730	Volcanic Eruption 4	0,628

Tabel B.4. Koefficienterne for Google Trends variablene til hvert lag opnået med PLS.

B.1.4 Udvidet neuralt netværk for VIX

Tabel B.5 viser en række mulige kombinationer for strukturen af det udvidede neurale netværk, der opstilles med Google Trends variable som input. Heraf er den optimale samt udvalgte model markeret med fed.

	AIC	MSE	MAE
1 VIX lag			
1 Google lag			
d=1	4750	28,882	3,232
d=2	3336	7,711	1,797
d=3	3285	11,917	2,088
2 Google lags			
d=1	4042	31,620	3,232
d=2	2593	287,615	4,404
d=3	2267	43,414	3,034
3 Google lags			
d=1	3652	65,516	5,140
d=2	3176	11,928	2,046
d=3	2343	129,117	6,664
2 VIX lags			
1 Google lag			
d=1	4901	7,662	1,754
d=2	**4085**	**6,232***	**1,521***
d=3	3439	11,347	1,846
2 Google lags			
d=1	3680	32,935	2,942
d=2	3179	12,846	1,778
d=3	2759	25,507	2,477
3 Google lags			
d=1	3618	25,245	2,760
d=2	3057	61,240	4,561
d=3	1907	37,030	3,021
3 VIX lags			
1 Google lag			
d=1	4883	8,987	1,866
d=2	3858	8,383	1,664
d=3	3087	23,272	2,299
2 Google lags			
d=1	3534	67,764	4,563
d=2	3260	14,737	2,139
d=3	2954	15,168	2,085
3 Google lags			
d=1	3557	42,402	3,721
d=2	3030	49,209	4,034
d=3	1883	76,600	4,113
5 VIX lags			
1 Google lag			
d=1	4543	27,543	3,089
d=2	2953	44,192	2,683
d=3	3003	21,905	2,173
2 Google lags			
d=1	3603	37,465	3,035
d=2	2880	23,505	2,337
d=3	2246	72,048	3,756
3 Google lags			
d=1	3608	25,227	2,698
d=2	2908	8,579	1,679
d=3	1957	120,928	4,917
10 VIX lags			
1 Google lag			
d=1	4330	12,322	1,929
d=2	2862	39,751	2,637
d=3	2953	33,801	3,136
2 Google lags			
d=1	4374	12,363	1,933
d=2	4062	15,649	2,105
d=3	2361	97,941	4,483
3 Google lags			
d=1	3811	31,648	3,113
d=2	2436	140,078	5,245
d=3	1819*	80,347	4,392

Tabel B.5. Resultater for en række neurale netværk med lags af VIX samt Google Trends variable som input. De bedste værdier er markeret med *, og den udvalgte model er markeret med fed

Tabel B.6 viser vægtene fra det udvalgte neurale netværk.

	Vægt forbundet til knude 1	Vægt forbundet til knude 2
VIX.1	5,436	23,602
VIX.2	3,313	17,710
Wildfire.1	0,595	-14,386
Avalanche.1	0,796	5,760
Blizzard.1	-1,791	-16,214
ClimateChange.1	0,864	0,019
Drought.1	-0,156	-9,987
DustStorm.1	-65,241	-126,908
Earthquake.1	-1,919	-32,550
FlashFlood.1	-1,529	-332,703
Flood.1	1,099	23,340
GlobalWarming.1	-1,243	-4,015
Hailstorm.1	-0,697	-12,845
HeatWave.1	0,425	9,460
Landslide.1	0,397	-24,563
Mudslide.1	0,516	2,414
Pandemic.1	0,506	-43,794
Sealevelrise.1	0,334	4,825
Sinkhole.1	2,010	6,152
Thunderstorm.1	-29,283	-25,933
Tornado.1	0,211	-145,668
TropicalCyclone.1	-4,064	-81,191
Tsunami.1	-2,093	-19,746
VolcanicEruption.1	0,021	0,291

Tabel B.6. Vægtene til de to knuder i det neurale netværk for VIX indekset.

B.2 RV indekset

I dette afsnit vises resultater fra modelleringen af RV indekset, som af pladsmæssige årsager er vist her.

B.2.1 Koefficienter for ADL-modellen for RV

Tabel B.7 og B.8 viser koefficienter, standardfejl, t-værdier samt p-værdier for ADL-modellen for RV.

	Koefficient	Standardfejl	t-værdi	p-værdi
AR 1	0,342	1,53e-02	22,312	4,17e-104*
AR 2	0,303	1,63e-02	18,634	2,19e-74*
AR 3	-4,89e-02	1,71e-02	-2,862	4,24e-03*
AR 4	0,123	1,70e-02	7,213e	6,52e-13*
AR 5	6,76e-02	1,72e-02	3,942	8,23e-05*
AR 6	-3,96e-02	1,70e-02	-2,328	2,00e-02*
AR 7	-5,67e-02	1,70e-02	-3,331	8,73e-04*
AR 8	2,30e-02	1,61e-02	1,421e	0,155
AR 9	0,180	1,50e-02	11,962	1,97e-32*
Wildfire 1	-2,23e-05	8,39e-05	-0,266	0,791
Avalanche 1	-6,34e-06	6,16e-05	-0,103	0,918
Blizzard 1	1,50e-05	1,17e-04	0,128	0,898
Climate Change 1	-2,00e-04	1,08e-04	-1,853	6,39e-02
Drought 1	-4,93e-05	7,61e-05	-0,648	0,517
Dust Storm 1	-6,46e-05	1,79e-04	-0,360	0,719
Earthquake 1	-7,37e-05	1,13e-04	-0,653	0,514
Flash Flood 1	-2,18e-05	1,40e-04	-0,156	0,876
Flood 1	8,72e-05	1,22e-04	0,718	0,473
Global Warming 1	2,61e-04	9,49e-05	2,751	5,97e-03*
Hailstorm 1	-4,17e-05	9,56e-05	-0,436	0,663
Heat Wave 1	-7,92e-05	1,12e-04	-0,709	0,479
Landslide 1	1,58e-05	8,80e-05	0,179	0,858
Mudslide 1	-1,86e-05	1,33e-04	-0,140	0,889
Pandemic 1	3,51e-03	2,01e-04	17,480	5,08e-66*
Sea Level Rise 1	-3,99e-05	4,34e-05	-0,920	0,358
Sinkhole 1	-1,53e-04	9,38e-05	-1,632	0,103
Thunderstorm 1	-9,21e-05	1,19e-04	-0,774	0,439
Tornado 1	-1,66e-04	1,12e-04	-1,474	0,141
Tropical Cyclone 1	-1,89e-05	6,01e-05	-0,315	0,753
Tsunami 1	3,92e-07	1,95e-04	2,01e-03	0,998
Volcanic Eruption 1	-1,11e-05	1,12e-04	-9,91e-02	0,921
Wildfire 2	4,71e-05	1,09e-04	0,432	0,666
Avalanche 2	-5,87e-05	6,40e-05	-0,918	0,359
Blizzard 2	-4,57e-05	1,53e-04	-0,299	0,765
Climate Change 2	2,10e-04	1,28e-04	1,646e	9,99e-02
Drought 2	7,10e-05	8,69e-05	0,816	0,415
Dust Storm 2	5,25e-05	2,35e-04	0,224	0,823
Earthquake 2	-9,63e-05	1,16e-04	-0,827	0,408
Flash Flood 2	7,02e-05	1,85e-04	0,380	0,704
Flood 2	-9,91e-05	1,68e-04	-0,590	0,555
Global Warming 2	-1,89e-04	1,10e-04	-1,708	8,77e-02
Hailstorm 2	-1,21e-04	1,04e-04	-1,161	0,246
Heat Wave 2	7,62e-05	1,49e-04	0,512	0,608
Landslide 2	-5,37e-05	9,99e-05	-0,538	0,591
Mudslide 2	3,94e-05	1,88e-04	0,210	0,834
Pandemic 2	-2,78e-03	2,99e-04	-9,305	2,14e-20*
Sea Level Rise 2	-2,60e-05	4,41e-05	-0,589	0,556
Sinkhole 2	1,44e-04	1,22e-04	1,176	0,240
Thunderstorm 2	3,42e-05	1,25e-04	0,274	0,784
Tornado 2	-5,79e-05	1,27e-04	-0,455	0,649
Tropical Cyclone 2	8,46e-05	6,62e-05	1,277	0,202
Tsunami2	3,28e-04	1,99e-04	1,652	9,86e-02
Volcanic Eruption 2	7,96e-05	1,31e-04	0,606	0,545
Wildfire 3	6,22e-07	1,09e-04	5,72e-03	0,995
Avalanche 3	6,09e-05	6,90e-05	0,882	0,378

	Koefficient	Standardfejl	t-værdi	p-værdi
Blizzard 3	-2,06e-05	1,59e-04	-0,130	0,897
Climate Change 3	-1,20e-04	1,28e-04	-0,934	0,350
Drought 3	2,71e-05	8,73e-05	0,310	0,757
Dust Storm 3	-2,61e-04	2,36e-04	-1,105	0,269
Earthquake 3	4,94e-05	1,16e-04	0,427	0,669
Flash Flood 3	-1,13e-04	1,88e-04	-0,601	0,548
Flood 3	1,52e-04	1,70e-04	0,892	0,372
Global Warming 3	1,72e-04	1,12e-04	1,541	0,123
Hailstorm 3	3,78e-05	1,05e-04	0,361	0,718
Heat Wave 3	-7,54e-05	1,49e-04	-0,506	0,613
Landslide 3	2,65e-05	1,10e-04	0,241	0,809
Mudslide 3	1,18e-05	1,86e-04	6,38e-02	0,949
Pandemic 3	2,17e-03	3,09e-04	7,002	2,94e-12*
Sea Level Rise 3	1,38e-05	4,42e-05	0,311	0,755
Sinkhole 3	1,22e-04	1,25e-04	0,980	0,327
Thunderstorm3	-7,30e-05	1,24e-04	-0,589	0,556
Tornado 3	4,98e-05	1,27e-04	0,392	0,695
Tropical Cyclone 3	-6,33e-05	6,63e-05	-0,954	0,340
Tsunami 3	-1,27e-05	1,99e-04	-6,37e-02	0,949
Volcanic Eruption 3	-1,27e-04	1,32e-04	-0,964	0,334
Wildfire 4	-6,08e-05	1,09e-04	-0,558	0,577
Avalanche 4	8,80e-05	6,90e-05	1,274	0,203
Blizzard 4	7,82e-05	1,60e-04	0,489	0,625
Climate Change 4	-7,29e-05	1,28e-04	-0,572	0,568
Drought 4	-1,56e-05	8,74e-05	-0,178	0,859
Dust Storm 4	9,62e-05	2,37e-04	0,406	0,685
Earthquake 4	-2,63e-05	1,14e-04	-0,231	0,818
Flash Flood 4	7,26e-06	1,88e-04	3,87e-02	0,969
Flood 4	-8,34e-05	1,70e-04	-0,491	0,624
Global Warming 4	-6,99e-05	1,11e-04	-0,627	0,531
Hailstorm 4	2,56e-05	1,05e-04	0,243	0,808
Heat Wave 4	1,66e-04	1,49e-04	1,115	0,265
Landslide 4	1,11e-04	1,13e-04	0,987	0,324
Mudslide 4	-1,71e-04	1,823e-04	-0,938	0,349
Pandemic 4	-1,47e-03	2,91e-04	-5,055	4,49e-07*
Sea Level Rise 4	3,25e-06	4,43e-05	7,33e-02	0,942
Sinkhole 4	-2,00e-04	1,25e-04	-1,600	0,110
Thunderstorm 4	-2,64e-05	1,21e-04	-0,218	0,828
Tornado 4	5,29e-04	1,27e-04	4,153	3,35e-05*
Tropical Cyclone 4	6,94e-05	6,61e-05	1,050	0,294
Tsunami 4	-4,42e-04	1,99e-04	-2,218	2,66e-02*
Volcanic Eruption 4	1,17e-04	1,32e-04	0,888	0,375
Wildfire 5	-4,66e-05	1,09e-04	-0,428	0,668
Avalanche 5	-1,14e-05	6,92e-05	-0,164	0,870
Blizzard 5	-8,07e-05	1,60e-04	-0,505	0,613
Climate Change 5	2,66e-04	1,27e-04	2,098	3,59e-02*
Drought 5	6,01e-05	8,74e-05	0,688	0,492
Dust Storm 5	-4,22e-06	2,38e-04	-1,78e-02	0,986
Earthquake 5	6,72e-05	1,14e-04	0,591	0,555
Flash Flood 5	3,18e-05	1,87e-04	0,170	0,865
Flood 5	-1,51e-05	1,70e-04	-8,871e-02	0,929
Global Warming 5	-2,17e-04	1,12e-04	-1,949	5,14e-02
Hailstorm 5	8,82e-05	1,05e-04	0,839	0,401
Heat Wave 5	-8,00e-05	1,47e-04	-0,543	0,587
Landslide 5	-2,20e-04	1,13e-04	-1,952	5,10e-02

Tabel B.7. Koefficienter for ADL-modellen for RV indekset, del 1. P-værdier under 0,05 er markeret med *.

	Koefficient	Standardfejl	t-værdi	p-værdi
Mudslide 5	2,86e-04	1,83e-04	1,561	0,119
Pandemic 5	1,76e-06	2,87e-04	6,14e-03	0,995
Sea Level Rise 5	-4,70e-06	4,44e-05	-0,106	0,916
Sinkhole 5	3,54e-05	1,25e-04	2,83e-01	0,778
Thunderstorm 5	6,19e-05	1,22e-04	0,506	0,613
Tornado 5	-2,35e-04	1,28e-04	-1,844	6,52e-02
Tropical Cyclone 5	-4,69e-05	6,61e-05	-0,710	0,478
Tsunami 5	-1,24e-04	1,99e-04	-0,623	0,533
Volcanic Eruption 5	-1,98e-04	1,32e-04	-1,504	0,133
Wildfire 6	3,13e-05	1,09e-04	0,287	0,774
Avalanche 6	-4,50e-05	6,90e-05	-0,652	0,514
Blizzard 6	-2,35e-05	1,59e-04	-0,148	0,883
Climate Change 6	-4,49e-05	1,27e-04	-0,354	0,723
Drought 6	-9,67e-05	8,76e-05	-1,104	0,270
Dust Storm 6	4,51e-06	2,37e-04	1,90e-02	0,985
Earthquake 6	6,94e-05	1,13e-04	0,614	0,539
Flash Flood 6	3,15e-05	1,86e-04	0,169	0,866
Flood 6	-5,63e-05	1,70e-04	-0,331	0,740
Global Warming 6	1,36e-04	1,11e-04	1,221	0,222
Hailstorm 6	-7,71e-05	1,05e-04	-0,734	0,463
Heat Wave 6	1,19e-04	1,47e-04	0,807	0,420
Landslide 6	4,34e-04	1,13e-04	3,848	1,21e-04*
Mudslide 6	-3,48e-04	1,83e-04	-1,905	5,68e-02
Pandemic 6	-7,97e-05	2,91e-04	-0,274	0,784
Sea Level Rise 6	-4,42e-05	4,43e-05	-0,998	0,319
Sinkhole 6	4,72e-05	1,25e-04	0,377	0,706
Thunderstorm 6	2,59e-05	1,25e-04	0,208	0,835
Tornado 6	-1,97e-04	1,27e-04	-1,547	0,122
Tropical Cyclone 6	-5,60e-05	6,63e-05	-0,844	0,399
Tsunami 6	3,36e-05	1,98e-04	0,169	0,866
Volcanic Eruption 6	3,31e-04	1,31e-04	2,519	1,18e-02*
Wildfire 7	3,70e-05	1,09e-04	0,340	0,734
Avalanche 7	-5,58e-05	6,98e-05	-0,800	0,424
Blizzard 7	8,21e-05	1,58e-04	0,520	0,603
Climate Change 7	-7,69e-05	1,27e-04	-0,606	0,544
Drought 7	6,76e-06	8,76e-05	7,72e-02	0,939
Dust Storm 7	-7,32e-05	2,36e-04	-0,310	0,757
Earthquake 7	-2,36e-04	1,12e-04	-2,110	3,49e-02*
Flash Flood 7	-1,00e-05	1,86e-04	-5,37e-02	0,957
Flood 7	9,16e-05	1,70e-04	0,540	0,589
Global Warming 7	-4,55e-05	1,11e-04	-0,410	0,682
Hailstorm 7	9,08e-06	1,05e-04	8,61e-02	0,931
Heat Wave 7	-3,39e-05	1,47e-04	-0,231	0,817
Landslide 7	6,84e-06	1,13e-04	6,05e-02	0,952
Mudslide 7	3,32e-05	1,82e-04	0,182	0,855
Pandemic 7	-1,07e-03	2,96e-04	-3,602	3,19e-04*
Sea Level Rise 7	2,70e-05	4,43e-05	0,603	0,547
Sinkhole 7	-3,57e-05	1,25e-04	-0,286	0,775
Thunderstorm 7	-3,61e-05	1,25e-04	-0,289	0,772
Tornado 7	1,62e-04	1,36e-04	1,185	0,236
Tropical Cyclone 7	1,77e-05	6,64e-05	0,267	0,790
Tsunami 7	-5,08e-05	1,90e-04	-0,268	0,789
Volcanic Eruption 7	-2,53e-04	1,31e-04	-1,927	5,41e-02
Wildfire 8	-1,01e-04	1,09e-04	-0,935	0,350
Avalanche 8	-6,89e-06	6,52e-05	-0,106	0,916
Blizzard 8	-4,39e-07	1,52e-04	-2,89e-03	0,998
Climate Change 8	1,02e-04	1,26e-04	0,807	0,420
Drought 8	5,91e-05	8,73e-05	0,677	0,498
Dust Storm 8	4,25e-05	2,34e-04	0,181	0,856
Earthquake 8	1,40e-05	1,11e-04	0,126	0,900
Flash Flood 8	1,01e-04	1,84e-04	0,547	0,584
Flood 8	-2,61e-04	1,68e-04	-1,556	0,130
Global Warming 8	-1,55e-05	1,10e-04	-0,141	0,888
Hailstorm 8	-7,11e-05	1,05e-04	-0,675	0,500
Heat Wave 8	-1,23e-04	1,47e-04	-0,834e-01	0,404
Landslide 8	-1,62e-04	1,14e-04	-1,430	0,153
Mudslide 8	1,67e-04	1,79e-04	0,934	0,350
Pandemic 8	1,86e-03	2,85e-04	6,529	7,43e-11*
Sea Level Rise 8	-1,32e-05	4,40e-05	-0,300	0,764
Sinkhole 8	-7,25e-05	1,22e-04	-0,593	0,553
Thunderstorm 8	-1,46e-05	1,24e-04	-0,118	0,906
Tornado 8	-5,80e-05	1,40e-04	-0,418	0,676
Tropical Cyclone 8	3,58e-05	6,63e-05	0,539	0,590
Tsunami 8	-5,27e-05	1,74e-04	-0,303	0,762
Volcanic Eruption 8	1,12e-04	1,31e-04	0,854	0,393
Wildfire 9	8,86e-05	6,34e-05	1,397	0,163
Blizzard 9	-3,54e-05	1,16e-04	-0,304	0,761
Climate Change 9	-2,23e-05	1,07e-04	-0,208	0,835
Drought 9	-4,79e-05	7,60e-05	-0,630	0,529
Dust Storm 9	-5,46e-06	1,79e-04	-3,05e-02	0,976
Earthquake 9	1,14e-04	1,07e-04	1,061	0,289
Flash Flood 9	-1,06e-04	1,38e-04	-0,768	0,442
Flood 9	2,39e-04	1,22e-04	1,959	5,01e-02
Global Warming 9	-2,07e-05	9,45e-05	-0,219	0,827
Hailstorm 9	7,26e-05	9,71e-05	0,747	0,455
Heat Wave 9	1,24e-04	1,12e-04	1,109	0,267
Landslide 9	5,75e-05	1,00e-04	0,575	0,566
Mudslide 9	-1,14e-04	1,25e-04	-0,913	0,361
Pandemic 9	-2,02e-03	1,99e-04	-10,112	9,84e-24*
Sea Level Rise 9	-1,64e-05	4,30e-05	-0,381	0,703
Sinkhole 9	4,04e-05	9,37e-05	0,431	0,667
Thunderstorm 9	-2,43e-05	1,18e-04	-0,206	0,837
Tornado 9	3,43e-06	1,22e-04	2,81e-02	0,978
Tropical Cyclone 9	-6,50e-06	6,06e-05	-0,107	0,915
Tsunami 9	-3,09e-05	1,65e-04	-0,187	0,851
Volcanic Eruption 9	1,03e-05	1,12e-04	9,15e-02	0,927

Tabel B.8. Koefficienter for ADL-modellen for RV indekset, del 2. P-værdier under 0,05 er markeret med *.

B.2.2 Koefficenter for Ridge regression på ADL-modellen for RV

Tabel B.9 og B.10 viser koefficienterne, standardafvigelserne, t-værdier samt p-værdier for Rigde regressionen på ADL-modellen for RV.

	Koefficient	Standardfejl	t-værdi	p-værdi
AR 1	0,234	1,43e-02	16,359	7,79e-06*
AR 2	0,224	1,43e-02	15,686	9,57e-06*
AR 3	0,51e-02	1,43e-02	3,559	8,11e-03*
AR 4	9,81e-02	1,44e-02	6,811	5,20e-04*
AR 5	6,57e-02	1,44e-02	4,557	3,04e-03*
AR 6	3,60e-03	1,44e-02	0,250	0,406
AR 7	1,49e-02	1,44e-02	1,039	0,173
AR 8	1,80e-02	1,43e-02	1,259	0,132
AR 9	0,121	1,41e-02	8,535	1,82e-04*
Wildfire 1	2,47e-06	1,40e-04	0,018	0,493
Avalanche 1	-1,10e-05	1,04e-04	-0,106	0,540
Blizzard 1	1,56e-05	1,96e-04	0,079	0,470
Climate Change 1	-4,23e-05	1,81e-04	-0,234	0,588
Drought 1	-1,02e-05	1,27e-04	-0,080	0,530
Dust Storm 1	-4,00e-05	3,00e-04	-0,133	0,550
Earthquake 1	-5,72e-05	1,89e-04	-0,303	0,613
Flash Flood 1	-3,78e-06	2,35e-04	-0,016	0,506
Flood 1	3,88e-05	2,04e-04	0,190	0,429
Global Warming 1	4,93e-05	1,59e-04	0,311	0,384
Hailstorm 1	-5,42e-05	1,60e-04	-0,338	0,626
Heat Wave 1	-2,41e-05	1,87e-04	-0,129	0,549
Landslide 1	-7,05e-07	1,47e-04	-0,005	0,502
Mudslide 1	2,97e-06	2,23e-04	0,013	0,495
Pandemic 1	1,13e-03	3,32e-04	3,400	9,62e-03*
Sea Level Rise 1	-3,83e-05	7,26e-05	-0,528	0,690
Sinkhole 1	-8,64e-05	1,57e-04	-0,550	0,697
Thunderstorm 1	-6,14e-05	1,99e-04	-0,309	0,615
Tornado 1	-1,17e-04	1,88e-04	-0,623	0,720
Tropical Cyclone 1	-8,48e-06	1,01e-04	-0,084	0,532
Tsunami 1	-4,87e-06	3,27e-04	-0,015	0,506
Volcanic Eruption 1	-4,39e-07	1,88e-04	-0,002	0,501
Wildfire 2	2,23e-05	1,82e-04	0,122	0,454
Avalanche 2	8,47e-07	1,07e-04	0,008	0,497
Blizzard 2	-3,35e-05	2,56e-04	-0,131	0,550
Climate Change 2	3,17e-05	2,14e-04	0,148	0,444
Drought 2	1,15e-05	1,45e-04	0,079	0,470
Dust Storm 2	-1,25e-05	3,93e-04	-0,031	0,512
Earthquake 2	-1,19e-04	1,95e-04	-0,613	0,717
Flash Flood 2	-9,23e-06	3,09e-04	-0,030	0,511
Flood 2	1,07e-05	2,81e-04	0,038	0,486
Global Warming 2	1,15e-05	1,85e-04	0,062	0,477
Hailstorm 2	-9,48e-05	1,74e-04	-0,545	0,695
Heat Wave 2	-2,20e-06	2,49e-04	-0,009	0,503
Landslide 2	-2,88e-05	1,67e-04	-0,173	0,565
Mudslide 2	-7,93e-06	3,14e-04	-0,025	0,510
Pandemic 2	2,30e-04	4,84e-04	0,477	0,327
Sea Level Rise 2	-3,17e-05	7,37e-05	-0,430	0,658
Sinkhole 2	7,50e-05	2,05e-04	0,367	0,364
Thunderstorm 2	-1,07e-05	2,08e-04	-0,051	0,519
Tornado 2	-5,49e-05	2,12e-04	-0,259	0,597
Tropical Cyclone 2	4,35e-05	1,11e-04	0,392	0,355
Tsunami 2	2,64e-04	3,32e-04	0,793	0,232
Volcanic Eruption 2	4,21e-05	2,20e-04	0,191	0,428
Wildfire 3	-5,07e-06	1,82e-04	-0,028	0,511
Avalanche 3	3,40e-05	1,16e-04	0,294	0,390
Blizzard 3	6,54e-06	2,66e-04	0,025	0,491
Climate Change 3	-6,94e-06	2,14e-04	-0,033	0,512
Drought 3	1,60e-05	1,46e-04	0,110	0,458
Dust Storm 3	-5,67e-05	3,95e-04	-0,143	0,554
Earthquake 3	1,44e-04	1,94e-04	0,743	0,246
Flash Flood 3	-2,13e-05	3,14e-04	-0,068	0,526
Flood 3	3,55e-05	2,84e-04	0,125	0,453
Global Warming 3	2,71e-05	1,86e-04	0,145	0,445
Hailstorm 3	8,91e-06	1,75e-04	0,051	0,481
Heat Wave 3	6,67e-06	2,49e-04	0,027	0,490
Landslide 3	2,26e-05	1,84e-04	0,123	0,454
Mudslide 3	-1,03e-06	3,12e-04	-0,003	0,501
Pandemic 3	2,52e-04	4,98e-04	0,506	0,317
Sea Level Rise 3	3,10e-06	7,40e-05	0,042	0,484
Sinkhole 3	6,21e-05	2,09e-04	0,297	0,389
Thunderstorm 3	-3,77e-05	2,07e-04	-0,182	0,569
Tornado 3	3,75e-05	2,12e-04	0,177	0,433
Tropical Cyclone 3	-1,42e-05	1,11e-04	-0,128	0,548
Tsunami 3	-9,69e-05	3,33e-04	-0,291	0,609
Volcanic Eruption 3	-4,48e-05	2,20e-04	-0,204	0,577
Wildfire 4	-4,92e-05	1,82e-04	-0,270	0,601
Avalanche 4	4,17e-05	1,16e-04	0,361	0,366
Blizzard 4	1,69e-05	2,67e-04	0,063	0,476
Climate Change 4	-2,39e-05	2,13e-04	-0,112	0,542
Drought 4	5,28e-06	1,46e-04	0,036	0,486
Dust Storm 4	-2,28e-05	3,96e-04	-0,058	0,522
Earthquake 4	-9,35e-05	1,90e-04	-0,491	0,678
Flash Flood 4	-5,22e-06	3,14e-04	-0,017	0,506
Flood 4	-2,01e-05	2,84e-04	-0,071	0,527
Global Warming 4	-9,13e-06	1,86e-04	-0,049	0,519
Hailstorm 4	1,37e-05	1,76e-04	0,078	0,471
Heat Wave 4	5,13e-05	2,49e-04	0,206	0,423
Landslide 4	7,08e-05	1,89e-04	0,376	0,361
Mudslide 4	-3,53e-05	3,06e-04	-0,116	0,544
Pandemic 4	-2,03e-04	4,68e-04	-0,433	0,659
Sea Level Rise 4	1,69e-06	7,41e-05	0,023	0,491
Sinkhole 4	-7,94e-05	2,09e-04	-0,379	0,640
Thunderstorm 4	-1,87e-05	2,30e-04	-0,092	0,535
Tornado 4	3,31e-04	2,12e-04	1,559	8,99e-02
Tropical Cyclone 4	3,13e-05	1,11e-04	0,283	0,394
Tsunami 4	-1,91e-04	3,33e-04	-0,573	0,704
Volcanic Eruption 4	2,48e-05	2,20e-04	0,113	0,457
Wildfire 5	-3,56e-05	1,82e-04	-0,196	0,574
Avalanche 5	1,35e-05	1,16e-04	0,116	0,456
Blizzard 5	-5,63e-05	2,67e-04	-0,210	0,579
Climate Change 5	4,78e-05	2,12e-04	0,225	0,415
Drought 5	1,13e-05	1,46e-04	0,0772	0,471
Dust Storm 5	-8,48e-06	3,97e-04	-0,021	0,508
Earthquake 5	-1,40e-05	1,90e-04	-0,074	0,528
Flash Flood 5	2,46e-05	3,13e-04	0,078	0,470
Flood 5	-2,50e-05	2,84e-04	-0,088	0,533
Global Warming 5	-2,04e-05	1,86e-04	-0,110	0,542
Hailstorm 5	5,54e-05	1,76e-04	0,315	0,383
Heat Wave 5	1,23e-05	2,47e-04	0,050	0,481
Landslide 5	-4,39e-05	1,89e-04	-0,232	0,587

Tabel B.9. Koefficienter for Ridge ADL-modellen for RV, del 1. P-værdier under 0,05 er markeret med *.

	Koefficient	Standardfejl	t-værdi	p-værdi
Mudslide 5	1,24e-05	3,06e-04	0,041	0,485
Pandemic 5	-1,97e-04	4,59e-04	-0,428	0,657
Sea Level Rise 5	-1,02e-05	7,43e-05	-0,138	0,552
Sinkhole 5	-1,64e-05	2,09e-04	-0,078	0,530
Thunderstorm 5	3,50e-06	2,05e-04	0,017	0,494
Tornado 5	-1,25e-04	2,12e-04	-0,590	0,710
Tropical Cyclone 5	-3,86e-05	1,11e-04	-0,349	0,630
Tsunami 5	-1,14e-04	3,32e-04	-0,343	0,627
Volcanic Eruption 5	-9,09e-05	2,20e-04	-0,413	0,652
Wildfire 6	-6,11e-06	1,82e-04	-0,034	0,513
Avalanche 6	-3,86e-05	1,16e-04	-0,334	0,624
Blizzard 6	2,41e-06	2,67e-04	0,009	0,497
Climate Change 6	2,19e-05	2,12e-04	0,103	0,461
Drought 6	-2,33e-05	1,47e-04	-0,159	0,560
Dust Storm 6	-6,88e-06	3,96e-04	-0,017	0,507
Earthquake 6	-3,50e-05	1,89e-04	-0,185	0,570
Flash Flood 6	5,76e-06	3,11e-04	0,019	0,493
Flood 6	-2,78e-05	2,84e-04	-0,098	0,537
Global Warming 6	4,13e-06	1,86e-04	0,022	0,492
Hailstorm 6	-7,53e-05	1,76e-04	-0,428	0,657
Heat Wave 6	4,18e-05	2,46e-04	0,170	0,436
Landslide 6	2,04e-04	1,89e-04	1,080	0,165
Mudslide 6	-7,95e-05	3,05e-04	-0,260	0,598
Pandemic 6	-2,69e-04	4,68e-04	-0,576	0,705
Sea Level Rise 6	-2,77e-05	7,41e-05	-0,374	0,638
Sinkhole 6	1,43e-05	2,09e-04	0,068	0,474
Thunderstorm 6	6,11e-06	2,09e-04	0,029	0,489
Tornado 6	-1,08e-04	2,12e-04	-0,508	0,684
Tropical Cyclone 6	-2,69e-05	1,11e-04	-0,242	0,591
Tsunami 6	-4,07e-05	3,32e-04	-0,123	0,546
Volcanic Eruption 6	1,61e-04	2,20e-04	0,731	0,249
Wildfire 7	1,98e-05	1,82e-04	0,109	0,459
Avalanche 7	2,78e-06	1,17e-04	0,024	0,491
Blizzard 7	4,11e-05	2,64e-04	0,156	0,441
Climate Change 7	3,50e-06	2,12e-04	0,016	0,494
Drought 7	-3,43e-06	1,47e-04	-0,023	0,509
Dust Storm 7	-2,74e-05	3,95e-04	-0,069	0,526
Earthquake 7	-1,16e-04	1,87e-04	-0,620	0,719
Flash Flood 7	1,92e-05	3,12e-04	0,062	0,477
Flood 7	-1,28e-05	2,84e-04	-0,045	0,517
Global Warming 7	-1,22e-05	1,86e-04	-0,065	0,525
Hailstorm 7	-6,83e-06	1,76e-04	-0,039	0,515
Heat Wave 7	-6,89e-06	2,46e-04	-0,028	0,511
Landslide 7	4,14e-05	1,89e-04	0,219	0,418
Mudslide 7	-1,63e-05	3,05e-04	-0,054	0,520
Pandemic 7	-1,36e-04	4,77e-04	-0,285	0,607
Sea Level Rise 7	7,80e-06	7,41e-05	0,105	0,460
Sinkhole 7	-1,29e-05	2,09e-04	-0,062	0,523
Thunderstorm 7	-3,20e-05	2,09e-04	-0,153	0,558
Tornado 7	3,98e-04	2,28e-04	1,747	7,06e-02
Tropical Cyclone 7	8,56e-06	1,11e-04	0,077	0,471
Tsunami 7	-8,31e-05	3,18e-04	-0,262	0,598
Volcanic Eruption 7	-1,03e-04	2,20e-04	-0,471	0,671
Wildfire 8	-1,33e-05	1,82e-04	-0,073	0,528
Avalanche 8	-1,02e-05	1,09e-04	-0,094	0,536
Blizzard 8	6,15e-06	2,54e-04	0,024	0,491
Climate Change 8	3,00e-05	2,11e-04	0,142	0,446
Drought 8	1,04e-05	1,46e-04	0,071	0,473
Dust Storm 8	1,97e-05	3,92e-04	0,050	0,481
Earthquake 8	-1,44e-05	1,86e-04	-0,078	0,530
Flash Flood 8	-3,97e-06	3,08e-04	-0,013	0,505
Flood 8	-4,39e-05	2,81e-04	-0,156	0,559
Global Warming 8	-6,22e-06	1,84e-04	-0,034	0,513
Hailstorm 8	-2,67e-05	1,77e-04	-0,151	0,557
Heat Wave 8	-2,62e-05	2,46e-04	-0,107	0,540
Landslide 8	-3,52e-05	1,90e-04	-0,186	0,570
Mudslide 8	1,55e-05	2,99e-04	0,052	0,480
Pandemic 8	-2,15e-05	4,61e-04	-0,047	0,518
Sea Level Rise 8	-7,41e-06	7,36e-05	-0,101	0,538
Sinkhole 8	-4,93e-05	2,04e-04	-0,241	0,591
Thunderstorm 8	-1,38e-05	2,07e-04	-0,067	0,525
Tornado 8	-1,85e-04	2,32e-04	-0,797	0,769
Tropical Cyclone 8	7,97e-06	1,11e-04	0,072	0,473
Tsunami 8	-5,88e-05	2,91e-04	-0,202	0,576
Volcanic Eruption 8	7,25e-05	2,19e-04	0,330	0,377
Wildfire 9	3,88e-05	1,40e-04	0,278	0,396
Avalanche 9	5,25e-05	1,06e-04	0,493	0,321
Blizzard 9	-9,05e-06	1,96e-04	-0,046	0,518
Climate Change 9	2,32e-07	1,79e-04	0,001	0,499
Drought 9	-6,63e-06	1,27e-04	-0,052	0,520
Dust Storm 9	-3,53e-06	2,99e-04	-0,012	0,505
Earthquake 9	5,52e-05	1,79e-04	0,308	0,385
Flash Flood 9	-1,74e-05	2,31e-04	-0,075	0,529
Flood 9	7,49e-05	2,05e-04	0,366	0,365
Global Warming 9	-1,70e-05	1,58e-04	-0,108	0,541
Hailstorm 9	4,98e-05	1,63e-04	0,305	0,386
Heat Wave 9	4,62e-05	1,87e-04	0,247	0,407
Landslide 9	9,24e-06	1,68e-04	0,055	0,479
Mudslide 9	-1,53e-05	2,10e-04	-0,073	0,528
Pandemic 9	-5,77e-04	3,24e-04	-1,780	0,932
Sea Level Rise 9	-1,09e-05	7,20e-05	-0,151	0,557
Sinkhole 9	-1,17e-06	1,57e-04	-0,007	0,503
Thunderstorm 9	-1,08e-05	1,97e-04	-0,055	0,521
Tornado 9	9,29e-05	2,04e-04	0,455	0,334
Tropical Cyclone 9	-2,94e-06	1,02e-04	-0,029	0,511
Tsunami 9	-3,67e-05	2,75e-04	-0,133	0,550
Volcanic Eruption 9	-1,32e-05	1,88e-04	-0,070	0,527

Tabel B.10. Koefficienter for Ridge ADL-modellen for RV, del 2. P-værdier under 0,05 er markeret med *.

B.2.3 Omskrevne PLS koefficienter for RV

Tabel B.11 og B.12 viser de omskrevne koefficienterne for ADL-modellen med PLS komponenter som eksterne regressorer.

	Koefficient		Koefficient		Koefficient
AR 1	0,337	Thunderstorm 2	1,30e-04	Volcanic Eruption 4	2,63e-05
AR 2	0,315	Tornado 2	-9,09e-05	Wildfire 5	3,96e-06
AR 3	-4,00e-02	Tropical Cyclone 2	1,57e-04	Avalanche 5	-2,30e-05
AR 4	0,121	Tsunami 2	5,34e-05	Blizzard 5	-5,23e-06
AR 5	7,21e-02	Volcanic Eruption 2	5,29e-05	Climate Change 5	2,23e-05
AR 6	-4,41e-02	Wildfire 3	-2,28e-05	Drought 5	-4,68e-05
AR 7	-4,58e-02	Avalanche 3	1,18e-04	DustStorm 5	9,33e-06
AR 8	1,26e-02	Blizzard 3	6,09e-05	Earthquake 5	-6,47e-06
AR 9	0,189	Climate Change 3	-1,79e-05	Flash Flood 5	1,75e-05
Wildfire 1	-8,13e-05	Drought 3	1,89e-04	Flood 5	1,03e-05
Avalanche 1	2,24e-04	DustStorm 3	-5,77e-05	Global Warming 5	-1,43e-04
Blizzard 1	1,00e-04	Earthquake 3	1,22e-04	Hailstorm 5	3,07e-05
Climate Change 1	-7,27e-05	Flash Flood 3	-3,29e-05	Heat Wave 5	3,02e-05
Drought 1	1,76e-04	Flood 3	-1,57e-06	Landslide 5	-1,05e-06
Dust Storm 1	-8,07e-05	Global Warming 3	1,23e-05	Mudslide 5	1,55e-05
Earthquake 1	2,58e-04	Hailstorm 3	-1,65e-04	Pandemic 5	-1,91e-04
Flash Flood 1	-9,10e-05	Heat Wave 3	-1,06e-04	Sea Level Rise 5	2,29e-05
Flood 1	-8,76e-05	Landslide 3	1,23e-05	Sinkhole 5	2,19e-05
Global Warming 1	1,78e-04	Mudslide 3	-1,05e-04	Thunderstorm 5	2,43e-05
Hailstorm 1	-3,38e-04	Pandemic 3	1,49e-03	Tornado 5	-3,90e-06
Heat Wave 1	-2,42e-04	Sea Level Rise 3	-4,03e-05	Tropical Cyclone 5	2,01e-05
Landslide 1	1,62e-05	Sinkhole 3	-2,24e-05	Tsunami 5	5,49e-06
Mudslide 1	-1,91e-04	Thunderstorm 3	-1,07e-04	Volcanic Eruption 5	6,39e-06
Pandemic 1	2,98e-03	Tornado 3	7,49e-05	Wildfire 6	-1,45e-05
Sea Level Rise 1	-1,91e-04	TropicalCyclone 3	-1,01e-04	Avalanche 6	-2,74e-05
Sinkhole 1	-1,33e-04	Tsunami 3	-4,71e-05	Blizzard 6	-2,75e-05
Thunderstorm 1	-1,99e-04	Volcanic Eruption 3	-4,80e-05	Climate Change 6	-3,13e-05
Tornado 1	1,30e-04	Wildfire 4	2,23e-05	Drought 6	-1,33e-04
Tropical Cyclone 1	-2,29e-04	Avalanche 4	-7,23e-05	Dust Storm 6	3,02e-05
Tsunami 1	-7,92e-05	Blizzard 4	-3,29e-05	Earthquake 6	-2,77e-05
Volcanic Eruption 1	-7,94e-05	Climate Change 4	2,22e-05	Flash Flood 6	-1,63e-05
Wildfire 2	5,99e-05	Drought 4	-7,65e-05	Flood 6	-5,24e-05
Avalanche 2	-1,49e-04	Dust Storm 4	2,86e-05	Global Warming 6	1,90e-04
Blizzard 2	-6,84e-05	Earthquake 4	-7,80e-05	Hailstorm 6	2,67e-05
Climate Change 2	4,27e-05	Flash Flood 4	2,87e-05	Heat Wave 6	-5,07e-06
Drought 2	-8,46e-05	Flood 4	2,12e-05	Landslide 6	-7,69e-06
Dust Storm 2	5,06e-05	Global Warming 4	-6,19e-05	Mudslide 6	3,28e-05
Earthquake 2	-1,86e-04	Hailstorm 4	1,06e-04	Pandemic 6	-3,42e-04
Flash Flood 2	5,80e-05	Heat Wave 4	7,54e-05	Sea Level Rise 6	-6,33e-05
Flood 2	6,45e-05	Landslide 4	-5,74e-06	Sinkhole 6	-5,64e-05
Global Warming 2	-5,60e-05	Mudslide 4	6,18e-05	Thunderstorm 6	2,36e-05
Hailstorm 2	2,29e-04	Pandemic 4	-9,31e-04	Tornado 6	-3,24e-05
Heat Wave 2	1,62e-04	Sea Level Rise 4	5,23e-05	Tropical Cyclone 6	3,81e-06
Landslide 2	-1,05e-05	Sinkhole 4	3,62e-05	Tsunami 6	1,92e-05
Mudslide 2	1,29e-04	Thunderstorm 4	6,49e-05	Volcanic Eruption 6	1,97e-05
Pandemic 2	-2,06e-03	Tornado 4	-4,15e-05	Wildfire 7	2,69e-05
Sea Level Rise 2	1,35e-04	Tropical Cyclone 4	7,02e-05	Avalanche 7	-4,07e-05
Sinkhole 2	9,06e-05	Tsunami 4	2,60e-05	Blizzard 7	-1,39e-05

Tabel B.11. Koefficienterne for Google Trends variablene til hvert lag opnået med PLS, del 1.

	Koefficient		Koefficient		Koefficient
Climate Change 7	2,40e-05	Blizzard 8	4,34e-05	Avalanche 9	-1,14e-04
Drought 7	2,67e-05	Climate Change 8	-1,96e-05	Blizzard 9	-5,56e-05
DustStorm 7	5,66e-06	Drought 8	4,93e-05	Climate Change 9	2,48e-05
Earthquake 7	-5,67e-05	Dust Storm 8	-3,13e-05	Drought 9	-9,83e-05
Flash Flood 7	2,47e-05	Earthquake 8	1,17e-04	DustStorm 9	4,43e-05
Flood 7	4,19e-05	Flash Flood 8	-3,13e-05	Earthquake 9	-1,38e-04
Global Warming 7	-6,97e-05	Flood 8	-3,51e-05	Flash Flood 9	3,88e-05
Hailstorm 7	6,92e-05	Global Warming 8	-1,11e-05	Flood 9	3,35e-05
Heat Wave 7	5,65e-05	Hailstorm 8	-1,38e-04	Global Warming 9	-8,31e-06
Landslide 7	-7,88e-07	Heat Wave 8	-9,40e-05	Hailstorm 9	1,72e-04
Mudslide 7	3,25e-05	Landslide 8	6,82e-06	Heat Wave 9	1,16e-04
Pandemic 7	-5,98e-04	Mudslide 8	-7,93e-05	Landslide 9	-9,46e-06
Sea Level Rise 7	6,98e-05	Pandemic 8	1,27e-03	Mudslide 9	1,01e-04
Sinkhole 7	5,12e-05	Sea Level Rise 8	-7,55e-05	Pandemic 9	-1,56e-03
Thunderstorm 7	3,48e-05	Sinkhole 8	-4,84e-05	Sea Level Rise 9	8,15e-05
Tornado 7	-2,05e-05	Thunderstorm 8	-7,70e-05	Sinkhole 9	5,22e-05
Tropical Cyclone 7	5,32e-05	Tornado 8	5,82e-05	Thunderstorm 9	9,99e-05
Tsunami 7	1,15e-05	Tropical Cyclone 8	-9,44e-05	Tornado 9	-7,25e-05
Volcanic Eruption 7	1,10e-05	Tsunami 8	-3,34e-05	Tropical Cyclone 9	1,14e-04
Wildfire 8	-3,59e-05	Volcanic Eruption 8	-3,29e-05	Tsunami 9	4,31e-05
Avalanche 8	8,97e-05	Wildfire 9	3,90e-05	Volcanic Eruption 9	4,29e-05

Tabel B.12. Koefficienterne for Google Trends variablene til hvert lag opnået med PLS, del 2.

B.2.4 Udvidet neuralt netværk for RV

Tabel B.13 viser de 45 modelkombinationer af neurale netværk for RV indekset, som er anvendt til at udvælge den optimale størrelse af netværket.

	AIC	MSE	MAE
1 RV lag			
1 Google lag			
d=1	52*	1,04e-08	6,74e-05
d=2	102	4,26e-08	1,33e-04
d=3	152	7,16e-06	5,82e-04
2 Google lags			
d=1	96	7,51e-09	7,23e-05
d=2	190	1,56e-07	1,94e-04
d=3	284	5,50e-05	8,70e-04
3 Google lags			
d=1	140	1,76e-08	7,89e-05
d=2	278	7,23e-04	3,70e-03
d=3	416	8,32e-07	2,96e-04
2 RV lags			
1 Google lag			
d=1	54	4,27e-07	2,45e-04
d=2	106	2,28e-07	2,34e-04
d=3	158	1,31e-07	2,06e-04
2 Google lags			
d=1	98	8,10e-09	6,53e-05
d=2	194	2,73e-04	2,37e-03
d=3	290	2,39e-08	1,04e-04
3 Google lags			
d=1	142	7,54e-09	6,11e-05
d=2	282	1,73e-07	2,05e-04
d=3	422	3,36e-08	9,75e-05
3 RV lags			
1 Google lag			
d=1	56	1,65e-08	1,03e-04
d=2	110	1,31e-04	3,92e-03
d=3	164	3,68e-04	2,74e-03
2 Google lags			
d=1	100	1,34e-08	8,20e-05
d=2	198	1,62e-04	1,75e-03
d=3	296	5,05e-03	7,39e-03
3 Google lags			
d=1	144	9,43e-09	6,82e-05
d=2	286	1,99e-07	1,63e-04
d=3	428	7,13e-07	3,63e-04
5 RV lags			
1 Google lag			
d=1	60	2,17e-07	2,26e-04
d=2	118	1,74e-06	3,50e-04
d=3	176	5,03e-08	1,35e-04
2 Google lags			
d=1	104	1,61e-07	1,58e-04
d=2	206	4,54e-07	9,03e-05
d=3	308	1,06e-05	1,08e-03
3 Google lags			
d=1	**148**	**6,67e-09**	**5,76e-05***
d=2	294	1,11e-08	6,70e-05
d=3	440	1,96e-07	1,78e-04
10 RV lags			
1 Google lag			
d=1	70	1,47e-07	2,01e-04
d=2	138	2,68e-07	2,36e-04
d=3	206	1,72e-06	3,50e-04
2 Google lags			
d=1	114	5,89e-09	5,83e-05
d=2	226	5,88e-09*	5,84e-05
d=3	338	4,17e-08	1,10e-04
3 Google lags			
d=1	158	1,49e-07	1,24e-04
d=2	314	2,78e-08	9,19e-05
d=3	470	1,35e-03	2,38e-03

Tabel B.13. Resultater for en række neurale netværk med lags af RV samt Google Trends variable som input. De bedste værdier er markeret med *, og den udvalgte model er markeret med fed.

Tabel B.14 viser vægtene fra det udvalgte neurale netværk.

	Vægt		Vægt		Vægt
Bias.1	-3,015	Tropical Cyclone.1	-2,149	Wildfire.3	-2,243
RV.1	-0,237	Tsunami.1	0,985	Avalanche.3	0,135
RV.2	-0,027	Volcanic Eruption.1	0,443	Blizzard.3	-1,350
RV.3	0,192	Wildfire.2	-0,722	Climate Change.3	-1,614
RV.4	0,214	Avalanche.2	-2,948	Drought.3	-0,199
RV.5	0,167	Blizzard.2	-1,817	Dust Storm.3	-3,816
Wildfire.1	-1,168	Climate Change.2	-0,495	Earthquake.3	0,373
Avalanche.1	-0,886	Drought.2	0,104	Flash Flood.3	-1,014
Blizzard.1	-0,763	Dust Storm.2	-4,341	Flood.3	-2,483
Climate Change.1	0,715	Earth quake.2	0,167	Global Warming.3	-0,317
Drought.1	-0,816	Flash Flood.2	-1,515	Hailstorm.3	-3,257
Dust Storm.1	-4,054	Flood.2	-4,631	Heat Wave.3	-3,009
Earthquake.1	-0,296	Global Warming.2	-0,437	Landslide.3	-0,867
Flash Flood.1	-0,668	Hailstorm.2	0,136	Mudslide.3	-4,604
Flood.1	-2,592	Heat Wave.2	-3,524	Pandemic.3	0,662
Global Warming.1	0,442	Landslide.2	-0,642	Sea Level Rise.3	-1,086
Hailstorm.1	-3,712	Mudslide.2	-2,769	Sinkhole.3	-0,137
Heat Wave.1	-5,268	Pandemic.2	1,698	Thunderstorm.3	-1,684
Landslide.1	-0,067	Sea Level Rise.2	-0,239	Tornado.3	1,019
Mudslide.1	-0,614	Sinkhole.2	-1,474	Tropical Cyclone.3	0,589
Pandemic.1	0,826	Thunderstorm.2	-2,328	Tsunami.3	-0,288
Sea Level Rise.1	-0,377	Tornado.2	-1,242	Volcanic Eruption.3	-0,192
Sinkhole.1	-0,326	Tropical Cyclone.2	-1,283	Bias.2	0,401
Thunderstorm.1	-0,7148	Tsunami.2	0,231	Vægt til output	0,401
Tornado.1	-0,081	Volcanic Eruption.1	-0,249		

Tabel B.14. Vægtene i det udvalgte neurale netværk for RV indekset.

The Frisch-Waugh-Lovell Therorem

C

Sætning C.1 (The Frisch-Waugh-Lovell Theorem)
Betragt regressionerne

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

og

$$M_{X_1}Y = M_{X_1}X_2\beta_2 + \varepsilon,$$

hvor $M_{X_1} = I - X_1(X_1^TX_1)^{-1}X_1^T$ *er projektions matricen, som skaber residualerne af en regression på* X_1*.*

Ordinary least squares estimaterne af β_2 *for disse to regressioner er numerisk identiske, og desuden er residualerne numerisk identiske, [Davison og MacKinnon., 2009].*