Aalborg Universitet

Forecast Analyse

Forecasting af udviklingen i de danske regioners ejendomspriser

Emil McLeman-Hasselgaard Studie nr: 20123355

Rune Lorenz Nielsen Studie nr: 20123617

Vejleder: Lasse Bork

Abstract

The Danish housing market saw a decline as a result of the financial crisis beginning in 2008. In Denmark the housing prices have been on the rise for the past several years. This situation is expected to have been brought about from the general economic developments. The purpose of this paper is to attempt to create a vector autoregressive model, founded in a series of macro-economic factors which are expected to affect the Danish housing market. The low interest rates constitute part of the explanation for the rise in housing prices. International and domestic interest rates have been declining for the past several years, which has resulted in lower user costs for property owners. Part of these user costs is property tax, and in order to pay this tax, the property owner must have a disposable income. Therefore, the model-estimation will be built on four variables: Housing prices, disposable income, real estate tax revenue and interest rate on mortgage bonds.

The selected data is examined for the presence of unit root through a Dickey-Fuller test, and the model-estimation is hence conducted. An estimation of simple autoregressive models and vector autoregressive models is carried out. The estimated models are used in forecasts for Denmark as a whole, and for the Danish regions, such as to form six VAR models and six AR models. The cause for necessity of estimating different models for the separate regions, is that the housing prices have risen faster in some parts of the country than others. The estimated models are thus applied, in order to create forecasts with 2 different forecast intervals.

The forecast values are then closely analysed, such as to facilitate an evaluation of the forecast accuracy of the models. Denmark as a whole, and the Danish regions – with the exception of Region Sjælland – get the most precise forecasts when the parsimonious AR models is applied. The best forecast models are subjected to analysis, to determine the future values for development in housing prices.
That which can be deduced, is that the generated VAR models have not been more successful in creating forecasts than the simple AR models. It is discussed whether the cause for the low accuracy in the VAR models can be found in the data from the period of observation, which was influenced by a housing bubble and a financial crisis.

Indhold

1 Indledning

Finanskrisen i 2008 førte til nedgang i det danske boligmarked, hvilket har ført til en periode med faldende priser for enfamiliehuse i Danmark. Boligmarkedet er betydningsfuldt for økonomien som helhed, da markedet påvirker den makroøkonomiske og finansielle stabilitet. De seneste år er der blevet observeret stigende boligpriser i Danmark, hvilket dog det ikke er fuldstændigt gældende, hvis der ses på de enkelte regioner. Dele af regionerne i Danmark er plaget af lange liggetider for ejendomme.
De stigende ejendomspriser der kun tilfalder nogle af regionerne, gør det interessant at analysere, om der kan ses en decideret forskel mellem regionerne. Ud fra empiri vides det allerede at huspriserne steg mest i Region Hovedstaden før finanskrisen. De lave renter der nu bliver observeret efter finanskrisen tillægger spørgsmål, om vi igen vil se store boligprisstigninger i Region Hovedstaden, eller om regionerne i f.eks. Jylland også nyder godt af lavrentemiljøet. Et alternativt udfald til at alle regionerne vil se stigning i boligpriserne, vil være at der i nogle regioner ikke vil forekomme stigninger i huspriserne, men derimod kunne der blive observeret et fald i priserne.
Boligpriserne, der bliver anvendt, er for enfamiliehuse, så der kommer diversifikation i regionerne.
Med dette projekt forsøges der at forecaste de fremtidige boligpriser på baggrund af 3 makroøkonomiske variabler. De makroøkonomiske variabler, der bliver anvendt, er den gennemsnitlige disponible indkomst, den gennemsnitlige rente på 30 års realkreditobligationer og de samlede boligskatteindtægter.
Boligpriserne afhænger i høj grad af renten, og hvis renterne forbliver lave kan det fører til positive stigninger i boligpriserne. Siden boligpriserne udvikler sig forskelligt regionerne imellem, bliver de enkelte variabler der anvendes i projektet observeret på regionalplan. En model til forecast af hele Danmark vil også blive inkluderet for at skabe et helheds billede. Projektet anvender en tidsperiode der starter i 1992 og fortsætter op til slutningen af 2015. Her ved gives der både plads til det økonomiske opsving der skete i 2000'erne, samt opsvingets kollaps i 2008. Det observerede data her derfor store udsving inkluderet, hvilket gør det muligt at tjekke robusthed af model estimationerne.

2 Metode

I dette afsnit vil metoden og opbyggelsen af projektet gennemgås, for at skabe overblik og vise tankegangen bag projektet.

En undersøgelse fra den danske Nationalbank erklærer at store udsving i de danske huspriser kan destabilisere den danske samfundsøkonomi[1]. For at sætte fokus på dette problem er det vigtigt at forstå mekanikkerne/drivkræften bag husprisudviklingen i Danmark. Da denne husprisudviklingen ikke er ens i

[1] HOUSE PRICE BUBBLES AND THE ADVANTAGES OF STABILISING HOUSING TAXATION, Klein Asbjørn side 48-49

hele Danmark, er det blevet valgt, at sætte fokus på de enkelte regionerne i Danmark. Datasættet der indeholder de valgte variabler for projektet, er observeret for hele Danmark og de fem regioner Danmark, hvor variabler for regionerne er inddelt i de respektive opdelinger fortaget af kommunalreformen fra 2007. Hele Danmark er blevet valgt for at give et overordnet overblik af, hvordan den generelle udvikling i de danske huspriserne har set ud. Opdelingen i regioner er blevet lavet på baggrund af, at kunne analysere eventuel variation i udviklingen af boligpriserne baseret på regionale tendenser. Variablerne der indgår i datasættet er fundet relevante, da de forventes at have en påvirkning på huspriserne. Relevansen er blevet udledt af en makroøkonomisk gennemgang af, hvilke makroøkonomiske variabler der har påvirkning på boligmarkedet i Danmark. Den makroøkonomiske gennemgang leder herefter over i en analyse af den fremtidige forventning af boligprisudvikling. Til analysen anvendes forecast af to forskellige modeller, hvor en almindelig autoregressiv proces sættes til at være benchmark. Her vil den anden udvalgte model blive holdt op imod benchmark-modellen.

Inden det dog er muligt at estimere de forskellige modeller skal data først testes for stationaritet. Dette gøres ved hjælp af en *Dickey-Fuller* test, der senere udvides til en *Augmented Dickey-Fuller* test. Dickey og Fuller opstiller tre forskellige ligninger, som data kan siges at følge eller være genereret af. Der anvendes en speciel metode til at finde den optimale data-genererings ligning, som data antages at være genereret af. Efter ligningen for modellen er fundet, indsættes der et højt antal lag i modellen, da augmented Dickey-Fuller tests er meget påvirkelige af lag-længden. Lag-længden bliver reduceret indtil det sidste lag for modellen er signifikant, denne metode refereres til som *general to specific*. Efter at den optimale model og lag-længde er fundet, kan data nu testes for *unit root*. Hvis der er *unit root* tilstede i dataet, kan det antages at data ikke er stationært.
Efter at der igennem augmented Dickey-Fuller testen er blevet sikret, at der ikke er *unit root* i dataet, kan modelestimeringen for de udvalgte modeller, der anvendes til at forecaste huspriserne, nu beregnes. Til dette, er der som beskrevet tidligere, blevet valgt en autoregressiv model til at være benchmark model, da en simpel model med få parametre ofte forecaster bedre end en model med flere parametre.

Alternativ modellen til den autoregressive proces er en vektor autoregressiv model (*VAR*). VAR modeller har betydeligt flere parametre der skal estimeres, i forhold til en simpel autoregressiv model (*AR*). VAR modellen er i stand til at inkludere flere variabler end en simpel AR. Der kan drages forhåbningsfulde forventninger til at VAR modellen, vil forklare den fremtidige udvikling i ejendomspriserne med en højere præcision end en parametersparsommelige model.
Efter model udvælgelsen har fundet sted og modellerne er blevet estimeret, kan selve forecastene nu beregnes. Dette gøres ved *one step ahead* metoden, hvor modellen der anvendes til forecast bliver

estimeret på et reduceret dataset, og herefter forecastet en periode frem. Dette bliver gentaget så længe der er overskydende data, der ikke bruges til model estimeringen, og derved bliver *out-of-sample forecast* genereret.

De første forecast estimeringer, der bliver dannet, forecaster for en relativ kort tidsperiode, som senere vil blive udvidet, for at teste om modellerne er robuste. Den udvidet tidsperiode, der udvælges, er perioden som ligger lige efter finanskrisen. Under finanskrisen var der forekommet et stor fald i huspriserne, og det kunne derfor være interessant at undersøge, om f.eks. VAR modellerne vil være i stand til at fange den nedadgående husprisudvikling.
Det grundlæggende analyse arbejde, der bliver fortaget, er lavet i Matlab, fordi der er mindre problematik i forhold til estimeringen, og det illustrative er mere håndterbart i Matlab.
Til sidst i projektet vil der konkluderes på analysen, samt svares på problemformuleringen.

3 Afgrænsning

Afsnittet afgrænsning vil indeholde de mest væsentlige afgrænsninger, der er fortaget i projektet. Der vil ikke blive inkluderet specifikke afgrænsninger for det udvalgte data i projektet, da dette bliver gjort under databeskrivelsen.

- Projektets primære fokus vil være på ejendomspriserne. Ejendomspriserne for enfamiliehuse bliver derfor anvendt som værende den variable, der gerne vil forecastes. Enfamiliehuse er blevet valgt på baggrund af, det antages, at der er en jævn fordeling af denne type ejendom i hele Danmark. Projektet afgrænser sig fra at anvende lejelighedspriserne. Afgrænsningen er blevet fortaget på baggrund af at nogle af yderkantsområderne i regionerne ikke vil have et stort udbud af lejligheder. Enfamiliehuse gør det muligt af få et bredere estimat af hele regionen, og det undgås derfor at få en centrering af udviklingen i huspriserne omkring byer med højt indbyggertal.

- Tidsperioden er blevet automatisk begrænset, da det ikke har været meget tilgængeligt data for de enkelte regioner. Derfor er tidsperiode for det observerede data sat til at være fra primo 1992 til ultimo 2015, fordi det er i denne periode, der er mest tilgængeligt data.

- Projektet anvender en autoregressiv model med ét lag som benchmark, da det ofte er tilfældet, at en simpel model med meget sparsomme parametre, er bedre til at forecast end en meget parametriseret model. Der gøres også brug af en mere parametriseret model. Her er en vektor autoregressiv model blevet valgt til forecast. VAR model er en standard VAR og ikke en strukturel, hvilket er blevet valgt, da

der kun vil blive fokuseret på forecasting af huspriserne. Denne afgræsning betyder også, at det ikke er muligt at tilføre choks til den danske økonomi samt de underliggende regioner.

- For at finde den bedste VAR model til forecast, bliver der foretaget en model udvælgelse. Model udvælgelsen gør brug af Akaike Information kriterium og Bayesian information kriterium, hvilket er nogle af de mest anvendte selektions kriterier til model udvælgelse. Akaike Information kriterium er et meget overordnet selektions kriterier, hvor modellen med de laveste residualer udvælgelses. Da fokus ligger på at forecaste, er det ekstra selektions kriterium i form af Bayesian information kriterium også blevet valgt til anvendelse i projektet, da Bayesian i forhold til Akaike vælger en mere sparsommelig model.

- Goyal og Welch's artikel fra 2008 bliver anvendt i projektet til at finde den kumulative sum af residualerne. Det er kun et enkelt afsnit, der bliver anvendt, hvor formelen der anvendes til at udregne CPE bliver udledt. Den resterende analyse Goyal og Welch skriver om i deres artikel bliver ikke anvendt, da det ikke har nogle relevans for projektet.

4 Makroøkonomiske faktorer:

Dette afsnit vil komme til at omhandle de makroøkonomiske faktorer, som anvendes til estimeringen af projektets VAR modeller. Det specifikke data, der benyttes i projektet, vil ikke blive gennemgået her, da dette vil komme senere i afsnittet om databeskrivelsen. Der vil i stedet blive fokuseret på de makroøkonomiske bevægelser, som ligger bag udviklingen i ejendomspriserne.

Her vil blive lagt ud med at beskrive boligmarkedet, hvor der, som der vil blive vist i næste afsnit, er en opadgående tendens.
Herefter vil der blive set på pengepolitikken, som der bliver ført i Danmark. Denne politik har stor betydning, for de muligheder aktørerne på boligmarkeder har for at danne likviditet. Dette vil lede videre til en kort undersøgelse af udviklingen i danskernes gæld, som vil blive sat i kontrast til deres formue, som værende pensionsformue og boligformue. Til sidst vil der blive set på user cost.

Skævhed i fordelingen af huspriserne

Der vil blive vist senere under afsnit 6 databeskrivelsen af variablerne for udviklingen i huspriserne, er der signifikant forskel på de gennemsnitlige huspriser i Region Hovedstaden og de gennemsnitlige huspriser i resten af landet. Dette forekommer, da skønt kun 26 procent af ejerboligkvadratmeterne i Danmark er

placeret i Region Hovedstaden, så udgør disse 40 procent af den samlede boligformue[2]. Dette er sket på baggrund af en moderate forøgelse af antallet af boliger og indvandring fra udlandet, hvilket har ført til en øget efterspørgsel på det københavnske boligmarked. Effekten af den øgede efterspørgsel har ført til at huspriserne og huslejen har været stigende, siden 90'erne og frem til finanskrisen[3]. Den øget efterspørgsel giver ligeledes mindre liggetider på ejendommene, hvilket fører til større volatilitet på ejendomsmarkedet i Region Hovedstaden end i de resterende regioner.

De stigende priser på boligerne og den stigende husleje besværliggøre det at få råd til at bo i Region Hovedstaden. Dette fører til en spredning af boligefterspørgslen, da det enkelte borger ikke er villig eller har mulighed for at betale de høje boligomkostninger. Spredningen af boligefterspørgslen giver bølgeeffekter over hele Danmark, men da udgangspunktet er København, kan den største effekt ses i Region Sjælland. Sættes fokus på boligprisudviklingen i regionerne, som ligger længere væk fra Region Hovedstaden, såsom Region Nordjylland eller Region Sydjylland, vil bølgeeffekterne være ebbet ud.[4]

Det kan være svært at se hvordan en stigning i boligpriserne i København kan have effekt på boligprisudviklingen i Region Nordjylland. Her skal der tages højde for de nationale forhold, hvor lovmæssige ændringer og udviklingen i renten på de danske realkreditobligationer, vil have en simultan effekt på boligprisudviklingen i hele Danmark. Her kan der også ses en substitutionseffekt, hvor stigningen i huspriserne fører til at efterspørgslen bevæger sig fra region til region, således prisforskellen bliver udjævnet. Derved vil stigninger i huspriserne i Region Sjælland fører til stigninger i huspriserne i Region Midtjylland.

Husprisernes træghed

I et nedadgående marked kan sælger på boligmarkedet vælge at vente med salget, indtil at huspriserne begynder at stige igen, og sælger kan få den ønskede pris. Dette gør, at boligmarkedet er et skævt marked, hvor sælger oftest har fordelene. Når efterspørgsel er høj, virker markedet til at være et meget dynamisk marked med mange salg og stigende huspriser, men når efterspørgsel er lav, betyder det ikke voldsomt faldende priser. Her ses i stedet længere liggetid for boligerne.

Dette gælder dog kun hvis boligejerens *user cost* ikke overstiger den disponible indkomst og friværdien i boligen. Hvis sælger allerede har investeret i anden bolig, som giver yderligere *user costs,* eller har af de

[2] Regionale aspekter på boligmarkedet, Hviid, Simon Juul side 48-49
[3] Regionale aspekter på boligmarkedet, Hviid, Simon Juul side 50
[4] Regionale aspekter på boligmarkedet, Hviid, Simon Juul side 51

diverse årsager, været nødt til at bo til leje i en anden bolig, har sælger øget incitament til at sælge boligen. Derfor er det dog muligt at se fald i udviklingen af huspriserne.

Den pengepolitiske rente

Renten på realkreditobligationerne hænger stærkt sammen med de pengepolitiske renter, da pengepolitikken, som der bliver ført, har stor indflydelse på langsigtet plan. Her er det i særdeleshed inflationsmålene, som har betydning for kreditorers vilje til at udlåne penge. Hvis der ikke er lagt faste lave inflationsmål, kan inflationen hurtigt komme til at overstige det afkast, som kreditorer forventer at få af deres udlån til debitorerne. Sker dette vil kreditorerne få et negativt afkast af deres investering, hvilket de selvfølgelig ikke er interesseret i. Derved vil kreditorer altid tage inflation med i deres beregninger, for at undgå at inflationen skal "spise" afkastet af investeringen. Ligeledes repræsenterer de pengepolitiske renter risikofrie passiver, som kreditorer kan vælge at investere i, frem for realkreditobligationer.

For at forstå den danske pengepolitik er det nødvendigt at forstå Danmarks fastkurssamarbejde med den europæiske centralbank, ECB. I Danmark har den danske krone været bundet op på Euroen siden den 1. januar 1999 med en centralkurs på 7,46038 kroner pr. euro. Dette skete som del af det politiske projekt *Exchange Rate Mechanism* 2, ERM2, der havde til formål at ensrettet pengepolitikken i euro-området for at forbedre handelsvilkårene på det indre markedet. Under ERM2 har medlemslandene indgået aftale om at holde centralkursen, hvor det dog er tilladt at afvige fra centralkursen med udsvingsbånd på plus/minus 15 procent. I Danmark er der blevet valgt at fastsætte udsvinsbåndet til plus/minus 2,25 procent. Dette viser en øget dedikation fra Nationalbankens side om at fastholde centralkursen, og sender signal til investorer såvel som spekulanter, at skønt der i Danmark ikke benyttes euro, så kan de forvente at valutakursen imellem den danske krone og Euroen forbliver den samme.[5]

På kortsigt vedligeholder Nationalbanken centralkursen ved at sælge eller opkøbe udenlandsk valuta, hvilket i praksis betyder, at hvis Nationalbanken kan se at der sker en appreciering af den danske krone i forhold til Euroen, så sælger de ud af den krone danske for at opkøbe euro. Dette fører til et øget udbud af danske kroner, samt en øget efterspørgsel af euro. Derved vil der ske en depreciering af den danske krone i forhold til Euroen, hvilket kontrastere apprecieringen og derved opnås centralkursen igen.

På længere sigt er det ikke tilstrækkeligt at opkøbe og sælge danske kroner for at holde centralkursen, og derfor justerer Nationalbanken på de pengepolitiske renter. Effekten af dette kan nemmest forklares ved hjælp af renteparitetet. Renteparitetet tager udgangspunkt i en valutaspekulant, som forsøger at maksimere sit afkast ved at investere i passiver på tværs af landegrænserne. Derfor har han fokus på den

[5] *Pengepolitik i Danmark*. I, Danmarks Nationalbank, side 9.

rente, som han kan få på passiverne i de givne lande, men også på de valutakursændringer, der forventes at ske. Hvis der er forventning om at den danske krone vil appreciering mod Euroen, vil denne valutaspekulant være villig til at tage en mindre rente på de danske passiver, til fordel for det afkast han vil få af valutakursændringerne. Derved opstår der en paritet mellem renten på de danske passiver R_{DKK} og renten på de europæikse passiver, $R_{€}$, samt forventning til valutakursen $E^{e}_{\frac{DKK}{€}}$:

$$R_{DKK} = R_{€} + (E^{e}_{\frac{DKK}{€}} - E_{\frac{DKK}{€}})/E_{\frac{DKK}{€}}$$ [6]

Grundet fastkurssamarbejdet er forventningen til den fremtidige valutakurs, at den må være lig den nuværende valutakurs, derfor:

$$E^{e}_{\frac{DKK}{€}} = E_{\frac{DKK}{€}}$$

Givet at dette er sandt, er renten på de danske passiver i pariteten ikke længere afhængig af forventningen til de fremtidige valutakurser. Derved kan valutaspekulanten kun skabe afkast af renten på passiverne, og renteparitetet ender således med at være:

$$R_{DKK} = R_{€}$$

Derfor kan det ses at Nationalbanken er nødsaget til at følge renteniveauet, som ECB sætter, for at fastholde de langsigtede valutakurser op ad centralkursen. Dette betyder dog ikke, at Nationalbanken og ECB's renter er nøjagtigt ens, da der ikke her er medtaget faktorer såsom risiko ved køb af statsobligationer. Her behøves ikke et tænkt eksempel, da de fleste nok vil være enige om at tiltroen til de danske statsobligationer er højere end tiltroen til nogen af de sydeuropæiske statsobligationer, såsom de græske. Derfor er de danske statsobligationer mere attraktive for investorer, da de repræsenterer større sikkerhed, og derfor er de villige til at tage et lavere afkast på disse, og dermed en lavere rente.

Eftersom Nationalbanken har overgivet deres mulighed for at justere de pengepolitiske renter til fastholdelse af fastkurspolitikken, er renterne ikke længere et instrument til at bekæmpe inflation. Derved kan der sættes spørgsmålstegn ved, om Nationalbanken har inflationen som et specifikt mål.

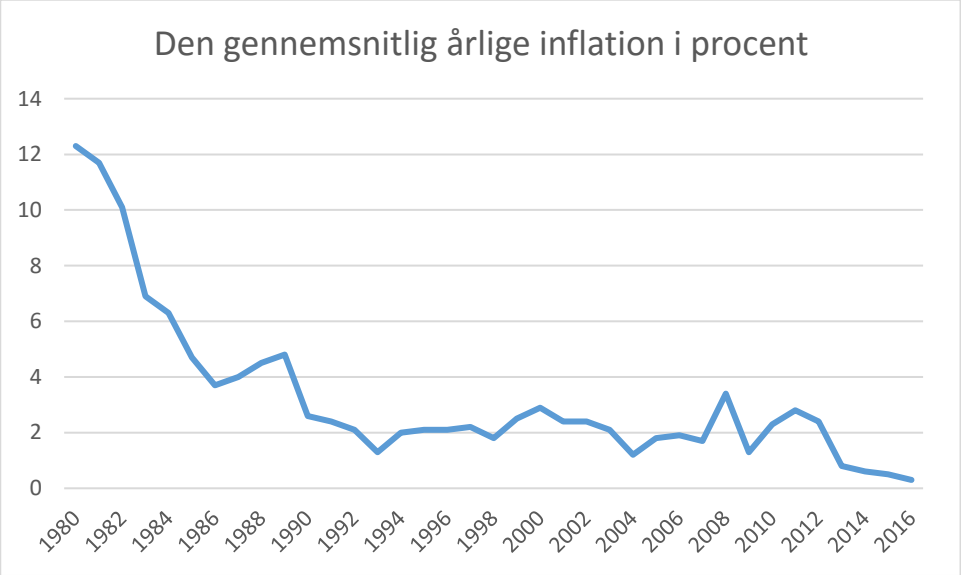
Nationalbanken anerkender dette spørgsmål, men mener dog ved at følge ECB's pengepolitik, følger de ligeledes ECB's inflationsmål på 2 procent[7]. Dette skyldes den store samhandel, der eksisterer inden for

[6] International Economics: Theory and Policy, 10th Edition, Krugman, Paul m.fl, Side 393-394.

[7] *Pengepolitik i Danmark*. I, Danmarks Nationalbank, side 9.

EU's grænser. For at se nærmere på om det er lykkes Nationalbanken at nå ECB's inflationsmål på 2 procent, er den gennemsnitlige årlige inflation illustreret grafisk i procent i den nedenstående figur 4.1:

Figur (4.1): Den gennemsnitlige årlige inflation i procent fra 1980 til 2016



Kilde: Dansk Statistik, PRIS9

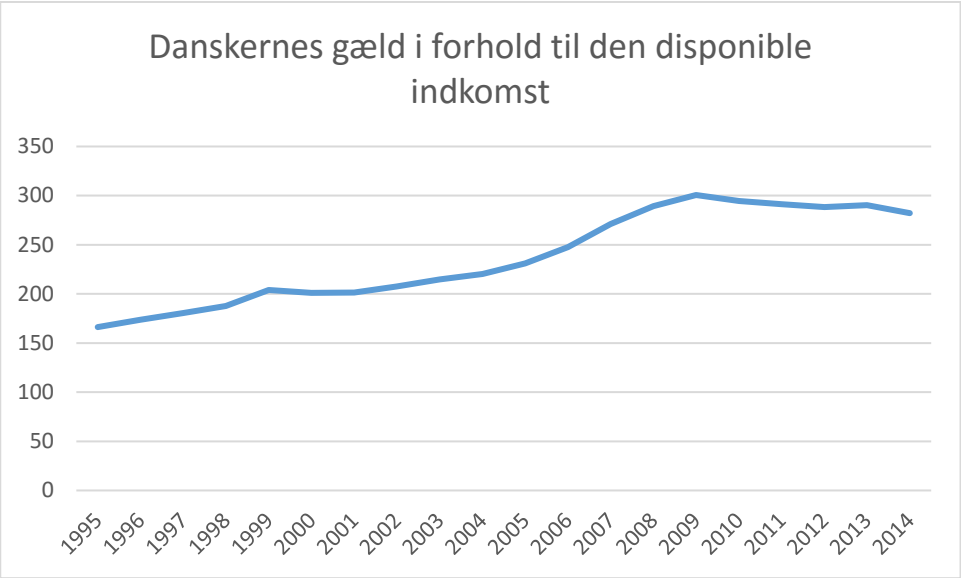
Som det kan ses i figur 4.1 har det tilnærmelsesvis lykkes Nationalbanken at holde inflationen på omkring de 2 procent. Skønt det ikke er muligt at holde inflationen helt fast på de 2 procent, er inflationen langt fra lige så høj, som den var i 80'erne. Dermed kan det siges, at det har igennem fastkurspolitikken lykkes for Nationalbanken at følge ECB's inflationsmål.

Gæld og formue:

En stor del af begrundelse for valget af data er omkostningerne ved at eje bolig, eller *user cost*, derfor vælges det i det følgende afsnit at kigge på danskernes gæld og formue i forhold til deres disponible indkomst.

I den nedenstående figur 4.2 er der illustreret danskernes samlede gæld i forhold til deres disponible indkomst. Både gælden og den disponible indkomst er her ikke regnet i faste priser, hvilket vil sige at der ikke er taget højde for inflation, da det ikke er blevet fundet nødvendigt i denne kontekst.

Figur (4.2): Danskernes gæld i forhold til den disponible indkomst 1995 til 2014



Kilde: Behandlet data fra Dansk Statistik, INDKP102 og DNMUDL

Som det kan ses i figur 4.2, har der generelt i perioden 1995 til 2014 været en stigning i danskernes bruttogæld. Stigningen i bruttogælden har været størst fra år 2001 og op til 2008, hvor finanskrisen indtræffer. Her har gælden steget fra 200 procent til 300 procent af den disponible indkomst. Dette er ikke sket på baggrund af et fald i den disponible indkomst, som der vil blive vist senere i databeskrivelsen ved figur 6.2. Det kan være at danskerne har haft gode forventninger til fremtiden, og derfor ikke set den øgede gældsætning, som værende et problem, da fremgangen i økonomien gjorde, at denne gæld kunne betales tilbage i fremtiden.
Dog topper bruttogælden i 2009, hvor der har været et fald i den disponible indkomst. Faldet i den disponible indkomst kan i sig selv have ført til et øget gældsoptag, da det enkelte individ kan have suppleret den faldende indkomst med lån. Hertil vil et fald i den disponible indkomst direkte have effekt på bruttogældens størrelse, da det relative forhold mellem gæld og indkomst forskydes.

Der kan stilles spørgsmålstegn ved, hvordan danskerne kan have en så stor bruttogæld. Her skal der ses på danskernes formue. For den almindelige dansker består deres formue af to forskellige former for formue. Den ene er danskernes boligformue, hvilket er den formue som danskerne har i form af fast ejendom. Den anden formue er danskernes pensionsopsparing eller pensionsformue.

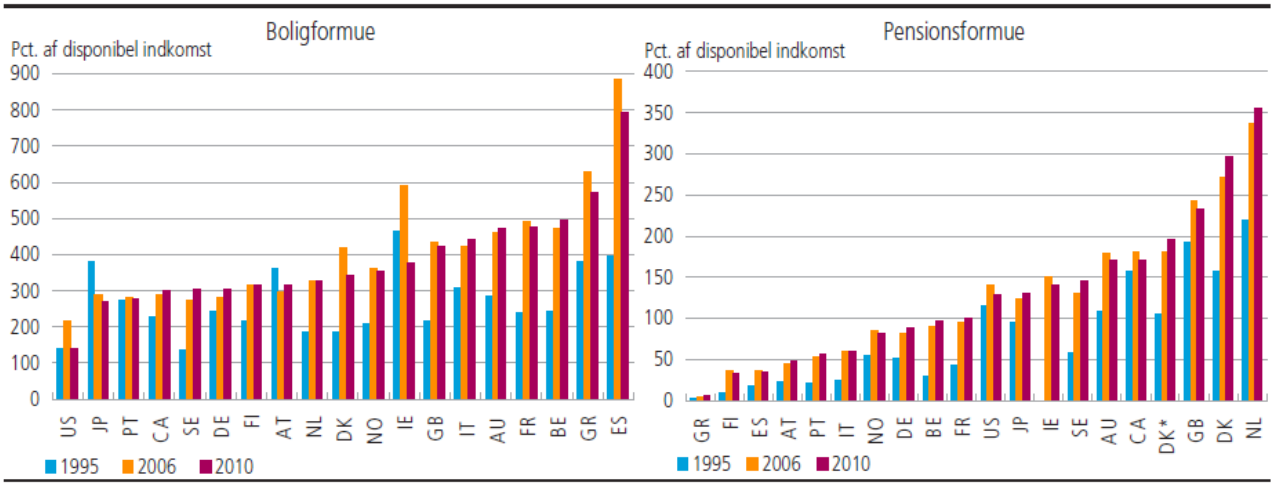
I Danmark er der, hvad Nationalbanken kalder for "et veludviklet finansielt system[8]", hvilket betyder, at de danske boligejere har bedre muligheder for at gældsætte deres formuer, end boligejere og købere i andre

[8] HUSHOLDNINGERNES FORMUE OG GÆLD, Danmarks Nationalbank

lande. Her har nyere typer af lån ligeledes givet danskere gode lånemuligheder. De afdragsfrie lån har gjort det muligt for boligkøbere, at minimere gældens effekt på deres *user cost* i en given periode, hvilket betyder, at de kan gøre en større investering i boligen. De variabelt forrentede lån giver dem muligheden for at optage billigere lån imod en større risiko. Her er endda udviklet variabelt forrentede lån med en maksimal rentesats, hvilket gør det muligt for debitor selv at bestemme størrelsen lånets risiko.

Nedenstående i figur 4.3 kan ses størrelse på disse formuer i årene 1995, 2006 og 2010 i forhold til danskernes disponible indkomst. Her er ligeledes andre OECD landes formuer.

Figur (4.3): Bolig- og pensionsformue målt i forhold til den disponible indkomst i forskellige OECD-lande



Kilde: (Nationalbanken, husholdningernes balance og gæld, s. 44)

I figur 4.3 kan det ses, at danskernes boligformue er fra 1995 til 2006 steget fra under 200 til over 400 procent af den disponible indkomst. Dette har sandsynligvis været forsaget af boligboblen, som opstod inden finanskrisen. Denne stigning i boligformuen har sandsynligvis dannet grundlagt for stigningen, der kunne i ses i bruttogælden i samme periode, da boligformue her giver et indblik i udviklingen af huspriserne i perioden. Den boligformue repræsenterer ejendomspriserne, og en stigning i disse betyder, at nye købere på boligmarkedet har været nødsaget til at optage større lån i perioden for at kunne anskaffe likviditet til at købe ejendom. Derfor kan stigningen i bruttogælden ikke ses om en hasarderet opførsel fra danskernes side, men derimod har det været nødvendigt at forøge bruttogælden, for at skabe likviditet til investeringen i boligformuen.

Hertil skal det tilføjes, at hovedparten af boliginvesteringer går til renovering af de eksisterende boliger og ikke til nybyggeri. Disse investeringer har vist sig at være meget følsomme over for konjektursvingninger,

og den store stigning i boligformuen op til finanskrisen, kom delvist på baggrund af disse investeringer. [9] Derved har det ikke kun været købere på boligmarkedet, som har investeret i boligformue. Her har de eksisterende ejere af boliger kunne se fordel i at investere i deres egen bolig, da denne investering under højkonjunkturen på kortsigt har givet et godt afkast. Her er der sandsynligvis blevet fundet likviditet til at udføre investeringerne i boligen, ved at benytte friværdien i boligen. Da investeringen i boligformuen giver en øget værdi af boligen, har det sandsynligvis ikke været svært at låne i friværdien. Den store stigning i boligformuen har givet kreditorer større incitament til at udlåne til de danske boligejere, da den høje boligformue kan dække eventuelle tab, og kan ses som sikkerhed for kreditoren. Dette kan delvist forklarer den lave rente på realkreditobligationerne, som der vil blive vist i afsnittet med databeskrivelsen.

User costs

Et begreb, som vil blive brugt gentagne gange i projektet, er *user costs*. Med *user costs* menes der de omkostninger, som er forbundet med at eje en bolig. Ifølge Nationalbanken skal dette principcelt ses som værende realrenten, et skatteelement og afskrivninger på boligen.[10] Stigende *user costs* betyder derfor, at det er forventeligt at have en negativ effekt på udviklingen i boligpriserne, da det bliver mindre attraktivt at eje en bolig. Herved giver *user costs* en stabiliserende effekt. Et eksempel på dette er realrenten, hvor stigende renter betyder, at det er dyrere at låne til køb af ejendom. Her vil købere på boligmarkedet kunne se frem til at betale en højere pris for boligen end sælgerne får, da renten skal inkluderes i den totale omkostning ved købet.

Med afskrivninger på boligen, menes der, den værdi som boligen taber over tid. Her udgør slitage en væsentlig del af disse afskrivninger, men nyere teknologi kan blive en essentiel del af den almindelige bolig. Hvis ikke ejeren af boligen investerer i vedligeholdelse af boligen, samt investerer i forbedringer af boligen, så vil boligen falde i værdi. Her kan de nye ejere af boligen, derfor se frem til en ekstra investering i boligen efter køb, og er derfor mindre villige til at give en høj pris for boligen.

Her er boligmarkedet særligt, da grunden som huset er bygget på, generelt ikke taber værdi. De den nye ejere har muligheden for at renovere den eksisterende bolig på grunden, eller rive den ned og bygge et ny ejendom. Derfor kan købere vælge boligen på baggrund af beliggenheden alene, og derved vil en boligs værdi altid være afhængig af beliggenheden. Et forfaldet hus i et attraktivt område kan derfor være meget værd, skønt ejeren eller forrige ejere ikke har investeret i vedligeholdelse af selvsamme bolig. Dette gør det særligt besværligt at måle afskrivningerne på boligen, og afskrivningerne vil derfor ikke blive taget med som del af dette projekts definition af *user cost*.

[9] Hviid, Sim Juul, Regionale aspekter, side 49
[10] *MONA - en kvartalsmodel af dansk økonomi* Danmarks Nationalbank, side 43

Skatteelementet er den beskatning, der er på at eje en bolig. I Danmark findes der 2 former for beskatning af boliger, ejendomsskatten og grundskylden, hvilket vil blive grundigere beskrevet i databeskrivelsen. Ejendomsskatten og grundskylden er en beskatning på afkastet, som ejeren af ejendommen kan få af sin investering i ejendommen, eller en beskatning af det ejeren sparer, ved at eje frem for at leje. Her gør ejendomsbeskatningen det mindre attraktivt at investerer i ejendomme fremfor at investere i aktiver, der eventuelt kunne stimulere økonomien[11].

Boligbeskatningen er en automatisk stabilisator, da stigninger i boligpriserne ligeledes fører til stigninger i boligbeskatningen, og fald i boligpriserne giver fald i ejendomsskatterne[12]. Herved har ejendomsskatterne en dæmmende effekt på volatiliteten i udviklingen af ejendomspriserne, da det bliver mindre attraktivt at eje en bolig, når boligpriserne stiger, og omvendt bliver det billigere at eje en bolig i et nedadgående marked.
De nu gennemgået makroøkonomiske faktorer, vil blive anvendt til at opstille en problemformulering.

5 Problemformulering:

Der ønskes at opstilles en vektor autoregressiv model, som er baseret på makroøkonomiske faktorer. Da det er fundet relevant at undersøge, om der er forskel i husprisudviklingen de danske regioner imellem, skal dataet til denne model kunne findes på regionalt plan. Heraf skal der dannes forecasts, som gerne skal give mere præcise og robuste resultater, der er bedre end forecasts, som er genereret ud fra en simpel parametersparsom autoregressiv model. Den endelige problemformulering for projektet er derfor følgende:

Er det muligt ved hjælp af en vektor autoregressiv model, at generere forecasts for udviklingen i de danske boligpriser, både på landsplan og regionalt, som er bedre end en simpel parametersparsom autoregressiv model?

6 Databeskrivelse

Det følgende afsnit vil omhandle de variable, som AR- og VAR modellerne, der vil blive forecastet på, vil komme til at indeholde. Heri vil der komme en gennemgang af, hvorfor netop disse variabler er blevet valgt, og hvordan udviklingen i variablen har været igennem observationsperioden. Her vil også blive set på

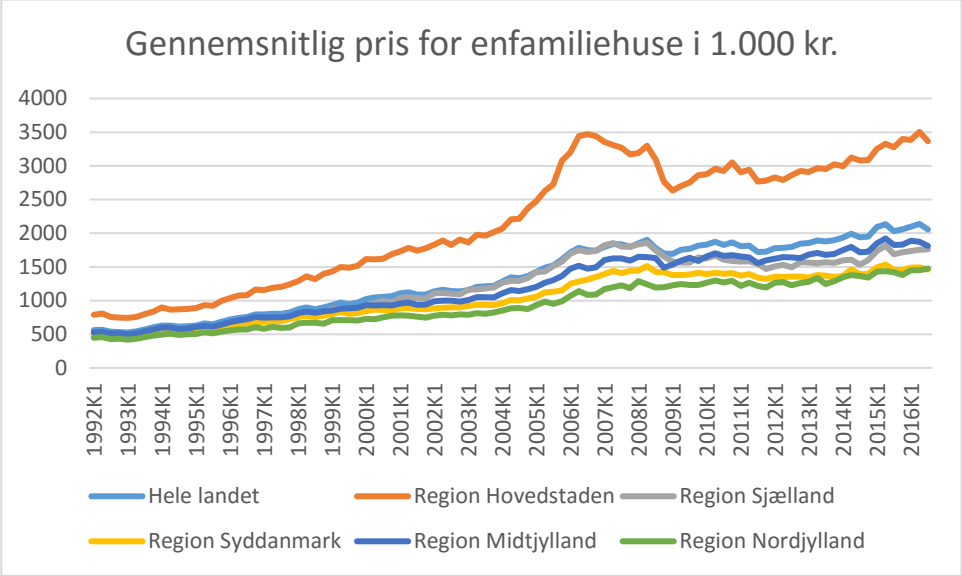
[11] HOUSE PRICE BUBBLES AND THE ADVANTAGES OF STABILISING HOUSING TAXATION, Hviid, Simon Juul, side 50
[12] HOUSE PRICE BUBBLES AND THE ADVANTAGES OF STABILISING HOUSING TAXATION, Hviid, Simon Juul, side 49 til 53

databehandlingen, som den enkelte variabel har gennemgået, samt et kritisk perspektiv på både valget og tilgangen til variablerne.

Huspriser

En essentielt del af modellen er den afhængige variabel *huspriser*. Hertil benyttes der data fra Dansk Statistik. Dette data indeholder 6 rækker med 96 observationer for de 5 regioner og hele Danmark. Disse observationer dækker over den gennemsnitlige pris for enfamiliehuse solgt i perioden ved almindelig frihandel. Observationerne er inddelt kvartalsmæssigt, og dækker over perioden første kvartal 1992 til tredje kvartal 2016. Nedenfor i figur 6.1 er det rå data vist grafisk:

Figur (6.1): Den gennemsnitlige pris for enfamiliehuse i hele Danmark og de fem regioner



Kilde: Dansk statistik, ejen77

Det mest iøjnefaldende ved figur 6.1 er, hvor meget niveauet for huspriserne i Region Hovedstaden ligger over gennemsnittet for huspriserne i hele Danmark og de resterende 4 regioner. Region Hovedstaden er tydeligvis den region, hvor boligboblen havde den største effekt, hvilket er illustreret af perioden 2005K1 til 2009K1. Dette kommer af, at boligpriserne i Region Hovedstaden er mere følsomme overfor ændringer i konjekturerne, og har derfor højere volatilitet, hvilket blev forklaret under makroøkonomiske faktorer i afsnittet om skævhed i fordelingerne af boligpriserne.

Her kan det også ses, at boligpriserne i Region Sjælland har under boligboblen gennemgået en større stigning, og har op til finanskrisen højere boligpriser end gennemsnittet for hele Danmark. Efter finanskrisen falder boligpriserne i Region Sjælland til under det gennemsnitlige niveau for hele Danmark,

hvilket viser at den høje volatilitet af boligpriserne i Region Hovedstaden, har større effekt på boligmarked i Region Sjælland end i resten af regionerne. Det er dermed bølgeeffekterne, der ligeledes blev gennemgået i skævhed i afsnittet om fordelingerne af boligpriserne, der her er illustreret.

Ses der bort fra perioden efter boligboblen, kan der spores en opadgående trend i udviklingen i huspriserne. Her skal der gøres opmærksom på, at der i figur 6.1 ikke er blevet taget højde for inflation, og der derfor vil være en mindre stigende trend, efter huspriserne er blevet omregnet til at være i faste priser.

Databehandling

For at finde huspriserne i faste priser bliver det rå data for huspriserne, *IH*, deflateret med forbrugerprisindekset PRIS112 fra Dansk statistik, *index*, hvor *index* på tidspunkt t=2015K1 er lig 100. Ejendomspriserne i faste priser, H_t, findes ved deflateringen, som foregår på følgende måde:

$$\frac{IH_t}{index_t} * 100 = H_t$$

Da *index* er opgjort på årlig basis, mens huspriserne er opgjort i kvartaler, er der manglende datapunkter for *index*. Derfor er det nødvendigt at interpolere observationerne for *index*, og derved generere de manglende datapunkter. Dette bliver gjort ved hjælp af *cubic-spline*, hvilket vil blive forklaret senere i afsnit (7) *Interpolering*, da denne metode til dannelse af datapunkter er essentielt for variablerne *indkomst* og *skat.*

Da VAR modellernes fokus ligger på udviklingen i ejendomspriserne, er det mere relevant at finde realvæksten i den givne periode, h_t, fremfor den faktiske værdi af ejendomspriserne. Derfor beregnes realvæksten, hvilket her bliver set som, differensen mellem logaritmen til den givne observations værdi og logaritmen til den observerede værdi i den foregående periode. Her benyttes logaritmen til H_t, for at normalisere dataet, således at observationer, som har værdier, der ligger langt fra den gennemsnitlige værdi af observationer, kommer tættere på gennemsnittet. Herved transformeres eksponentielle trends ligeledes til lineære trends.

$$h_t = logH_t - logH_{t-1}$$

Grundet problemer med stationaritet i det observerede data for nogle af de andre variable, vil der under konstruktionen af VAR modellerne og AR modeller blive fjernet endnu en af de første observationer, således datasættets begyndelse ligger i 3. kvartal 1992.

Kritik af data:

Da variablen beskriver gennemsnitsprisen, kan den derfor kritiseres for ikke at kigge på de faktiske ejendomsværdier, men derimod omsætningen på ejendomsmarkedet. Omsætningen på netop ejendomsmarkedet er speciel, som beskrevet tidligere, da sælger af boligen ofte kan vælge at blive siddende i boligen, indtil sælgeren får tilbudt den pris for boligen, som sælger forlanger. Derved kan der på ejendomsmarkedet opstå trægheder, når markedet er i en nedgående retning, som vil i dette data give en højere pris på boligmarkedet end den faktiske værdi.

Boligboblen i perioden 2005K1 til 2009K1 kan give problemer i forhold til at estimere VAR modellerne, da der her er tale om en unik situation, hvor huspriserne stiger med mere end, hvad der kunne forventes i forhold til den udvikling, der kan spores i de øvrige perioder.

Et andet væsentligt kritikpunkt er interpoleringen af indekset, som bruges til deflateringen. Da indekset, der benyttes, er forbrugerprisindekset, som er baseret på den gennemsnitlige pris for varer og tjenester[13], giver indekset ikke en klar forståelse af, hvor meget den gennemsnitlige dansker vælger at bruge på bolig. Dog giver indekset en klar fornemmelse af de inflationære træk, der er i samfundet, og da deflationen benyttes for at direkte tage den inflationære trend ud af beregningerne, er det forbrugerprisindekset, som er blevet valgt til dette projekt.

Andet data med relevans:

Her kunne det have været relevant at kigge på ejerlejlighedspriser, da ejerlejligheder har en meget større tendens til at ligge i byerne, og i særdeleshed storbyerne, kontra yderområderne hvor det er parcelhusene, som er dominerende[14]. Derved kunne der blive sat et andet perspektiv i projektet med storbyerne kontra yderområdet. Dog ville det ikke være muligt at kigge på den disponible indkomst, da den enkelte person kunne bo i en ejerlejlighed i byen og samtidigt arbejde i et yderområde. Her ville denne persons indkomst blive tilskrevet indkomst niveauet i byen, skønt indkomsten kommer udefra. Ligeledes ville en person kunne bo i et parcelhus i yderområdet og arbejde i storbyen. Herudover kan det formodes at parcelhuspriserne i byerne følger ejerlejlighedspriserne. Dette ville betyde, at der ville blive observeret en falsk bølgeeffekt, hvor prisstigninger på boligerne i storbyerne ville påvirke det observeret data for boligpriserne i yderområderne. Derfor giver det mere mening at måle ejendomsprisudviklingen på regionalt basis.

[13] FORBRUGERPRISINDEKS, Danmarks Statistik.
[14] Hviid, Sim Juul, Regionale aspekter, side 48

Indkomst

Variablen *indkomst* dækker over danskernes gennemsnitlige disponible indkomst i perioden 1992k3 til 2015k4. Hertil benyttes der data fra Dansk Statistik. Dette data indeholder ligesom variablen for huspriser 6 rækker med 94 observationer for de 5 regioner og hele Danmark.

Danskernes disponible indkomst burde have en stor effekt på de danske huspriser, da det er med udgangspunkt i indkomsten, at køberne på ejendomsmarkedet finder størrelsen på investeringen i ejendommen. Særligt på det danske ejendomsmarked er det økonomiske råderum relevant, da investeringen ofte forekommer igennem lån, hvilket medfører en fast fremtidig udgift i form af afdrag og rentebetalinger. Hertil kommer ligeledes ejendomsskatten og grundskylden, der giver fast udgift ved at eje en bolig, hvilket som forklaret tidligere bliver i projektet defineret som *user costs*.
Derfor forventes det at den disponible indkomsten har en positiv indflydelse på huspriserne, hvilket vil sige at en stigning i den disponible indkomst vil føre til en stigning i ejendomspriserne.

Databehandling

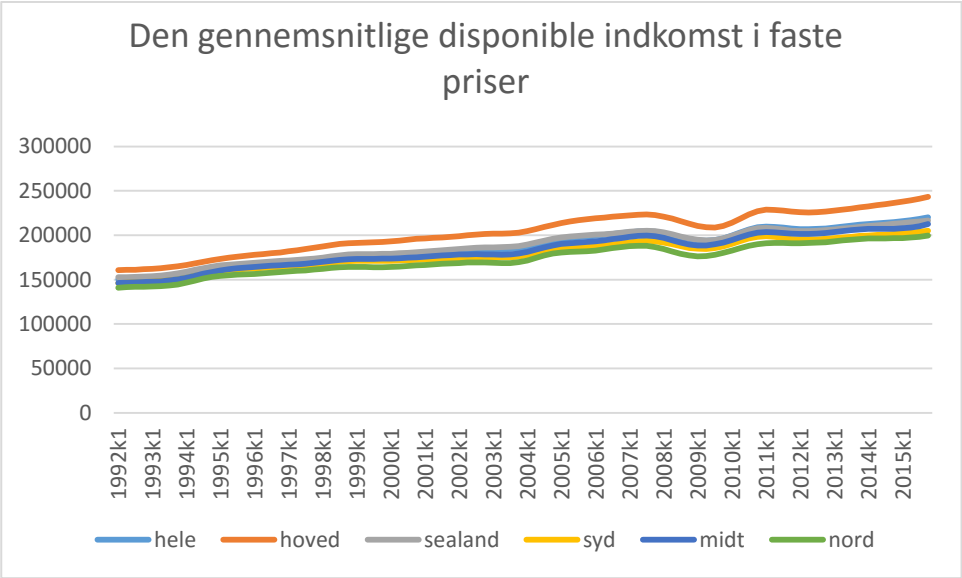
Da ejendomspriserne i modellen bliver bearbejdet i faste priser, er det nødvendigt at ligeledes omregne indkomsten til faste priser. Her benyttes det samme indeks, *index*, som deflator, men med den lille ændring at der her ikke forekommer en interpolering af indekset inden deflationen. Dette skyldes, at observationerne i det rå data for både *index* og indkomsten er på årligt basis.

$$\frac{IY_t}{index_t} * 100 = Y_t$$

Da den gennemsnitlige disponible indkomst er opgjort på årligt basis, er det nødvendigt at foretage en interpolering af det eksisterende data, for at omregne de årlige observationer til kvartalsmæssige observationer. Denne interpolering sker ved hjælp af *cubic spline*, som der sagt tidligere, vil blive forklaret senere.

Efter at dataet for den gennemsnitlige disponible indkomst er blevet deflateret til faste priser, og de manglede observationer er blevet fundet ved hjælp af interpoleringen, er dataet klar til databehandlingen, der er nødvendig for at estimere VAR modellerne. Nedenstående i figur (6.2) er dette data blevet illustreret grafisk, for at give et billede af udviklingen i den disponible indkomst.

Figur (6.2): Den gennemsnit disponible indkomst i faste priser fra 1992K1 til 2015K1



Kilde: Behandlet data fra Dansk Statistik, INDKP102

Skønt der er taget højde for inflation i figur (6.2), viser figuren at der er en opadgående trend i danskernes disponible indkomst. Denne opadgående trend i dataet formodes at være grunden til problemerne med stationaritet, som der vil blive taget højde for senere i projektet. Det kan også ses i figur (6.2), at niveauet for den gennemsnitlige disponible indkomst i Region Hovedstaden er højere, og har en kraftigere stigning, end niveauet for de resterende regioner og dermed også gennemsnitlige disponible indkomst for hele Danmark.

For at finde udviklingen i den gennemsnitlige disponible indkomst, findes differensen mellem logaritmen til den givne observationsværdi for den gennemsnitlige disponible indkomst på tidpunkt *t,* og logaritmen til den observerede værdi for den gennemsnitlige disponible indkomst i den foregående periode *t-1*. Her benyttes logaritmen til Y_t, af samme grunde som forklaret tidligere under databehandlingen af huspriserne.

$$y_t = logY_t - logY_{t-1}$$

Grundet stationaritet i det observerede data for udviklingen i den gennemsnitlige disponible indkomst, har det været nødvendigt at tage differencen mellem udviklingen i den gennemsnitlige disponible indkomst på tidspunkt *t,* og udviklingen i den gennemsnitlige disponible indkomst i den foregående periode *t-1*. Nødvendigheden af dette trin vil blive gennemgået senere i afsnittet om *Augmentet Dickey-Fuller* tests.

$$\Delta y_t = y_t - y_{t-1}$$

Derved bliver der ikke længere set på udviklingen i den gennemsnitlige disponible indkomst, men derimod accelerationen i den gennemsnitlige disponible indkomst. Det betyder, at en negativ værdi for Δy_t ikke nødvendigvis betyder et fald i den gennemsnitlige disponible indkomst, men derimod et fald i væksten af den disponible indkomst. Dette kan virke forvirrende for fortolkningen af variablen, men hvis den enkelte boligkøber kan se et fald i stigningen af dennes indkomst, vil vedkommenens forventninger til den fremtidige indkomst ligeledes falde. Forventningerne til den fremtidige indkomst har stor indflydelse på vilje til at investere i bolig, og derved størrelse af investeringen i ejendommen. Derfor forventes det, at accelerationen i udviklingen af den disponible indkomst har en signifikant effekt på udviklingen i huspriserne.

Kritik af data

Interpoleringen, som der er blevet foretaget for at danne de manglede observationspunkter, er sket ved at benytte det årlige data for danskernes disponible indkomst. De manglede observationspunkter er herefter dannet imellem årlige observationspunkter, og kommer til at repræsentere de kvartalsmæssige observationer. Dermed dannes variablen indkomst på baggrund af data, som har en fire gange så stor værdi end den oprindelige observerede værdi, da indkomsten ikke er blevet fordelt udover året. I stedet er der fundet de værdier, som den gennemsnitlige årlige indkomst er estimeret til, havde den været målt på det givne tidspunkt. Dette forventes dog ikke at havde nogen effekt på estimationen af VAR modellerne, da det her efter databehandlingen er accelerationen i den gennemsnitlige disponible indkomst, som der arbejdes med.

I VAR modellerne for de enkelte regioner, kan der opstå *spild-over*-effekter, da en beboer i en region kan have arbejde i en anden region. Da variablen *indkomst* dækker over beboerne i regionen, og ikke arbejdspladserne, vil indkomsten fra den anden region blive tildelt regionen, som beboeren bor i. Skønt indkomst-mulighederne i den anden region er bedre, har beboeren fundet den første region mere eftertragtet. Dermed forventes der ikke, at disse *spild-over*-effekter vil have den store effekt, da personer, der arbejder med lavt lønnet arbejde, kan finde de lave huspriser i den anden region mere attraktive.

Andet relevant data

Det har været med i overvejelsen af datavalget, at benytte den totale gennemsnitlige indkomst. Her kunne indkomstvariablen have givet et renere indblik i væksten i samfundet, da den totale gennemsnitlige indkomst ikke tager højde for finanspolitiske tiltag, som har effekt på den disponible indkomst. Dette indblik stammer fra virksomhedernes vilje til at investere i dansk arbejdskraft.

Fordelen ved at benytte den disponible indkomst fremfor den totale gennemsnitlige indkomst er, at der kan forekomme underliggende ændringer i indkomststrukturen, som den totale gennemsnitlige indkomst

ikke tager højden for, men den disponible indkomst gør. Her kan der for eksempel være tale om ændringer i indkomstskatten eller afdragsordninger, som kan forøge den disponible indkomst, og dermed give et større økonomisk råderum for den enkelte. Det forøgede økonomiske råderum kan få betydning for huspriserne, da købere på ejendomsmarkedet vil have mere likviditet til at dække deres *user costs*, og derfor er mere villige til at give en højere pris for en ny bolig. Da den totale gennemsnitlige indkomst bliver beregnet før skat, vil disse skatteændringer ikke blive en del af estimationen af VAR modellen, og derved opfanger VAR modellen ikke de underliggende ændringer i indkomststrukturen, som kan have en essentielt betydning for udviklingen i ejendomspriserne.

Når der fokusere på danskernes mulighed for at investere i ejendom, så ville det havde været meget relevant at medregne danskernes formue. Dette er der dog ikke blevet taget højde for under denne indkomstvariabel. Boligkøbernes formue har stor indflydelse på, hvilken pris de er villige til at betale for en ejendom, da denne giver dem øget likviditet til at investere. Denne formue kan også have effekt på længere sigt, da personer med en stor pensionsopsparing vil have bedre mulighed for at betale for de fremtidige *user cost*, der er ved at eje boligen.

Dog betyder en højere indkomst, at mulighed for at øge opsparingsraten, og dermed formuen, er bedre. Ligeledes forventes danskernes pensionsformue at følge deres disponible indkomst. Derfor vælges der, at blive fokuseret på den disponible indkomst i denne projekt.

Renten

Da de fleste boligkøbere finder likviditet til investeringen i boligkøb ved hjælp af realkreditlån, har renten på de 30-års realkreditobligationer stor betydning for udviklingen på det danske boligmarked, da renten på disse obligationer udgør en stor del af boligejeres *user cost*.

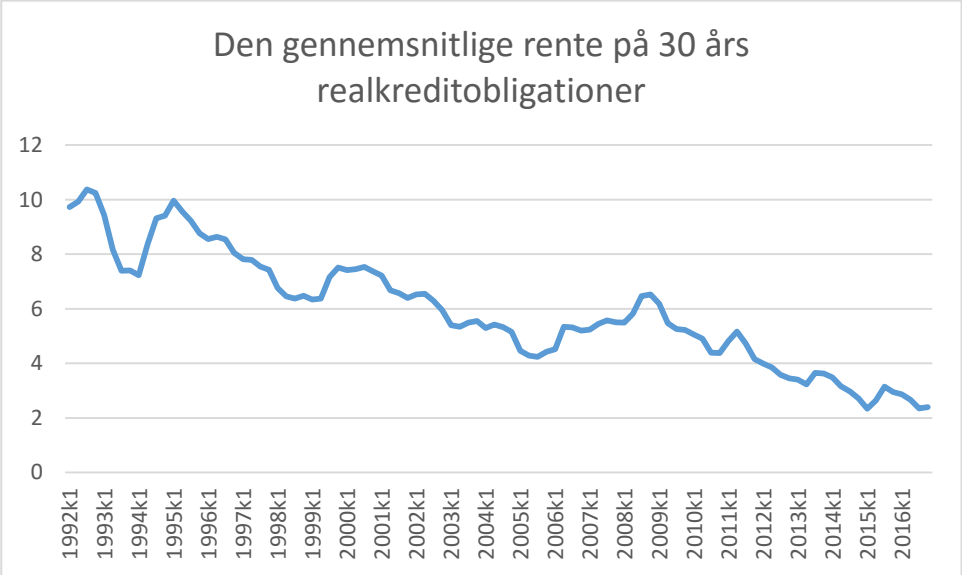
Dataet indeholder kun en række med 96 observationer, da renten formodes at gælde generelt for hele Danmark, og derved også for de enkelte regioner. Datagrundlaget er baseret på dels den gennemsnitlige rente for 30 års realkreditobligationer fra Dansk Statistik, DNRENTM, hvor observationerne er månedsobservationer. Da det resterende data er kvartal observationerne, er det derfor nødvendigt at omregne renten, således renten også bliver per kvartal. Dette gøres ganske simpelt ved at tage gennemsnittet af observationerne i månederne af et kvartal. Da dette data kun har observationer frem til 11. måned i 2012, har det været nødvendigt at finde data fra 2012 4. kvartal og frem. Her benyttes data fra Realkreditrådet[15], hvilket er på ugentligt basis. For at ligeledes omregne observationerne til kvartalsmæssige observationer, benyttes samme metodologi som tidligere. Her beregnes et kvartal fra start

[15] Obligationsrenter. Realkreditrådet.

i den påbegyndte uge til og med den sidste uge, som har udløb i samme kvartal. Dermed medregnes den sidste uge i kvartalet ikke, hvis ugens udløb ligger i det næste kvartal.

Nedenfor i figur (6.3) er den gennemsnitlige rente på de 30 års realkreditobligationer illustreret grafisk, for at give et bedre overblik over udviklingen i perioden.

Figur (6.3): Den gennemsnitlige rente på 30 års realkreditobligationer fra 1992K1 til 2016K1



Kilde: Data fra Dansk Statistik, DNRENTM, og Realkreditrådet.

Som det kan ses i figur (6.3) har renteniveauet i perioden været stærkt faldende. Dette giver boligkøbere lukrative forhold, hvis de mangler likviditet til deres investering i ejendom, da den faldende rente på realkreditobligationerne giver mindre *user costs* ved køb af ejendommen.

Da den danske Nationalbank er indgået i et fastkurssamarbejde med den europæiske centralbank, ECB, stammer de lave rente på statsobligationer fra ECB's pengepolitik, som forventes at fortsætte fremadrettet.[16] Grundet den lave vækst i dele af euro-området, har ECB valgt at sænke renten på deres statsobligationer, og derved undgå deflation.

Forventningen om at renten på de risikofrie passiver forbliver lav i den nærmeste fremtid, giver kreditorer øget incitament til at udlåne til de danske boligkøbere, frem for at investere i statsobligationer der giver et lavere afkast. Derfor er forventningen til renten på realkreditlånene, at den forbliver lav.

[16] *Årsberetning 2016,* Den Europæiske Centralbank, side 43

Databehandling

For at finde udviklingen i den gennemsnitlige rente på 30 års realkreditobligationer, beregnes differensen mellem logaritmen til den givne observations værdi for renten på 30 års realkreditobligationer på tidspunkt *t,* og logaritmen til den observerede værdi for rente på 30 års realkreditobligationer i den foregående periode *t-1*. Her benyttes logaritmen til R_t, af samme grunde som forklaret tidligere under databehandlingen af huspriserne.

$$r_t = logR_t - logR_{t-1}$$

Kritik af data

Da formålet med at benytte de 30 års realkreditobligationsrenter, er at finde *user costs* ved boligkøb, er en væsentlig kritik af at benytte 30 års realkreditobligationsrenter, at disse ikke korrekt afspejler omkostningerne ved at optage lån. Her er der nemlig ikke taget højde for de omkostninger, der ligger i at udstede en realkreditsobligation. I Danmark står realkreditinstitutterne for forvaltningen af realkreditobligationerne. Denne ydelse er selvfølgelig ikke gratis, og realkreditinstitutterne tager sig betalt ved dels at pålægge gebyrer ved udstedelse af realkreditobligationerne, men også ved at pålægge debitorerne en ekstra rente på hovedstolen. Denne rente bliver kaldt for ydelsesbidraget. Ved at sammenlægge renten på realkreditobligationerne, ydelsesbidraget og omkostningerne ved oprettelsen af lånene vil det være muligt at finde realrenten. Realrenten er hermed den totale forrentning af boligejernes lån, og et mere præcist mål for *user costs*. Dog har dette ikke været muligt at finde data, der beskriver realrenten inden for den givne tidsperiode, og da ydelsesbidraget og omkostningerne ved udstedelse af realkreditobligationerne anses for at være en mindre del af *user cost*, er renten på realkreditobligationer blevet fundet passende.

Andet relevant data:

Der kunne også være valgt at arbejde med risikofrie aktiver eller passiver såsom statsobligationer. Da investorer ofte benytter disse renter til at vurdere, hvilket afdrag de skal tage i bytte for den risiko, der eksisterer. Dette ville give et indblik i pengepolitiske tiltag, som kan give kreditorer øget incitament til at udlåne, og derved give mere fordelagtige vilkår for debitorerne.

Fordelen ved at tage renten på de 30 års realkreditobligationer kontra statsobligationer er, at det kan nemt forestilles, at der er større tillid til den danske stat, end der er til de danske boligejere. Denne forskel i tilliden har stor betydning for niveauet for den givne rente. Hvis tilliden til debitor er lav, vil kreditor have lavere incitament til at udlåne penge til debitor. Derfor vil kreditor kræve en højere rente på den udstedte

obligation, og derved få et højere afkast for den øgede risiko, som kreditor føler, der er blevet påtaget. Derfor vil renten på realkreditsobligationer altid være højere end renten på de risikofrie passiver. Hvis denne renteforskel var konstant, ville renteforskellen ikke være noget problem, da det er udviklingen i renten, som der fokuseres på. Problemet ligger i, at tilliden til de danske realkreditobligationer ikke er en konstant størrelse, og derfor vil der forekomme udsving i realkreditobligationsrenten, som ikke vil kunne ses på renten af de risikofrie aktiver. Derfor vil renten på de risikofrie passiver ikke korrekt afspejle de lånevilkår, som boligkøbere oplever, og derved vil det ikke være den korrekte effekt af renteudviklingen, som vil blive estimeret i modellen.

Når der bliver snakket om rentesvingninger og tillid, er det ikke muligt at undgå realkreditobligationer, som har et refinansieringselement. Her kunne det have været relevant at kigge på renten på de såkaldte F1- og F2-lån, som bliver refinansieret henholdsvis hvert og hvert andet år. Da disse lån er ganske almindelige i Danmark, ville denne variable rente være stærkt relevant. Her kunne det være muligt at danne en rente-variabel, som indeholdte både renten for de fast forrentede og de variabelt forrentede lån. Dog ville det blive noget kompliceret at beregne, hvor stor indflydelse de forskellige lån, bør have på denne variabel. Hertil kan det forventes at renten på disse lån følger renten på de fastforrentede realkreditobligationer, og derfor fokusere dette projekt kun på de 30 års realkreditobligationer.

Skat

Til at måle effekten af ejendomsskatterne benyttes variablen *skat*. Dette data indeholder ligesom variablen for huspriser 6 rækker med 96 observationer for de 5 regioner og hele Danmark. Datagrundlaget stammer fra Dansk Statistik, hvor periode 2007 til 2016 kommer fra databasen *EJDSK1*. Da dette data ikke rækker længere tilbage end 2007, har det været nødvendigt at finde tidligere datapunkter fra en anden database. Her er perioden 1992 til 2006 dækket af *ESKATX*, hvor variablerne for de 5 regioner bliver dannet ved at lægge de samlede ejendomsskatteindtægter fra de enkelte kommuner i de fremtidige regioner sammen. Dette skift i databaser, er sandsynligvis sket på baggrund af kommunalreformen i 2007.

Ejendomsværdibeskatning har en direkte stabiliserende effekt på boligprissvingninger, da en fast procentmæssig beskatning af boligernes værdi har en kontraktiv effekt på et opadgående marked, og en lempende effekt på et nedadgående marked. Dette betyder i praksis, at det bliver mere attraktiv at eje en bolig, skønt boligformuen er faldende, og mindre attraktiv at eje bolig, mens boligformuen er stigende.

I Danmark er der 2 boligskatter, som bliver pålagt de danske boligejere. Den ene er grundskylden, der er en beskatning på grunden alene, hvoraf skatteindtægterne til kommunen, som den pågældende bolig ligger i. Størrelsen på grundskylden bliver fastlagt af kommunen, men skal dog ligge imellem 16 og 34 promille. Den

anden skat er ejendomsværdiskat. Siden 2003 har skattestoppet betydet at ejendomsvurderingerne, som ligger til grund for ejendomsværdiskatten, har været fastfrosset på 2002 priser. Dermed har de danske boligejere ikke set en nominel stigning i deres ejendomsværdiskat. Hermed er den stabiliserende effekt, som ejendomsværdibeskatningen burde have haft, blevet fjernet. [17]

Endvidere har fastfrysningen af ejendomsværdiskatten negative effekter på boligmarkedet, hvis efterspørgslen falder. Da fastfrysningen af ejendomsværdiskatten betyder at den procentmæssige beskatning falder i takt med de nominelle priser på ejendommene stiger, vil den procentmæssige beskatning stige når huspriserne falder. Hvilket betyder, at i et nedgående ejendomsmarked vil beskatningen fylde en større del af *user costs* for boligejere, og dermed vil det blive mindre attraktivt at købe bolig.

Databehandling

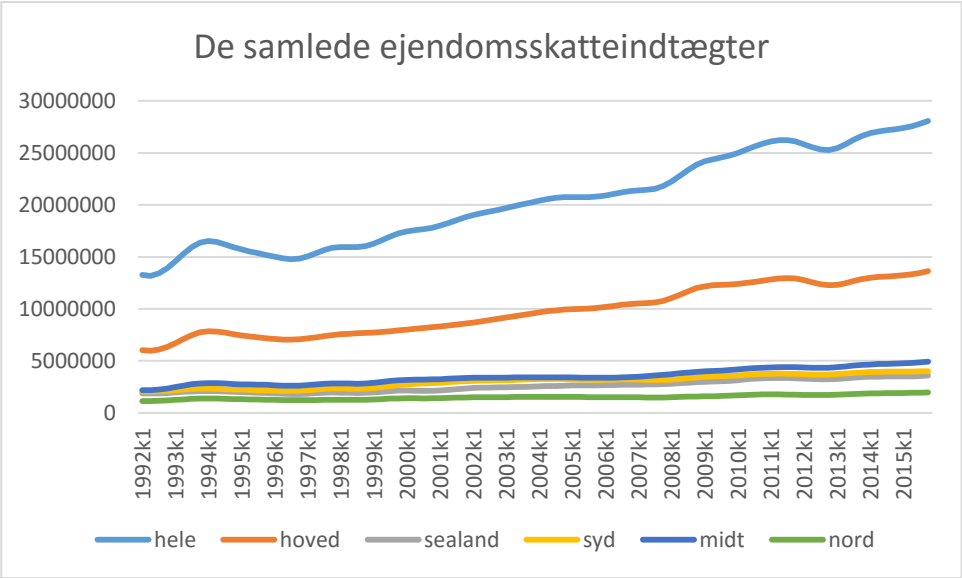
Først beregnes ejendomsskatter *IS* om til faste priser, ved at deflatere ejendomsskatterne med samme forbrugerprisindeks, som er blevet benyttet til variablerne for udviklingen i huspriserne og den gennemsnitlige disponible indkomst.

$$\frac{IS_t}{index_t} * 100 = S_t$$

Da ejendomsskatterne bliver opgjort årligt, betyder det, at der ikke er nok observerede værdier for skat til at estimere VAR modellerne. Derfor har det, ligesom ved variablen for indkomst, været nødvendigt at foretage en interpolering af det observerede data, for at omregne de årlige observationer til kvartalsmæssige observationer. Denne interpolering sker ligeledes ved hjælp af *cubic spline*, som vil blive forklaret senere.

[17] Ejendomsværdiskat og ejendomsskat (grundskyld). Skatteministeriet.

Figur (6.4): De samlede ejendomsskatteindtægter i perioden



Kilde: Data fra Dansk Statistik, ESKATX og EJDSK1.

Hele Danmark har naturligt nok de største samlede skatteindtægter, da de samlede skatteindtægter for hele Danmark består af en opsummering af dataet fra de 5 regioner. Udover dette kan det ses at ejendomsskatterne fra Region Hovedstaden er signifikant større, hvilket skal ses i lyset af ejendomspriserne i området er større end de resterende regioner, som det kan ses i figur (6.1). Hertil er Region Hovedstaden den mest befolkningsrige region, og derfor er antallet af ejendomme som er skattepligtige højere end i de resterende regioner.

Med undtagelse af perioden efter finanskrisen kan der ses en jævn stigning i skatteindtægterne. Det er påfaldende at ejendomsskatteindtægterne stiger under finanskrise efter at boligboblen er sprunget. Her kan finanskrise betydet, at kommunerne har oplevet at lavkonjekturen har ført til et fald i deres skatteindtægter, og derved har kommunerne sat grundskylden op for at kontrastere dette.

Databehandling

Igen skal der findes udviklingen i det generelle ejendomsskatteniveau, hvilket findes ved at tage differensen mellem logaritmen til den givne observations værdi for de samlede ejendomsskatteindtægter på tidspunkt *t,* og logaritmen til den observerede værdi for de samlede ejendomsskatteindtægter i den foregående periode *t-1*. Her benyttes logaritmen til S_t, af samme grunde som forklaret tidligere under databehandlingen af huspriserne.

$$s_t = logS_t - logS_{t-1}$$

Grundet problemer med stationaritet i det observerede data for udviklingen i ejendomsskatterne, har det været nødvendigt at tage differencen mellem udviklingen i ejendomsskatterne på tidspunkt *t* og udviklingen i ejendomsskatterne i den foregående periode *t-1*. Nødvendigheden af dette trin vil blive gennemgået senere i afsnittet om *augmentet Dickey-Fuller* tests.

$$\Delta s_t = s_t - s_{t-1}$$

Ligesom ved udviklingen i den disponible indkomst, bliver der ikke længere set på udviklingen i de samlede ejendomsskatteindtægter, men derimod accelerationen i udviklingen i de samlede ejendomsskatteindtægter. Her kan voldsomt stigende ejendomsskatter skabe en dårlig stemning på boligmarkedet, hvor udgifterne ved at eje ejendom kan stige mere end forventet. Dette giver usikkerhed og faldende ejendomspriser.

Kritik af data

Ligesom ved variablen *indkomst*, kan der føres kritik af den interpolering der benyttes på variablen *skat*. Her er der ligeledes blevet estimeret de værdier, som de samlede ejendomsskatteindtægter ville have haft, havde de været målt på det givne tidspunkt, og ikke de faktiske ejendomsskatteindtægter der fundet sted. Dog er fokusset igen på udviklingen i de samlede ejendomsskatteindtægter, og derfor forventes dette ikke at have nogen effekt.

Da der ses på de samlede ejendomsskatteindtægter, betyder det, at skat på ejerlejligheder er tilstede i dette data, skønt priserne på ejerlejligheder ikke er en del af variablerne. Ejerlejlighedsejere betaler mindre grundskyld end parcelhusejere, da ejerlejlighedsejere deler grunden, og dermed grundskylden, med andre beboere i et lejlighedskompleks. Set i lyset af fastfrysningen af skatterne på ejendomsværdien, samt en fleksibel beskatning på grundskylden, betyder det at ejerlejlighedsejere ikke oplever stor effekt af ændringer i de samlede ejendomsskatteindtægter. Derfor er stigningen i de samlede udgifter til ejendomsskatter hovedsageligt tilgået parcelhusejere, og derfor anses problemet med at medtage ejendomsskatteindtægter fra ejerlejlighed ikke for at være alarmerede.

Andet relevant data

Som del af overvejelserne bag valget af data til variablen for *skat,* er det blevet overvejet, om det kunne være muligt at se på det relative forhold mellem de totale boligværdier og de totale skatteindtægter i de enkelte regioner og kommuner.

Da ejendomsvurderingerne, som ligger til grund for ejendomsværdiskatten, har været fastfrosset på 2002 priser, betyder det, at kommuner der har oplevet høj vækst i ejendomspriserne siden 2002, har mærket en mindre effekt af ejendomsskatten end kommunerne, som har oplevet mindre vækst eller fald i

ejendomspriserne. Derfor kan det relative forhold imellem skatteindtægterne og ejendomsværdierne være stærkt nedgående i kommunerne med høj stigning i ejendomsværdierne, skønt grundskylden i disse kommuner kan være stigende, og derved kan boligejere i disse kommuner kunne se en stigning i deres *user costs*.

7 Interpolering

At observere data for udvalgte variabler er ofte vanskeligt. Det er særligt vanskeligt at finde nok data til at kunne estimere modeller på baggrund af. Det kan derfor være nødvendigt at bruge en metode til at udvide sit datasæt. Interpolering er en af disse metoder, og bruges til at forbinde 2 data punkter. I projektet er der for flere forskellige variabler som har en begrænset mængde af data. Eksempelvis er noget af det anvendte data, observeret på årlig basis, mens andre variabler er på kvartal eller månedlig basis. Som det bliver tydeliggjort i databeskrivelsen, er det blevet valgt at det observerede data skal være på kvartalsmæssigt basis. Det har derfor været nødvendigt at anvende interpolering til at udlede de manglede data punkter. Den valgte interpoleringsmetode vil i dette afsnit blive gennemgået. Til sidst i afsnittet bliver der lavet et underafsnit, der forklarer, hvordan interpoleringen er lavet i R-Studio.

Til interpolering af data er *Cubic spline* valgt, da det er en simpel metode til at udlede de manglende datapunkter mellem allerede eksisterende observerede datapunkter. De observerede datapunkter fungerer som noder/tenor punkter, hvor et tredjegradspolynomium bliver estimeret mellem to af node punkterne. Hvert enkelt polynomium har sine egne koefficienter og er konstrueret således, at begyndelsen og slutningen af polynomiet rører noderne[18]. Cubic spline giver derfor mange "*sub-funktioner* ", hvor hvert enkelt af sub funktionerne er med til at generere de manglende datapunkter, som det ikke har været muligt at observere. Cubic spline kan derfor refereres til som en *piecewise* funktion, som har formen:

$$S(x) = \begin{cases} s_1(x) & if \quad x_1 \leq x < x_2 \\ s_2(x) & if \quad x_2 \leq x < x_3 \\ & . \\ & . \\ s_{n-1}(x) & if \quad x_{n-1} \leq x < x_n \end{cases} \qquad (7.1)^{19}$$

Det kan ses i funktionen (7.1), at $S(x)$ består af mange funktioner, der forbinder hvert enkelt node punkt, der er blevet sat til at skulle estimeres. s_i er et tredjegradspolynomium, som er defineret ved:

[18] Cubic Spline Interpolation. Sky McKinley and Megan Levine. Side 1-4
[19] Cubic Spline Interpolation. Sky McKinley and Megan Levine. Side 1-4

$$s_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \qquad (7.2)$$

Hvor $i = 1,2,\ldots, n-1$

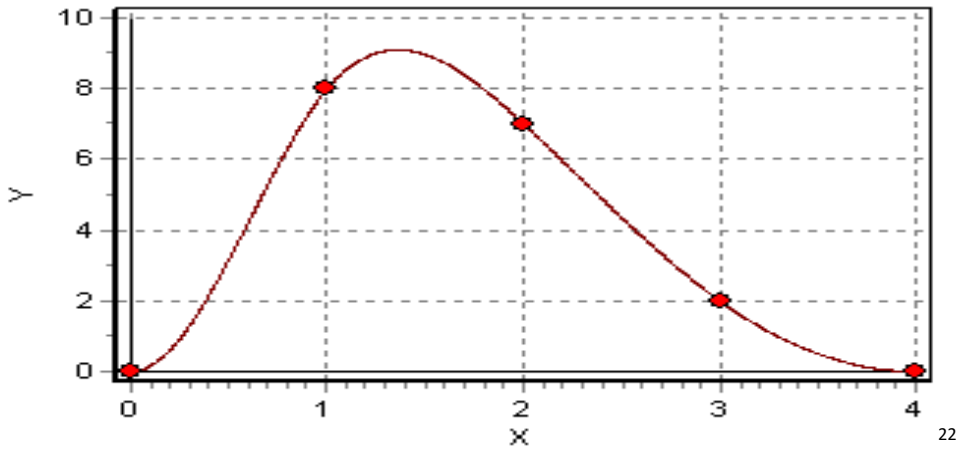
Cubic spline funktionen (7.2) har fire forskellige krav, som den skal overholde[20]:

- **Krav 1:** *Værdien for hver enkelt polynomium er ens ved node punkterne*.
- **Krav 2:** *Første differentiale for hvert polynomium skal være ens ved node punkterne.*
- **Krav 3:** *Anden differentiale for hvert polynomium skal være ens ved node punkterne.*
- **Krav 4:** *Anden differentiale for hvert polynomium er kontinuerligt mellem node punkterne.*

Krav 4 er i sig selv altid opfyldt. Overvejes polynomiet $y = ax^3 + bx^2 + cx + d$, er anden differentialet givet ved $y'' = 6ax + 2b$, hvilket er en lineær funktion og er ved dens egen definition, kontinuerlig mellem node punkterne[21].

I nedenstående figur (7.1) er et eksempel på, hvordan det grafisk ser ud, når Cubic spline anvendes til at forbinde de valgte node punkter.

Figur (7.1): Cubic spline eksempel



[22]

I figur (7.1) er der observerede værdier for variablen X med en numerisk forskel på 1. De observerede værdier er markeret med en rød prik. Det kan ses i illustrationen, at værdien ikke kendes mellem punkterne X, men det kan Cubic spline interpoleringen hjælpe med.

[20] *Fitting the term structure of interest rates,* Pienaar , Rod og Moorad Choudhry, *side 4-8*
[21] *Fitting the term structure of interest rates,* Pienaar , Rod og Moorad Choudhry *side 4-8*
[22] Interpolation Methods. Orcaflex.

Den forbinder de observerede punkter ved hjælp af et tredje grads polynomium, kan ses i ligning (7.2). Når polynomiet er estimeret kan værdierne mellem f.eks. X til tiden 1 og 2 trækkes ud. Hvis eksemplet fortsættes kan **krav 1** tjekkes op på.

Fra ligning (7.1) ved vi, at S(x) interpolerer alle data noderne, og det kan derfor konkluderes at:

$$S(x_i) = y_i \qquad (7.3)$$

Det kan ses i figur (7.1) at den første node $x_0 = 0$. Indsættes denne værdi i ligning (7.2), og notationen ændres til y_i fra ligning (7.3), derved:

$$y_i = a_i(0)^3 + b_i(0)^2 + c_i(0) + d_i$$

$$y_i = d_i$$

Det kan ses, at hvis y_i er det samme som d_i, vil polynomiet rører noden x_0. For at polynomiet også rører det andet node punkt i x_{i+1}, tager polynomiet følgende form:

$$y_{i+1} = a_i(x_{i+1} - x_i)^3 + b_i(x_{i+1} - x_i)^2 + c_i(x_{i+1} - x_i) + d_i$$

Dette kan skrives som:

$$d_{i+1} = a_i X_i^{\,3} + b_i X_i^{\,2} + c_i X_i + d_i \qquad (7.4)^{23}$$

Hvor $X_i = x_{i+1} - x_i$

Krav 2 for cubic spline kræver, at første differentiale skal være ens ved node punkterne, hvilket vil sige, y_i' skal give det samme som første differentiel af y_{i+1}' ved node punktet x_{i+1}.

Forholdet ved x_{i+1} kan skrives som:

$$3a_i X_i^{\,2} + 2b_i X_i + c_i = 3a_{i+1} X_{i+1}^{\,2} + 2b_{i+1} X_{i+1} + c_{i+1}$$

Fra gennemgangen af krav 1 ved vi at $X = 0$, som også er starten for det næste polynomium. Nul kan nu indsættes i X_{i+1} : $3a_{i+1}0^2 + 2b_{i+1}0 + c_{i+1} = 3a_i X_i^{\,2} + 2b_i X_i + c_i$

Hvilket giver: $c_{i+1} = 3a_i X_i^{\,2} + 2b_i X_i + c_i \qquad (7.5)^{24}$

Krav 3 er opfyldt for cubic spline, hvis anden differentiale for y_i'' vurderet ved node x_{i+1} er det samme som anden differentiale af y_{i+1}''.

Sammenlignes de to funktioner:

[23] *Fitting the term structure of interest rates,* Pienaar , Rod og Moorad Choudhry, *side 4-8*
[24] *Fitting the term structure of interest rates,* Pienaar , Rod og Moorad Choudhry, *side 4-8*

$$6a_iX_i + 2b_i = 6a_{i+1}X_{i+1} + 2b_{i+1}$$

Vi ved igen at $X = 0$ fra krav 1, og værdien indsættes igen på X_{i+1} plads:

$$6a_iX_i + 2b_i = 6a_{i+1}0 + 2b_{i+1}$$
$$6a_iX_i + 2b_i = 2b_{i+1}$$
$$6a_iX_i = 2b_{i+1} - 2b_i$$

a_i isoleres nu så følgende ligning opnås:

$$a_i = \frac{b_{i+1} - b_i}{3X_i} \qquad (7.6)^{25}$$

Interpolering R-Studio

Teorien, der er blevet gennemgået i forrige afsnit, vil blive implementeret i R-Studio, hvor en funktion bliver anvendt til at lave interpoleringen for det data, hvor der er manglende datapunkter.

I R-Studio benyttes en R-pakke ved navn "imputeTS". Pakken gør det muligt at anvende funktionen *na.interpolation*. Der er forskellige interpoleringsmetoder inkluderet i denne funktion, her er interpolerings metode *cubic spline* valgt, da den er blevet vurderet til at være den mest optimale. Interpoleringen i R-Studio fungerer ved, at det observerede datasæt indlæses. Her bliver der inkluderet *NaN* observationer mellem de faktiske observerede værdier. Funktionen *na.interpolation* fylder disse *NaN* observationer ud med estimerede værdier, hvor *cubic spline* bliver anvendt til at estimere disse værdier. Det nye interpolerede datasæt er gennemgået under afsnittet for data beskrivelse.

8 Dickey Fuller test

Inden det observerede data kan anvendes til model arbejde, skal det først gennemgå en diagnostisk analyse, hvor der testes for om dataet er stationært. Betegnelsen stationaritet anvendes til at beskrive en tidsserie, som ikke har en *unit root*. Testen, der bliver anvendt i projektet, er en ***augmented*** *Dickey-Fuller* test, men først vil der komme en gennemgang af en ***standard*** *Dickey-Fuller* test, hvorefter den omskrives til en *augmented Dickey-Fuller* test.

Dickey-Fullers første antagelse består af, at data er genereret af nedenstående ligning:

$$y_t = a_1 y_{t-1} + \varepsilon_t \qquad (8.1)$$

[25] *Fitting the term structure of interest rates,* Pienaar , Rod og Moorad Choudhry, *side 4-8*

Hvor en subtraktion af y_{t-1} på hver side af ligning (8.1) fortages. Efter subtraktion fås den tilsvarende ligning $\Delta y_t = \gamma y_{t-1} + \varepsilon_t$ hvor $\gamma = a_1 - 1$. I ligning (8.1) var hypotesen, at der var *unit root*, hvis $a_1 = 1$, hvilket vil svarer til at teste hypotesen $\gamma = 0$ [26]. Dickey og Fuller konstruerer tre ligninger, der kan blive brugt til at teste, om der er *unit root*. De tre regressionsligninger er angivet i det nedenstående, og anvendes til at undersøge tilstedeværelsen af de deterministisk led c og δt. c er en konstant, som også bliver refereret til som "*Drift term*". δt er en lineær trend, der er afhængig af tiden t, og δt bliver refereret til som et lineært tidstrend element.

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t \tag{8.2}$$

Den første regression Dickey og Fuller anvender, er angivet i ligning (8.2). Hvis *unit root* er tilstede i ligning (8.2) er det en ren *random walk* model.

$$\Delta y_t = c + \gamma y_{t-1} + \varepsilon_t \tag{8.3}$$

Regressionen i ligning (8.3) er tillagt et *drift term* c.

$$\Delta y_t = c + \delta t + \gamma y_{t-1} + \varepsilon_t \tag{8.4}$$

Den sidste ligning Dickey og Fuller opsætter, er angivet i ligning (8.4). Regressionen foresætter, at data er genereret ved hjælp af et *drift term* c og et lineær trend led δt [27].

Ved Dickey-Fuller testen estimeres en eller flere af ligningerne, hvorefter den tilhørende t-værdi for γ findes. Hvis den t-værdien ikke er signifikant kan H_0-hypotesen $\gamma = 0$ ikke forkastes, hvilket betyder der er *unit root* i dataserien $\{y_t\}$. Der skal dog tages højde for, at de tre ligninger angivet ovenover, alle forkaster H_0-hypotesen ved forskellige kritiske værdier, så alt efter hvilken ligning der bliver anvendt, skal den tilhørende kritiske værdi for denne ligning benyttes.

Dickey-Fuller testene omskrives nu til ***augmented*** *Dickey-Fuller* tests. Dette gøres meget simpelt ved at introducere lags til det de tre tidligere omtalte ligninger.

AR

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \tag{8.5}$$

[26] MODELS WITH TREND. Enders, Walter. Side 181-255
[27] MODELS WITH TREND. Enders, Walter. Side 181-255

ARD $$\Delta y_t = c + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \qquad (8.6)$$

TS $$\Delta y_t = c + \delta t + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \qquad (8.7)$$

I forhold til de kritiske værdier, der anvendes til at afgøre om, der eksisterer en *unit root*, forbliver stadig de samme.

Dickey og Fuller introducerede i 1981 tre ekstra F-statistiker, der kan bruges til at teste, hvorvidt der skal være drift og/eller en lineær tids trend i den *augmented Dickey-Fuller* test. F-statistikkerne kaldes for ϕ_1, ϕ_2 og ϕ_3 [28].

Dickey og Fuller starter med at teste ϕ_3, hvor det antages, at H_0-hypotesen er givet ved ligning (8.6), og alternativ hypotesen er givet ved ligning (8.7), hvilket vil sige at $\gamma = \delta = 0$.

ϕ_3 anvendes til at teste joint H_0-hypotesen: $\gamma = \delta = 0$

Hvis F-værdien for ϕ_3 er lavere end den kritiske værdi, kan H_0-hypotesen ikke forkastes. Hvis dette er tilfældet, antages det, at data er genereret af ligning (8.6). ϕ_1 kan herefter testes, for at undersøge om der skal være en konstant med i modellen for dataet.

ϕ_2 undersøger om $\gamma = c = \delta = 0$. Her er H_0-hypotesen, at data er genereret af ligning (8.5), mens alternativ hypotesen er, at data er genereret af ligning (8.7). Hvis F-værdien for ϕ_2 er højere end den kritiske værdi for ϕ_2, forkastes H_0-hypotesen og alternativ hypotesen accepteres. Dette vil sige, at data er genereret af ligning (8.7), da mindst et af elementerne $\gamma = c = \delta \neq 0$.

Det kan herefter være interessant estimere ϕ_3, således at der kan testes for, om tidstrenden δ kan undlades, så ligningen kun har drift c med.

ϕ_2 tester joint H_0-hypotesen at $\gamma = c = \delta = 0$

ϕ_1 anvendes af Dickey og Fuller til at teste om hvorvidt data er genereret af ligning (8.5) eller (8.6). H_0-hypotesen antager at $\gamma = c = 0$, hvilket vil sige, at hvis F-værdien er lavere end den kritiske værdi for ϕ_1, kan H_0-hypotesen ikke forkastes. Data kan derfor antages at være genereret af ligningen (8.5).

[28] MODELS WITH TREND. Enders, Walter. Side 181-255

ϕ_1 anvendes til at teste joint H_0-hypotesen[29]: $\gamma = c = 0$

ϕ_1, ϕ_2 og ϕ_3 er alle konstrueret således, at en almindelig F-test kan anvendes, hvilket er angivet i ligning (8.8):

$$\phi_i = \frac{[SSR(restricted) - SSR(unrestricted)]/r}{SSR(unrestricted)/(T-k)} \qquad (8.8)$$

SSR(restricted) = squared residuals for den begrænset model.

SSR(unrestricted) = Sum of squared residuals for den ubegrænset model.

r = Antal af restriktioner.

T = Antal af brugbare observationer.

k = Antal parameter estimeret.

SSR er den estimerede værdis afvigelse fra den observerede værdi. *SSR* gør det muligt at undersøge, hvor stor uoverensstemmelse der er mellem empiriske observationer og den valgte models estimering. Små *SSR* værdier indikerer, at der er et godt *fit* mellem data og model. Hvis *SSR* værdierne for den begrænsede model og den ubegrænsede model er meget tæt, er det ensbetydende med at restriktionerne ikke er bindende. Hvordan *sum of squared residuals* udregnes, kan ses i nedenstående ligning (8.9), hvor y_i angiver den observerede værdi, og $\hat{y}_i$ er den estimerede værdi fra den pågældende model.

SSR $$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (8.9)$$

De kritiske værdier for ϕ_i, som er udregnet af Dickey og Fuller, kan ses i de understående tabeller.

Tabel (8.1)[30]: Kritiske værdier for ϕ_1

Signifikant level	0.10	0.05	0.025	0.01
Observationer T		ϕ_1		
25	4.12	5.18	6.30	7.88
50	3.94	4.86	5.80	7.06
100	3.86	4.71	5.57	6.70
250	3.81	4.63	5.45	6.52

[29] MODELS WITH TREND. Enders, Walter. Side 181-255
[30] SUPPLEMENTARY MANUAL TO ACCOMPANY. Enders, Walter. Side 120

500	3.79	4.61	5.41	6.47
∞	3.78	5.59	5.38	6.43

Tabel (8.2)[31]: Kritiske værdier for ϕ_2

Signifikant level	0.10	0.05	0.025	0.01
Observationer T		ϕ_2		
25	4.67	5.68	6.75	8.21
50	4.31	5.13	5.94	7.02
100	4.16	4.88	5.59	6.50
250	4.07	4.75	5.40	6.22
500	4.05	4.71	5.35	6.15
∞	4.03	4.68	5.31	6.09

Tabel (8.3)[32]: Kritiske værdier for ϕ_3

Signifikant level	0.10	0.05	0.025	0.01
Observationer T		ϕ_3		
25	5.91	7.24	8.65	10.61
50	5.61	6.73	7.81	9.31
100	5.47	6.49	7.44	8.73
250	5.39	6.34	7.25	8.43
500	5.36	6.30	7.20	8.34
∞	5.34	6.25	7.16	8.27

De udregnede ϕ_i'er sammenlignes med de udregnede Dickey-Fuller kritiske værdier for de pågældende ϕ_i-værdier, som kan ses i de ovenstående tabeller. De kritiske værdier gør det muligt at bestemme, hvilke restriktioner der gør sig gældende. H_0-hypotesen for ϕ_i er som nævnt tidligere, at data er genereret ved det begrænsede model, mens alternativ hypotesen er, at data er genereret af den ubegrænsede model.

[31] SUPPLEMENTARY MANUAL TO ACCOMPANY. Enders, Walter. Side 120
[32] SUPPLEMENTARY MANUAL TO ACCOMPANY. Enders, Walter. Side 120

Små værdier for ϕ_i betyder derfor, at H_0 accepteres og restriktionerne ikke er bindende, mens høje værdier for ϕ_i gør, at H_0-hypotesen kan forkastes og restriktionerne er bindende.

Modellen, der er blevet udvalgt på baggrund af de estimerede ϕ_i værdier, kræver også en bestemt lag-længde. Det er vigtigt at finde det rigtige antal lags for den *augmented Dickey-fuller* test. Hvis der bruges for få lags, vil residualerne fra regressionen af den udvalgte model ikke bevæge sig som en hvid støj. Samtidig vil modellen ikke være i stand til at fange den faktiske "*error*" proces, derfor vil estimationen af γ og den tilhørende standardfejl være dårlig.
Inkluderes der for mange lags i *augmented Dickey-fuller* test, bliver testens evne til at forkaste H_0-hypotesen af en eksisterende *unit root* forringet. Forringelsen sker på baggrund af de ekstra parametre, der skal estimeres, hvilket fører til tab af frihedsgrader. Lag-længden bliver meget ordinært fundet ved hjælp af ***general to specific*** metoden. Initial lag-længden vælges fra start til at være en relativ lang lag-længde for den udvalgte model, hvorefter der sker en reducering af lag-længden. Efter hver reducering tjekkes signifikant niveauet for den sidste parameter. Denne proces fortsættes indtil det sidste lag er signifikant. Når den optimale lag-længde er fundet, undersøges residualerne om de er uafhængig fordelt. Korrelationen kan tjekkes ved at plotte residualerne og/- eller ved brug af en *Ljung-Box* test[33].

Efter den optimale lag-længde er fundet for den udvalgte model, kan data tjekkes for stationaritet ved brug af *augmented Dickey-Fuller* test. Matlab anvendes til at udøve testen. Her undersøges det om H_0-hypotesen $\gamma = 0$ holder. I Matlab returneres enten et 0 eller et 1 tal, hvis facit er 0, er der *unit root* og data er derfor ikke stationær. Den underlæggende estimation af hvorvidt $\gamma = 0$ gøres ved hjælp af en standard t-test:

$$t = \frac{(a_1 - 1)}{\sigma'}$$

Augmented Dickey -Fuller tests

Den gennemgåede teori for *augmented Dickey-Fuller* test vil nu blive anvendt til at analysere, hvorvidt dataet, der anvendes i projektet, er stationært. Det vil først blive undersøgt, hvilken af de tre modeller der bedst forklarer det observerede data. Det observerede data er blevet gennemgået i afsnittet med databeskrivelsen, og *augmented Dickey-Fuller* tests vil blive konstrueret i Matlab. Til slut i afsnittet vil en overordnet konklusion på, om det observerede data er stationært.

[33] MODELS WITH TREND. Enders, Walter. Side 181-255

Huspriser

Det observerede data beskriver udviklingen i huspriserne i hver enkelt af de danske regioner samt hele Danmark. Datasættet består 95 enkelte observationer, hvor differencen og log er taget af data, for hver region og hele Danmark.

Ligningerne (8.5), (8.6) og (8.7) bliver anvendt i Matlab, for at kunne køre en regression på det observerede *diff(log(data))* for hver af de enkelte regioner samt hele Danmark. Regressionen anvender kun et lag for alle modellerne til at starte med, og efter den optimale model for dataet er fundet, vil det blive analyseret hvilken lag-længde, der skal anvendes.

Fra de tre *AR, ARD* og *TS* regressioner udledes *sum of squared residuals* for de enkelte modeller således, at ϕ_i værdierne kan findes. *SSR*-værdierne for huspriserne kan ses i nedenstående tabel (8.4):

Tabel (8.4): *SSR*-værdierne for huspriserne

			SSR			
Reg	**Hele**	**Hoved**	**Sealand**	**Syd**	**Mid**	**Nord**
TS	0,004607	0,016481	0,008384	0,002674	0,007921	0,008027
AR	0,009849	0,023859	0,011012	0,004409	0,012485	0,010677
ARD	0,002343	0,014475	0,004899	1,18E-05	0,00599	0,006102

Som det fremgår af teorien, kan de forskellige *SSR* allerede nu anvendes til at analysere, hvorvidt restriktionerne er bindende eller ej. Hvis der kigges nærmere på regressionen *TS* for hele Danmark, giver *SSR* os værdien 0,004607 og for *AR* værdien 0,009849. Værdierne er ikke tæt på at være de samme, derfor kan det udledes at AR modellen angivet i ligning (8.5) ikke er den optimale at bruge i *augmented Dickey-Fuller* test. ϕ_i værdierne udregnes nu på baggrund af *SSR*, således den optimale model til anvendelse af *augmented Dickey-Fuller* test kan findes. ϕ_i -værdierne for de observerede huspriser kan ses i tabel (8.5)

Tabel(8.5): ϕ_i -værdierne for de observerede huspriser

			ϕ(i)			
Phi	**Hele**	**Hoved**	**Sealand**	**Syd**	**Mid**	**Nord**
ϕ1	145,7431	29,50079	56,774	16988,53	49,3313	34,11801
ϕ2	34,51859	13,57919	9,507491	19,68125	17,47854	10,01439
ϕ3	-22,6016	-5,60091	-19,1211	-4,58E+01	-11,2117	-11,0331

Den første værdi, der er interessant at se på, er ϕ_2 for hele Danmark. Hvis vi sammenligner F-værdien $\phi_2 = 34{,}51859$ med den kritiske værdi 4,88 fra tabel (8.2) ved ~ 100 observationer, kan de ses at H_0-hypotesen bliver forkastet og alternativ hypotesen accepteres. H_0-hypotesen er, at data er genereret af den begrænsede model, mens alternativ hypotesen er, at data er genereret af den ubegrænsede. Den begrænsede model til ϕ_2 er ligning (8.5), og den ubegrænsede model er ligning (8.7). At alternativ hypotesen accepteres betyder, at en af elementerne $\gamma = c = \delta$ er forskellige fra 0. Alle ϕ_2 værdierne for de seks regressioner i tabel (8.5) er højere end den kritiske værdi fra tabel (8.2), hvilket som sagt betyder, at alternativ hypotesen accepteres for dem alle. For nu antager vi at alt data er genereret af ligning (8.7). Værdien for ϕ_3 udregnes nu for at teste om $\delta = \gamma = 0$. ϕ_3 for hele Danmark er lavere end den kritiske værdi 6,49 for 5 procents signifikant niveau angivet i tabel (8.3). H_0-hypotesen kan derfor ikke forkastes, og derfor kan antagelsen om, at data for hele Danmark er genereret af ligning (8.7), ændres til at data er genereret af ligning (8.6). At denne antagelse er sand, kan også ses på ϕ_1 værdien, da værdien er højere end den kritiske værdi på 4,71. Alternativ hypotesen accepteres derfor, hvilket vil sige at data er genereret af den ubegrænsede model, der er angivet i ligning (8.6). Hvis vi analyser på ϕ_i for alle regionerne, kan det ses, at ϕ_3 ikke er signifikant for nogle af dem, hvilket betyder at H_0-hypotesen accepteres, og data er derfor genereret af *ARD*-modellen i ligning (8.6). For at lave et sidste tjek, er ϕ_1 værdierne også udregnet for alle regionerne. Det viser sig, at de alle er signifikante ved et 5 procents signifikant niveau, da alle ϕ_1 værdierne i tabel (8.5) er over den kritiske værdi på 4,71. Det kan konkluderes, at det observerede data for regionerne og hele Danmark er genereret ved ARD modellen angivet i ligning (8.6).

Den optimale model for at teste om dataet for huspriserne er stationært er nu valgt. Inden testen dog kan køres, skal det undersøges hvor mange lag $p*$ *ARD*-modellen skal arbejde med.
Lag-længden for modellen findes ved at opstille et loop i Matlab for *augmented Dickey-Fuller* test ved brug af **general to specific** metodologien. Ligning (8.6) starter med at blive estimeret med 6 lags, herefter udtrækkes p-værdien. Hvis p-værdien for parameteren ved det 6. lag er signifikant, betyder det at *ADF* testen skal anvende 6 lags. Er dette ikke tilfældet bliver modellen genestimeret med lag-længden $p*-1$. Hele processen bliver fortaget i Matlab indtil det sidste lag element er signifikant forskelligt fra nul.

Tabel (8.6): Huspris parametres p-værdi

p-værdi	Hele	Hoved	Sealand	Syd	Mid	Nord
c	0,01095	0,057499	0,069653	0,089048	0,01232	0,006861
ɣ	0,478219	0,004685	0,123009	0,067218	0,524999	0,269101
B(t-1)	0,652692	0,001141	0,425281	2,45E-02	0,940529	0,778257
B(t-2)	0,730437		0,05983	0,00455	0,066773	0,476455
B(t-3)	0,289022		0,016224	6,28E-05	0,005282	0,028328
B(t-4)	0,017533					
B(t-5)						
B(t-6)						

I tabel (8.6) er p-værdierne for den optimale lag-længde angivet. Det kan ses, at for hele Danmark er de to sidste lags 5 og 6 ikke medtaget, hvilket betyder, de ikke har været signifikante, og det er først for ved det 4. lag at p-værdien er blevet signifikant. For regionen Hovedstaden er den optimale lag-længde ét lag, og for de resterende regioner er den optimale lag-længde for *ADF* testen 3 lags.
De 6 modeller, som er blevet estimeret, er blevet testet for korrelation i residualerne. Testen er blevet foretaget ved hjælp af plots af residualerne, samt Ljung-Box tests. Begge testene konkluderer, at der ikke er korrelation i residualerne.
Når den passende lag-længde for *ARD*-modellen, der anvendes i *ADF* testen, er fundet kan data nu blive testet for stationaritet. *Augmented Dickey-Fuller* testen for huspriserne med tilhørende t-værdier og kritiske værdier kan findes til sidst i afsnittet i tabel (8.13).

Indkomst

Indkomsten er observeret for hele Danmark samt regionerne. Dataet indeholder 94 observationer, hvor differencen er taget to gange af logaritmen af dataet.
Indkomst viste sig i første *augmented Dickey-Fuller* test ikke at være stationært, og det var derfor nødvendigt at tage difference én ekstra gang. Den første *augmented Dickey-Fuller* test som blev lavet på enkelt differencen af logaritme af indkomst dataet, vil derfor ikke blive gennemgået.

Selv om differencen er taget to gange, er proceduren stadig den samme. *SSR* for de forskellige regressioner er ikke taget med, men de kan ses i Matlab, hvis det interesserer læser.

Tabel (8.7): ϕ_i -værdierne for de observerede huspriser

ϕ(i)

Phi	Hele	Hoved	Sealand	Syd	Mid	Nord
ϕ1	0,014351	0,014012	0,002512	0,001172	0,023433	0,004828
ϕ2	0,007977	0,008701	0,00161	-0,00033	0,010457	0,002169
ϕ3	-0,00223	-0,00081	-6,95E-05	-1,66E-03	-0,00749	-0,00152

Tabel (8.7) angiver ϕ_i -værdierne, og det kan ses, at for hele Danmark og alle regionerne er værdierne meget lave. $\phi_2 = 0{,}007977$. Holdes denne værdi op mod Dickey og Fullers kritiske værdi 4,88 i tabel (8.2) for 5 procents signifikantniveau, kan H_0-hypotesen at $\gamma = c = \delta = 0$ ikke forkastes. Selv ved 10 procents signifikantniveau, hvor den kritiske værdi er 4,16 kan H_0-hypotesen ikke forkastes. Dette er også tilfældet ved alle regionerne. Det kan derfor allerede nu antages, at data er genereret ved AR-regressionen angivet i ligning (8.5), da H_0-hypotesen ikke kan forkastes, er data genereret af den begrænsede model.

Efter den optimale model for indkomst dataet er fundet, skal den optimale lag-længde findes. Dette gøres igen ved brug af **general to specific** metodologien. Modellen der startes ud med, er en AR model med 15 lags. Herefter justeres lag-længden indtil p-værdien for den sidste laggede parameter er signifikant. De optimale lag-længder er angivet i nedenstående tabel (8.8).

Tabel (8.8): Indkomst parametres p-værdi

p-værdi	Hele	Hoved	Sealand	Syd	Mid	Nord
ɣ	2,20E-08	2,87E-07	2,25E-07	4,74E-07	5,28E-10	7,40E-14
B(t-1)	1,02E-19	9,23E-19	5,46E-19	4,63E-19	1,37E-19	2,10E-20
B(t-2)	0,849412	0,573615	0,862233	0,922801	0,470522	0,693151
B(t-3)	0,006738	0,006162	0,005686	0,004511	0,01581	0,016699
B(t-4)	3,32E-09	7,54E-08	2,20E-09	4,29E-10	4,04E-08	1,28E-08
B(t-5)	3,51E-14	1,69E-12	6,00E-14	1,02E-14	8,96E-14	4,15E-12
B(t-6)	0,313038	0,168841	0,277481	0,365895	0,189764	0,445247
B(t-7)	0,108025	0,092074	0,093432	0,089943	0,129634	0,261583
B(t-8)	1,84E-06	7,43E-05	1,89E-06	3,25E-07	1,68E-06	0,000268
B(t-9)	1,10E-10	2,08E-08	2,92E-10	3,34E-11	3,76E-11	8,84E-07

B(t-10)	0,211285	0,138871	0,183397	0,234263	0,086033	
B(t-11)	0,329015	0,291643	0,298588	0,308432	0,322072	
B(t-12)	0,003919	0,035334	0,004979	0,001318	0,022708	
B(t-13)	9,66E-06	0,000818	4,01E-05	3,47E-06	0,000172	

De fleste modeller for dataet for hele Danmark og regionerne er signifikant ved det 13. lag, dog med undtagelse af indkomst dataet for Region Nordjylland. Det kan ses at AR modellen for Region Nordjylland, når der anvendes i *ADF* test, skal bruge 9 lags. De optimale lag-længder for indkomst data for hele Danmark og regionerne er angivet med grønt i tabel (8.8). De seks estimerede modeller er blevet undersøgt for korrelation i residualerne, og det kan konkluderes ud fra plot og Ljung-Box test, at der ikke er korrelation i residualerne.
Nu kan de fundene lag-længder anvendes i den *augmented Dickey-Fuller* test, som tidligere sagt laves i Matlab. Selve den *augmented Dickey-Fuller* test for stationaritet vil blive gennemgået til sidst i afsnittet og kan ses i tabel (8.13).

Skat

Data for variablen skat for regionerne og hele Danmark består af 94 observationer, hvor differencen er taget to gange af logaritmen til dataet. Igen har det været nødvendigt, ligesom ved indkomst dataet, at tage differencen to gange, fordi den første *augmented Dickey-fuller* test viste, at dataet ikke var stationært.

Tabel (8.9): ϕ_i -værdierne for variablen skat

ϕ(i)

Phi	Hele	Hoved	Sealand	Syd	Mid	Nord
ϕ1	-0,01362	-0,0153	-0,00337	0,010574	-0,01606	-0,00623
ϕ2	-0,01328	-0,01452	-0,00272	0,00541	-0,02532	-0,01209
ϕ3	-0,00646	-0,00664	-0,00074	-2,34E-03	-0,02211	-0,01198

Tabel (8.9) angiver alle ϕ_i -værdierne for F-testen i ligning (8.8). For af nemmest at kunne afklare hvilken model, der vil være den optimale for *augmented Dickey-Fuller* test, ses der først på ϕ_2. Alle ϕ_2 værdierne er meget lave, og sammenholdes de med den kritiske værdi for ϕ_2, der er angivet i tabel (8.2) 4,88, kan H_0 -hypotesen ikke forkastes. Det antages, at data for regionerne og hele Danmark, er genereret af den begrænsede AR model i ligning (8.5). Den optimale lag-længde findes nu for AR modellen.

General to specific metodologien anvendes igen, hvor start lag-længden for modellen er sat til 15 lags. P-værdierne for parametrene er angivet i understående tabel.

Tabel (8.10): Skat parametres p-værdi

p-værdi	Hele	Hoved	Sealand	Syd	Mid	Nord
ɣ	5,60E-17	3,81E-17	4,68E-07	2,89E-13	1,27E-18	1,66E-12
B(t-1)	1,66E-20	6,91E-21	6,28E-18	3,77E-17	1,36E-19	1,78E-18
B(t-2)	0,478993	0,761773	0,702667	5,25E-02	0,222661	0,28251
B(t-3)	0,137513	0,113039	0,02405	0,111932	0,285473	0,057363
B(t-4)	8,22E-10	1,21E-10	8,57E-11	4,22E-11	5,08E-09	7,89E-11
B(t-5)	1,46E-13	5,78E-14	4,94E-13	1,94E-12	1,80E-12	5,84E-12
B(t-6)	0,176959	0,355349	0,151784	0,008772	0,080508	0,095598
B(t-7)	0,638773	0,626288	0,317845	0,34025	0,809286	0,408827
B(t-8)	0,013249	0,000603	0,000226	0,001291	0,162509	0,000824
B(t-9)	0,000149	1,11E-06	2,74E-07	7,79E-06	0,018621	3,62E-06
B(t-10)			0,201162	0,010829		0,225177
B(t-11)			0,618754	0,542852		0,757225
B(t-12)			0,079246	0,166571		0,156317
B(t-13)			0,007438	0,026461		0,020307
B(t-14)				0,049783		

Det kan ses i tabel (8.10), at dataet for skat for hele Danmark og regionerne har meget forskellige lag-længder. Dette sker pga. den første initial lag-længde blev sat til 10, men det viste sig, at selvom lag for modellen var blevet reduceret indtil det sidste lag var signifikant, var der stadig korrelation i residualerne. Det var derfor nødvendigt at starte med en længere lag-længde, hvilket denne gang blev sat til 15. De seks forskellige modeller er igen er blevet testet ved hjælp af Ljung-Box test og plot af residualerne, og det har vist at der ikke længere er korrelation i residualerne. Facit for stationaritet testene vil blive gennemgået til sidst i afsnittet i tabel (8.13).

Renten

Dataet for renten består af 95 observationer, hvor differencen er taget af logaritmen af data. Renten er kun observeret for hele Danmark.

For at kontrollere at data for renten er stationært, skal det testes for hvilken model, det kan antages at data er genereret af. Dette gøres ved at se på de nedenstående ϕ_i -værdier i tabel (8.11).

Tabel (8.11): ϕ_i -værdierne for variablen renten

ϕ(i)

Phi	Hele
ϕ1	11,5158
ϕ2	7,393244
ϕ3	-0,24055

Først testes det om hvorvidt ϕ_2 er signifikant. Ved test af ϕ_2 er det antaget, at ligning (8.5) er den begrænsede og ligning (8.7) er den ubegrænsede. Her er $\phi_2 = 7{,}393244$, hvilket er højere end den kritiske værdi 4,88 angivet i tabel (8.2). H_0-hypotesen forkastes, og alternativ hypotesen, at data er genereret ved hjælp af den ubegrænsede model i ligning (8.7), accepteres. Inden det antages at dette er korrekt, testes ϕ_3 for at undersøge om der skal være en tidstrend med i modellen. Ligning (8.6) antages at være den begrænsede model, og ligning (8.7) er den ubegrænsede. Hvis H_0-hypotesen forkastes, skal der være en tidstrend med i modellen. Her er $\phi_3 = -0{,}24055$, hvilket er under den kritiske værdi 6,49 for ϕ_3 angivet i tabel (8.3). H_0-hypotesen kan derfor ikke forkastes, og data er derfor genereret af den begrænsede model i ligning (8.6). Den optimale *ARD* model undersøges nu for hvor mange lags, der skal være med i *augmented Dickey-Fuller* test.

Tabel (8.12): Rente parametres p-værdi

p-værdi	Hele
c	0,066404
ɣ	0,20127
b1	0,017376

Det kan ses i tabel (8.12), at det første lag, hvor parameteren blev signifikant, er ved ét lag. Den optimale lag-længde for *ARD* modellen for *ADF*-testen blev fundet ved hjælp af **General to specific** metodologien. Initial lag-længden for ARD modellen er blevet sat til 6 lags i Matlab.
Det sidste der mangler, er at teste, hvorvidt data for renten er stationært. Den *augmented Dickey-Fuller*

test køres i Matlab, og det testes her om γ er signifikant, hvilket betyder, at der ikke er *unit root* eksisterende og data derfor er stationært. Testen for stationaritet kan ses i nedenstående tabel (8.13).

Stationaritet

Den optimale model samt de tilhørende optimale lag-længder er blevet fundet i det ovenstående. Data for huspriserne, indkomst, skat og renten vil nu blive testet for eksistensen af *unit root*. Den *augmented Dickey-fuller* testet er blevet udregnet i Matlab. Her er t-værdierne for parameteren γ blevet udregnet for alle modellerne, samt den tilhørende kritiske værdi. I tabel (8.13) er hver enkelt t-værdi for γ samt tilhørende kritiske værdi for huspriserne, indkomst, skat og renten opdelt efter regioner samt hele Danmark blevet illustreret. De kritiske værdier er for 5 procents signifikantniveau.

Tabel (8.13): γ T- og kritisk værdi for alle de udvalgte modeller.

ADF	**Hele**	**Hoved**	**Sealand**	**Syd**	**Mid**	**Nord**
Huspriser	γ	γ	γ	γ	γ	γ
C-value	-2,8939	-2,89264	-2,89348	-2,89348	-2,89348	-2,89348
T-value	-4,215	-4,69278	-3,7053	-3,2069	-4,1848	-4,86966
Indkomst	γ	γ	γ	γ	γ	γ
C-value	-1,94456	-1,94456	-1,94456	-1,94E+00	-1,94446	-1,94453
T-value	-4,88587	-4,49545	-4,68005	-4,81E+00	-10,7173	-5,1828
Skat	γ	γ	γ	γ	γ	γ
C-value	-1,94453	-1,94453	-1,94456	-1,94E+00	-1,94453	-1,94456
T-value	-5,2142	-5,58059	-4,13787	-2,29E+00	-4,42249	-3,96144
Rente	γ					
C-value	-2,89264					
T-value	-7,08532					

H_0-hypotesen for *augmented Dickey-Fuller* test for *unit root* er at $\gamma = 0$. For at kunne forkaste H_0-hypotesen og acceptere alternativ hypotesen, skal t-værdien være lavere end den kritiske værdi. De kritiske værdier er baseret på hvor mange observationer, der er tilgængelige, samt hvilken model der er blevet estimeret. Det kan ses, at den kritiske værdi for AR modellen er højere end for ARD modellerne. Den første *ADF* test, som er blevet fortaget, er for huspriserne for hele Danmark. Her er t-værdien udregnet til -4,215, hvor den kritiske værdi, som kan ses i kolonnen *C-value*, er angivet til at være -2,8939. T-værdien er lavere end den kritiske værdi, så derfor kan H_0-hypotesen om tilstedeværelsen af *unit root* forkastes. Dataet for

huspriserne for hele Danmark er derfor stationært. Det gør sig også gældende for dataet for regionernes huspriserne, hvilket kan ses i tabel (8.13). Her har alle regionerne en lavere t-værdi end den kritiske. Alternativ hypotesen accepteres, da dataet ikke indeholder *unit root*. Indkomst dataet for regionerne og hele Danmark har den tilhørende kritiske værdi ~1,944, hvor alle t-værdier er lavere end den kritiske værdi. Derfor er indkomst dataet stationært. Indkomst dataet har for Region Syd- og Midtjylland stadig korrelation i residualerne, og selv ved tilføjelsen af flere lags blev der stadig observeret korrelation.
Dataet for skat har nogenlunde den samme kritiske værdi som for indkomst. Det kan ses i tabel (8.13), at de tilhørende t-værdier alle er lavere end den kritiske, så derfor er dataet for skat også stationært, da H_0 hypotesen $\gamma = 0$ forkastes. Den sidste variable renten som der er blevet observeret data for mangler nu. Renten er kun observeret for hele Danmark, med den udregnede t-værdi -7,085. T-værdien for renten er meget lavere end den kritiske værdi, så H_0-hypotesen forkastes, og alternativ hypotesen om, at data ikke indeholder *unit root,* accepteres. Alt det observerede data, der anvendes til analysen, kan konkluderes at være stationært.

9 Autoregressive modeller

Dette afsnit kommer til at handle om AR modeller, da der senere i projektet vil blive dannet AR(1) modeller for hele Danmark og de 5 regioner. AR(1) modellerne bliver anvendt som benchmark i projektet. Data som indeholder en række af tal og tid, betragtes som en tidsserie. Tidsserier har lige store tidsrum imellem observations intervallerne, således der kan omstilles modeller på baggrund af dataet. Tidsserie modellerne benyttes til at analysere data for at fange tendenser og underliggende effekter, som herefter kan bruges til analyser og prognoser. Autoregressive modeller er en af de mest benyttet type modeller til bearbejdning af tidsseriedata. I en AR model forventes det, at den forrige observation, X_{k-1}, har effekt på den observerede værdi på tidspunkt *k*. Denne effekt måles ved parameteren *p.*
Derved kan der opstilles en simpel model for tidsserien:

$$X_k = pX_{k-1} + \varepsilon_k. \qquad k = 0, \pm 1, \pm 2, \dots$$

Hvor *X* er den observerede værdi til tidspunkt *k*, *p* er en forklarende parameter for effekten af den *laggede* observations værdi og ε er et fejlled, for de effekter det laggede led ikke kan forklare. Ofte bliver det antaget at fejlleddene i modellerne er tilfældige og ikke korrelerede, der følger en normalfordeling, hvilket vil sige, at fejlledet har en middelværdi på nul og en konstant varians σ^2.

Ligesom X_k afhænger af den tidligere observations værdi X_{k-1}, afhænger X_{k-1} ligeledes af tidligere observations værdi X_{k-2}. Hvis der tages udgangspunkt i en AR(1), kan løsningen til X_k findes ved rekursiv substitution[34]:

$$X_k = pX_{k-1} + \varepsilon_k$$

$$= p(pX_{k-2} + \varepsilon_{k-1}) + \varepsilon_k$$

$$= p^2X_{k-2} + p\varepsilon_{k-1} + \varepsilon_k$$

$$= p^3X_{k-3} + p^2\varepsilon_{k-2} + p\varepsilon_{k-1} + \varepsilon_k$$

$$\dots$$

$$= p^kX_0 + \cdots + p\varepsilon_{k-1} + \varepsilon_k$$

Her er X0 initial værdien.

Fitte data til en model

For beregne parameteren *p* er det nødvendigt at fitte dataet til en AR(1) model. Dette gøres ved at minimere fejlledet *ε* i forhold til parameteren *p*, og derved få den estimerede værdi så tæt som muligt på den observerede værdi. Her tages den afledte af de summerede kvadrater af fejlledet:

$$\frac{\partial}{\partial p}\sum_{k=2}^{n}(X_k - pX_{k-1})^2 = 2\sum_{k=2}^{n}(X_k - pX_{k-1})\,(-X_{k-1})$$

Da vi er interesserede i at minimere fejlledet *ε,* sættes det afledte til 0.

$$2\sum_{k=2}^{n}(X_k - pX_{k-1})\,(-X_{k-1}) = 0$$

$$\sum_{k=2}^{n}\left(-X_kX_{k-1} + pX_{k-1}^2\right) = 0$$

$$-\sum_{k=2}^{n}(X_kX_{k-1}) + p\sum_{k=2}^{n}pX_{k-1}^2 = 0$$

$$\sum_{k=2}^{n}(X_kX_{k-1}) = p\sum_{k=2}^{n}pX_{k-1}^2$$

[34] *AR(1) TIME SERIES PROCESS,* Horvath, Zsuzsanna og Ryan Johnston side 3-5

Ved at isolere *p* fås udtrykket:

$$\hat{p} = \frac{\sum_{k=2}^{n}(X_k X_{k-1})}{\sum_{k=2}^{n} pX_{k-1}^2}$$

Det næste trin i modelleringen af dataet er at finde fordelingen af residualerne, for at kunne determinere om residualerne følger en normalfordeling, og dermed er hvid støj. For at finde denne fordeling ses der på den oprindelige ligning for en AR(1):

$$X_k = pX_{k-1} + \varepsilon_k$$

Her isoleres fejlleddene således:

$$\hat{\varepsilon}_k = X_k - \hat{p}X_{k-1}$$

Her kan et histogram af de estimerede fejlled være en mulighed, da det forventes at fejlleddene følger en normal fordeling, hvis de er hvid støj. Dette princip vil blive benyttet senere for forecasts residualerne under evalueringen af forecastene.[35]

Brug af AR(1) i projektet

I projektet benyttes AR(1) modeller til at teste kvaliteten af forecastene baseret på VAR modellerne, som vil blive gennemgået i afsnittet med forecast evaluring. AR(1) modellerne dannes på baggrund af ejendomsprisudviklingen i den enkelte region. Her fjernes de sidste 14 observationer af dataet, således at den dannede AR(1) model kan testes *out-of-sample*. Her laves *one step ahead forecasts*, som kan blive sat op imod i de 14 undladte observationer, og dermed kan modellen blive vurderet i forhold til disse observationer. I det følgende afsnit vil forecasts teori blive gennemgået med udgangspunkt i Walter Enders Applied Econometric Time Series[4 edition] , hvilket betyder at notationerne for AR modellen ændres til Walter Enders AR notationer.

10 Forecasting

Objektet i modelarbejde er ikke at modellere en model, der er repræsentativ for det økonomiske observerede data. Derimod er det målet at finde simpel modeller, der vil være i stand til at fange opførslen i en tidsserie, således at modellen har et tilstrækkelig basis for forecasting. I dette afsnit vil egenskaberne af forecasting blive gennemgået. Først bliver der lagt fokus teorien bag forecasting af en AR model, som i projektet bliver sat til at være benchmark. Forecasting teorien af en AR model blive gennemgået ud fra Applied Econometric Time Series[4 edition], skrevet af Walter Enders.

[35] *AR(1) TIME SERIES PROCESS,* Horvath, Zsuzsanna og Ryan Johnston, side 6-8

Enders antager at data genereringsprocessen af tidligere og nuværende realisationer af $\{y_t\}$ og $\{\varepsilon_t\}$ sekvenser er kendt af forskeren. Enders ligger først fokus på en AR model med et lag, noteres som AR(1), i forklaringen af forecasting:

AR(1) $$y_t = a_0 + a_1 y_{t-1} + \varepsilon_t \qquad (10.1)$$

y_t er de observerede værdier for den variable, der ønskes at blive modelleret. a_0 er en konstant, og a_1 er den estimeret parametre værdi for den laggede værdi af y_{t-1}. ε_t er fejlleddet for den estimerede AR(1) model. AR(1) modellen i ligning (10.1) opdateres nu én periode frem, således at ligningen ændres til:

$$y_{t+1} = a_0 + a_1 y_t + \varepsilon_{t+1}$$

Hvis koefficienterne for a_0 og a_1 er kendt, muliggør det at forecaste en periode frem y_{t+1}, hvilket er begrænset af informationen, der er tilgængelig i periode t:

$$E_t[y_{t+1}] = a_0 + a_1 y_t \qquad (10.2)$$

Enders bruger her notationen $E_t[y_{t+j}]$ som en forkortelse for den begrænsede forventning af y_{t+j}, given den information som er tilgængelig i periode t. j angiver hvilken periode, der bliver forecastet. Begrænsningerne, der forelægger forventningen, kan mere formelt skrives som:

$$E_t[y_{t+j}] = E(y_{t+j} \mid y_t, y_{t-1}, y_{t-2}, ..., \varepsilon_t, \varepsilon_{t-1}, ...) .$$

Hvis ligning (10.1) opdateres fra en periode frem til to perioder:

$$y_{t+2} = a_0 + a_1 y_{t+1} + \varepsilon_{t+2}$$

Hvilket betyder, at forecasts af y_{t+2} er betinget af periode t således:

$$E_t[y_{t+2}] = a_0 + a_1 E_t[y_{t+1}]$$

$E_t[y_{t+1}]$ kan erstattes ved at bruge ligning (10.2):

$$E_t[y_{t+2}] = a_0 + a_1(a_0 + a_1 y_t)$$

Det kan udledes, at forecast for y_{t+1} kan bruges til at forecast y_{t+2}. Dette princip betyder, at forecast kan blive konstrueret ved brug af frem af rettet iteration. Overordnet kan forecastet for y_{t+j} bruges til at forecaste en periode frem y_{t+j+1}. Siden princippet fra ovenstående ligninger kan bruges, kan y_{t+j+1} som

skrives: $y_{t+j+1} = a_0 + a_1 y_{t+j} + \varepsilon_{t+j+1}$ omskrives til:

$$E_t[y_{t+j+1}] = a_0 + a_1 E_t[y_{t+j}] \qquad (10.3)$$

Det kan udledes fra ligning (10.2) og ligning (10.3) at det er muligt af estimere en hel serie af *j-step-ahead* forecast ved at bruge frem af rettet iteration. Hvis processen fortsættes fra ligning (10.3), findes ligningen, som Enders kalder for **forecast funktionen**. Funktionen udtrykker alle *j-step-ahead* forecast som en funktion af information fra periode t:

$$E_t[y_{t+j}] = a_0(1 + a_1 + a_1^2 + \ldots + a_1^{j-1}) + a_1^j y_t$$

Ligning (10.1) er blevet estimeret på baggrund af observeret data, hvilket betyder, de forecastede værdier, der blive estimeret for denne model, er *out of sample* forecast. To-periode forecastet y_{t+2} anvender forecastet for den tidligere periode, der findes ved ligning (10.2), hvilket gør det svært at yderligere konkludere, om AR(1) modellen giver fornuftige forecasts.
Projektet anvender flere forskellige modeller til at forecaste y. Det er derfor nødvendig at bruge en metode, der gør det muligt at sammenholde de forskellige modellers forecast.
Det observerede data der anvendes i projektet har 94 observationer, men i stedet for at estimere vores benchmark model på alle observationerne, reduceres datasættet til f.eks. 80 observationer. Der er nu 14 observationer, der ikke bruges til model estimeringen, men dog er kendt. Når f.eks. en AR(1) model er blevet estimeret på det reduceret datasæt, kan modellen bruges til at forecaste for en periode frem, hvor den estimeret forecast værdi for notationen $N_{1,i} = \hat{y}_{80+i}$. Siden vi allerede kender den observerede værdi for periode y_{80}, kan forecast-fejlen for AR(1) modellen konstrueres. AR(1) modellen kan nu blive genestimeret ved brug af det observerede data, hvor en ekstra observation tages med, således at modellen bliver estimeret på 86 observationer. Denne proces fortsættes til at alle de estimeret forecast $N_{1,14}$ er fundet.

AR modellen, som bliver anvendt til benchmark, holdes op mod en VAR model. VAR modellen benyttes, da der er flere forskellige tidsserier. Der kan her refereres til den overordnede makroøkonomiske gennemgang, hvor det tydeligt vides, at huspriser bliver påvirket af andre variabler. Det vil derfor være interessant at estimere en VAR model, således at de forskellige variablers observationssekvens påvirkning på huspriserne kan udledes. Først vil teorien bag en VAR model gennemgås, hvorefter fokusområdet for VAR forecast vil blive teoretiseret.

VAR modellen, der blive brugt til gennemgang af teorien, er en to-variable VAR(1). Vi kan her lade tidsserien $\{y_t\}$ være påvirket af tidsserien for $\{z_t\}$ og vice versa. Det simple bivariant system kan opskrives som:

$$y_t = b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \varepsilon_{yt} \quad (10.4)$$

$$z_t = b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \varepsilon_{zt} \quad (10.5)$$

Det er antaget, at data er stationært, hvilket er blevet testet i afsnit 8 i projektet for det observerede data, der anvendes til at modellere en VAR. ε_{yt} og ε_{zt} antages at være hvid støj og ikke korreleret med standardafvigelsen σ_y og σ_z. Ligning (10.4) og (10.5) er begge af lag ordren én, siden der kun er inkluderet ét lag. Systemet for en VAR model tillader at inkorporeret feedback, fordi både y_t og z_t påvirker hinanden. Ligningerne, som bliver påvirket af begge variabler, er ikke i stand til at blive estimeret ved hjælp af OLS pga. der er en simultan effekt mellem de to. OLS estimaterne vil lide under at koefficienterne og fejlleddene er korreleret, og der derfor opstår bias estimater. For at undgå dette problem, bliver ligningerne (10.4) og (10.5) transformeret til et ligningssystem på følgende matrix form:

$$\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix}\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}\begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix}$$

Dette kan skrives mere kompakt:

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \varepsilon_t$$

Hvor $$B = \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix},\ x_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix},\ \Gamma_0 = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix}, \Gamma_1 = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

$$\text{og } \varepsilon_t = \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix}$$

Ved at Pre-multiplicere det forkortede ligningssystem med B^{-1}, opnås en standard VAR form:

$$x_t = A_0 + A_1 x_{t-1} + e_t \quad (10.6)$$

Notationerne i ligning (10.6) angiver: $A_0 = B^{-1}\Gamma_0$, $B^{-1}\Gamma_1$ og $e_t = B^{-1}\varepsilon_t$.

Notationerne for VAR modellen bliver nu ændret. Ligningen (10.6) kan nu bruges til at definere a_{i0} som

element *i* af vektoren A_0. a_{ij} bliver defineret som elementet i række *i* , kolonne *j* af matrixen A_1.

Fejlleddet e_{it} bliver defineret som elementet *i* af vektoren e_t. De nye notationer bruges nu til at omskrive ligning (10.6) til:

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}z_{t-1} + e_{1t} \qquad (10.7)$$

$$z_t = a_{20} + a_{21}y_{t-1} + a_{22}z_{t-1} + e_{2t} \qquad (10.8)$$

Der er forskel mellem de to ligningssystemer opskrevet i ligning (10.4) og (10.5), og det udledte ligningssystem i ligning (10.7) og (10.8). Ligningssystemet opskrevet i ligning (10.4) og (10.5) er defineret som en strukturel VAR, mens ligningerne (10.7) og (10.8) er refereret til som en standard VAR form. Til modelleringen af ligningssystemet i projektet, der bliver gennemgået i afsnittet om modeludvælgelsen brugen af standard VAR form.

Efter at VAR modellen er blevet estimeret, kan den blive brugt til multiforecasting. VAR(1) modellen gennemgået ovenover og angivet i (10.6), og vil blive brugt som et eksempel på, hvordan en VAR forecaster. De estimerede koefficienter for A_0 og A_1 kan nu blive brugt til at lave *one step ahead* forecast. *One step ahead* forecastet kan blive udledt ved at bruge relationen:

$$E_T[x_{t+1}] = A_0 + A_1x_T \qquad (10.9)$$

T angiver størrelsen på det observerede data, som VAR modellen er blevet estimeret på. Two step ahead forecastet findes ved at bruge relationen:

$$E_T[x_{T+2}] = A_0 + A_1E_T[x_{T+1}]$$

Ligning (10.9) kan nu substitueres ind, således at forecast værdien for one step ahead anvendes til at estimere forecast for *two step ahead*.

$$E_T[x_{T+2}] = A_0 + A_1[A_0 + A_1x_T] \qquad (10.10)$$

De observerede værdier er blevet reduceret til 80 ligesom ved AR eksemplet. Hver gang et *one step ahead* forecast er blevet udført, bliver en ny VAR(1) model estimeret på det observerede data, hvor der er inkluderet en ekstra observationsværdi $T+1$. Det er derfor ikke nødvendigt, hvis VAR modellerne estimeres på et reduceret datasæt, at anvende *one step ahead* forecast til at estimere *two step ahead* forecast, som det bliver forslået i ligning (10.10).

11 Modeludvælgelse

I dette afsnit vil modeludvælgelsen finde sted. Projektet anvender en AR model som benchmark model, og holder en VAR model op mod denne benchmark, for at teste hvilken model der estimerer de bedste forecast. Inden evalueringen kan finde sted, skal udvælges af hvor mange lags de enkelte modeller skal indeholde. Metoderne der anvendes til at finde den bedste lag-længde, er Akaike Information kriterium (AIC) og Bayesian information kriterium (BIC) for hver af de enkelte VAR modeller.

Det er vigtigt at undersøge, hvorledes en model fittes til data. Ved at tillægge ekstra lags til en model f.eks. en VAR, så vil *sum of squares* for modellens estimerede residualer reduceres, således modellen bliver mere præcis. Dog er det ikke uden konsekvenser at tillægge et ekstra lag til en model. Et ekstra lag betyder, at flere koefficienter skal estimeres, hvilket fører til tab af frihedsgrader[36]. Antallet af koefficienter, der skal estimeres, er afhængig af hvor mange variabler, der er med i VAR modellen.

Det primære fokus i modellen er, at kunne anvende de estimerede modeller til forecast, og hvis der inkluderes unødvendige koefficienter i modellen, kan det påvirke forecast præstationen af den estimerede model. To meget udbredte selektions kriterier er Akaike og Bayesian. Begge metoder vil blive anvendt i projektet til at afgøre den ideelle lag-længde for VAR modellerne for hele Danmark og de underlæggende regionerne.
AIC findes ved at estimere *sum of squared residuals* se afsnit 8 ligning (8.9) for specifik definition. De to selektions kriterier, henført tidligere nævnte, er givet i de understående ligninger[37]:

$$AIC = T * \ln(SSR) + 2n \qquad (11.1)$$

$$BIC = T * \ln(SSR) + n * \ln(T) \qquad (11.2)$$

Hvor T angiver mængden af brugbare observationer. n er antallet af estimerede parameter i modellen. n afhænger af hvor mange lags, der er inkluderet, samt om der er en konstant herimellem. $\ln(SSR)$ er den naturlige logaritme af *sum of squared residuals* fra den estimerede model. Der findes også andre formler for udregning af AIC og BIC, men for nu bruges Walter Enders' formler for AIC og BIC[38].

T værdien i AIC og BIC skal holdes fast, for at få et præcist estimat. Eftervirkningerne af AIC estimeret med en foranderlig T, er at AIC ikke længere finder den bedste model, men derimod den model med det laveste antal observationer. Grundlaget for dette, er at T har en direkte reducerende effekt på AIC, hvilket

[36] SAMPLE AUTOCORRELATIONS OF STATIONARY SERIES. Enders, Walter. Side 69-70
[37] SAMPLE AUTOCORRELATIONS OF STATIONARY SERIES. Enders, Walter. Side 69-70
[38] SAMPLE AUTOCORRELATIONS OF STATIONARY SERIES. Enders, Walter. Side 69-70

også kan ses i ligning (11.1), hvor T er ganget på $\ln(SSR)$. Den samme eftervirkning gøre sig også gældende for BIC. Den konstante værdi for T findes ved at bruge antallet af fri observationer for alternativ modellen, hvilket vil sige den model, der er blevet estimeret med flest lags. Udvælgelsen af den bedste model foregår ved at finde den model, hvor AIC og BIC er lavest. Forbedring af en models fit af data forbedrer også AIC og BIC. Projektet estimerer en VAR model med 1 til 7 lags, målet er her at finde den VAR model med den bedste AIC og BIC værdi. Tilføjelsen af et ekstra lag betyder at n stiger, samt *sum of squared residuals* reduceres. Dog hvis de ekstra regresser ikke har nogle forklarende kraft, vil tilføjelsen af dem få AIC og BIC til at stige. Grundlaget for at BIC inkluderes, er at den naturlige logaritme tages af T, hvilket betyder at BIC vil vælge en mere parametersparsom model end AIC[39].

Akaike Information kriterium og Bayesian information kriterium udregnes i Matlab. Matlab funktionen for AIC og BIC er mere udvidet end Walter Enders, derfor vil AIC og BIC funktionerne for Matlab blive illustreret:

$$\mathrm{AIC} = T*\log(\det(\frac{1}{T}\sum_{1}^{T}\varepsilon(t,\hat{\theta}_N)(\varepsilon(t,\hat{\theta}_N))^T)) + 2n + T*(n_y*(\log(2\pi)+1))$$

$$BIC = T*\log(\det(\frac{1}{T}\sum_{1}^{T}\varepsilon(t,\hat{\theta}_N)(\varepsilon(t,\hat{\theta}_N))^T)) + T*(n_y*\log(2\pi)+1) + n*\log(T)$$

Funktionerne er blevet omskrevet, således at de før gennemgået notationer i projektet holdes de samme. $\varepsilon(t)$ er en vektor med prædiktionsfejl. $\hat{\theta}_N$ er repræsentativ for de estimerede parametre i modellen. n_y er antallet af model outputs. Det kan ses, at de første led er det samme som i Enders, dog er determinanten taget af residualerne. AIC i Matlab tager også antallet af model outputs med i evalueringen af modellen. Det samme gør sig gældende for BIC. Igen er BIC mere fokuseret på parameter-sparsommelighed, fordi den igen har det sidste led med, hvor den tager logaritmen til antallet af brugbare observationer T.

Outputtet fra AIC og BIC, som er udregnet i Matlab, vil nu blive gennemgået. VAR modellerne, der bliver estimeret til at fitte data, er i lag rækkevidden 1 til 7 lags.
VAR modellerne er blevet estimeret ved hjælp af loop funktionen i Matlab. Det har derfor været muligt at estimere alle VAR modellerne på en gang. Fra de estimerede modeller er AIC og BIC blevet udledt ved hjælp af de ovenstående funktioner. I tabel (11.1) er alle de udledte AIC og BIC estimater illustreret.

[39] SAMPLE AUTOCORRELATIONS OF STATIONARY SERIES. Enders, Walter. Side 69-70

Tabel (11.1): VAR(1:7) AIC og BIC værdier

VAR hele danmark	AIC	BIC	VAR Hovedstanden	AIC	BIC	VAR Midtjylland	AIC	BIC
VAR(1)	-2089,2	-2041,81	VAR(1)	-2026,37	-1978,98	VAR(1)	-2043,24	-1995,86
VAR(2)	-2287,38	-2202,54	VAR(2)	-2254,65	-2169,81	VAR(2)	-2248,19	-2163,35
VAR(3)	-2248,2	-2126,32	VAR(3)	-2215,55	-2093,67	VAR(3)	-2218,92	-2097,04
VAR(4)	-2218,58	-2060,09	VAR(4)	-2194,52	-2036,03	VAR(4)	-2182,68	-2024,19
VAR(5)	-2173,6	-1978,93	VAR(5)	-2152,97	-1958,3	VAR(5)	-2139,51	-1944,84
VAR(6)	-2260,32	-2029,91	VAR(6)	-2222,21	-1991,81	VAR(6)	-2220,84	-1990,44
VAR(7)	-2255,51	-1989,82	VAR(7)	-2214,87	-1949,17	VAR(7)	-2221,43	-1955,74
VAR Nordjylland	AIC	BIC	VAR Sjælland	AIC	BIC	VAR Sydjylland	AIC	BIC
VAR(1)	-2005,37	-1957,98	VAR(1)	-1990,64	-1943,25	VAR(1)	-2031,4	-1984,01
VAR(2)	-2211,22	-2126,38	VAR(2)	-2207,53	-2122,69	VAR(2)	-2246,78	-2161,94
VAR(3)	-2170,09	-2048,21	VAR(3)	-2161,98	-2040,11	VAR(3)	-2198,31	-2076,44
VAR(4)	-2132,87	-1974,38	VAR(4)	-2129,81	-1971,32	VAR(4)	-2164,71	-2006,22
VAR(5)	-2096,85	-1902,18	VAR(5)	-2102,19	-1907,52	VAR(5)	-2126,88	-1932,21
VAR(6)	-2188,64	-1958,23	VAR(6)	-2202,04	-1971,64	VAR(6)	-2223,28	-1992,88
VAR(7)	-2186,12	-1920,43	VAR(7)	-2168,15	-1902,46	VAR(7)	-2205,72	-1940,02

I tabel (11.1), hvor de enkelte VAR modellers AIC og BIC er blevet illustreret, findes den laveste værdi for begge selektions kriterier. For Hele Danmark vælger AIC en VAR(2) som værende den bedste model, med en AIC-værdi på -2287,38. Den mere sparsommelige BIC vælger også en VAR(2) for hele Danmark, da den laveste værdi er for en VAR med to lags -2202,54. Den sammen konklusion kan drages ved VAR modellerne for regionerne, her er de bedste modeller også en VAR(2). De laveste AIC og BIC værdier er i tabel (11.1) markeret med grøn farve.

VAR model for hele Danmark

Ved hjælp af Akaike Information kriterium og Bayesian information kriterium blev en VAR(2) model udvalgt, som værende den bedste til at fitte data for hele Danmark. Estimeringen af modellen bliver fortaget i Matlab, og output fra estimeringen vil blive gennemgået i dette underafsnit.

Matlab har en allerede eksisterede funktion, der kan bruges til at opsætte en VAR model, men med en illustrativ hensigt vil VAR modellen for hele Danmark, skrives op på en standard VAR form. VAR teorien fra afsnittet ***Forecasting*** vil blive anvendt til dette formål. VAR modellen, der er blev udledt i afsnittet, var en VAR(1), med to variabler, denne model vil nu blive omskrevet til en VAR(2) med 4 variabler, således at den stemmer overens med det observerede data for projektet. Ligning (10.6) i ***Forecasting*** afsnittet vil blive anvendt til at opskrive VAR(2) modellen.

$$x_t = A_0 + A_1 x_{t-1} + A_2 x_{t-2} + e_t$$

Ligningssystemet for den ovenstående VAR(2) vil nu blive præsenteret. I afsnittet *Databeskrivelse* er det observerede data, som anvendes til at estimere VAR(2) modellen, gennemgået. Huspriserne får notationen h_t. Indkomst får notaionen Δy_t. Skat får notationen Δs_t. Renten får notationen r_t.

$$h_t = a_{10} + a_{11}h_{t-1} + a_{12}\Delta y_{t-1} + a_{13}\Delta s_{t-1} + a_{14}r_{t-1} + b_{11}h_{t-2} + b_{12}\Delta y_{t-2} + b_{13}\Delta s_{t-2} + b_{14}r_{t-2} + e_{1t}$$

$$\Delta y_t = a_{20} + a_{21}h_{t-1} + a_{22}\Delta y_{t-1} + a_{23}\Delta s_{t-1} + a_{24}r_{t-1} + b_{21}h_{t-2} + b_{22}\Delta y_{t-2} + b_{23}\Delta s_{t-2} + b_{24}r_{t-2} + e_{2t}$$

$$\Delta s_t = a_{30} + a_{31}h_{t-1} + a_{32}\Delta y_{t-1} + a_{33}\Delta s_{t-1} + a_{34}r_{t-1} + b_{31}h_{t-2} + b_{32}\Delta y_{t-2} + b_{33}\Delta s_{t-2} + b_{34}r_{t-2} + e_{3t}$$

$$r_t = a_{40} + a_{41}h_{t-1} + a_{42}\Delta y_{t-1} + a_{43}\Delta s_{t-1} + a_{44}r_{t-1} + b_{41}h_{t-2} + b_{42}\Delta y_{t-2} + b_{43}\Delta s_{t-2} + b_{44}r_{t-2} + e_{4t}$$

Det præsenteret ligningssystem for hele Danmark skrives nu op på vektor form, således at den stemmer overens med ligning (10.6) fra ***Forecasting*** afsnittet.

$$\begin{bmatrix} h_t \\ \Delta y_t \\ \Delta s_t \\ r_t \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \\ a_{30} \\ a_{40} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} h_{t-1} \\ \Delta y_{t-1} \\ \Delta s_{t-1} \\ r_{t-1} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} \begin{bmatrix} h_{t-2} \\ \Delta y_{t-2} \\ \Delta s_{t-2} \\ r_{t-2} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ e_{4t} \end{bmatrix} \quad (11.3)$$

Ligningen (11.3) er den opsatte vektor autoregressive model, der bliver estimeret i Matlab. Funktionen der bliver anvendt til at opstille VAR(2) modellen i ligning (11.3), kaldes for ***varm.*** Funktionen tillader at notere hvor mange variabler, der skal inkluderes i modellen, samt hvor mange lags modellen skal indeholde. VAR(2) modellen kan nu estimeres for hele Danmark, ved hjælp af funktion ***estimate***. ***estimate*** funktionen tillader også at trække residualerne ud fra den estimerede VAR(2) model, således der kan fortages analyse af hvor godt modellen fitter dataet. Dette bliver der gjort brug af i næste afsnit *evaluering af forecast.*

Ligning (11.4): Estimeret VAR(2) for hele Danmark.

$$\begin{bmatrix} h_t \\ \Delta y_t \\ \Delta s_t \\ r_t \end{bmatrix} = \begin{bmatrix} 0{,}008039 \\ -3{,}0947e-06 \\ -0.000110 \\ -0.017283 \end{bmatrix} + \begin{bmatrix} 0{,}0799 & 2{,}3876 & 0{,}5026 & -0{,}1483 \\ -0{,}0022 & 1{,}6076 & 0{,}0025 & 0{,}00039 \\ 0{,}0044 & 0{,}0636 & 1{,}5005 & -0{,}0034 \\ 0{,}3686 & -4{,}0840 & -0{,}4499 & 0{,}3987 \end{bmatrix} \begin{bmatrix} h_{t-1} \\ \Delta y_{t-1} \\ \Delta s_{t-1} \\ r_{t-1} \end{bmatrix}$$
$$+ \begin{bmatrix} -0{,}0128 & -0{,}9789 & -0{,}5280 & 0{,}00076 \\ -0{,}00091 & -0{,}8927 & 0{,}0093 & -0{,}0020 \\ -0{,}00018 & -0{,}0294 & -0{,}8021 & 0{,}0021 \\ 0{,}4717 & -0{,}7375 & -0{,}9754 & -0{,}1071 \end{bmatrix} \begin{bmatrix} h_{t-2} \\ \Delta y_{t-2} \\ \Delta s_{t-2} \\ r_{t-2} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ e_{4t} \end{bmatrix}$$

Den estimerede VAR(2) model for hele Danmark er gengivet i ligning (11.4). Det er svært at uddrage en effekt fra indkomst og skat, da differencen er taget to gange af det observerede data. Det kan dog ses at

huspriserne er positiv påvirket af høje værdier i Δy_{t-1} og Δs_{t-1}. r_{t-1}, som angiver ændringen i renten, har et negativ fortegn ved det første lag (-0,1483), hvilket betyder at renten har en negativ effekt på huspriserne. At renten påvirker huspriserne negativt, stemmer overens med den makroøkonomiske teori. Variablerne ved lag to, hvor husprisen er den afhængige variabel, har ændret fortegn. Den estimerede h_t bliver påvirket negativt, hvis de observerede værdier for lag to har været høje. Det er vigtig at VAR(2) modellen ikke overfortolkes, da hovedmålet med modellen er at anvende den til *out of sample* forecast.

12 Evaluering af forecast

I afsnittet for modeludvælgelse blev Akaike og Bayesian selektions kriterium anvendt til at finde den rigtige lag-længden for VAR modellerne for regionerne og hele Danmark. Begge selektions kriterierne havde den laveste værdi ved VAR modeller med to lags.
De nøje udvalgte VAR(2) modeller blev her efter estimeret, samt blev VAR(2) modellen for hele Danmark, i underafsnittet *VAR model hele Danmark,* gennemgået. VAR(2) modellen for hele Danmark bliver brugt som initial værdier for forecast funktionen i Matlab, således det første step ahead forecast kan udregnes, denne proces fortsættes indtil det valgte forecast interval er beregnet. Det samme gøres for de enkelte regioner i Danmark. Forecast for benchmark modellen bliver estimeret på samme måde, dog har der ikke været brug for at finde den samme optimale lag-længde som ved VAR modellerne. AR(1) modellen er som sagt valgt som benchmark, da modellen har meget sparsomme parametre estimeringer. Her endes der til sidst op med 6 model forecast ved brug af VAR(2) modellen, og 6 model forecast ved brug af AR(1) modellen.
De estimerede forecast for regionerne og hele Danmark vil nu blive gennemgået i de næste to afsnit. Det første afsnit er med et forecast interval på 14 steps ahead. Det andet afsnit anvender et forecast interval på 26 *steps ahead*.

Efter der nu er fortaget forecast af de 6 VAR modeller og de 6 AR modeller, vil disse forecasts blive testet mod hinanden, for at evaluere forecastene fra VAR modellerne. Denne evaluering er nødvendig at foretage, for at kunne vurdere om VAR modellerne, der er blevet geneneret på baggrund af de 4 makroøkonomiske faktorer, som der er forventet at skulle have effekt på ejendomsmarkedet, er bedre end en simpel AR(1) model af udviklingen i ejendomspriserne alene. Hvis ikke der er en signifikant forskel på modellernes evne til at forecaste de fremtidige værdier for udviklingen i ejendomspriserne, kan det ikke endeligt bevises at VAR modeller er bedre til at forecaste. Hvis dette skulle være tilfældet, fører genereringen af VAR modeller til en unødvendig overparametrisering, og dermed tab af frihedsgrader.

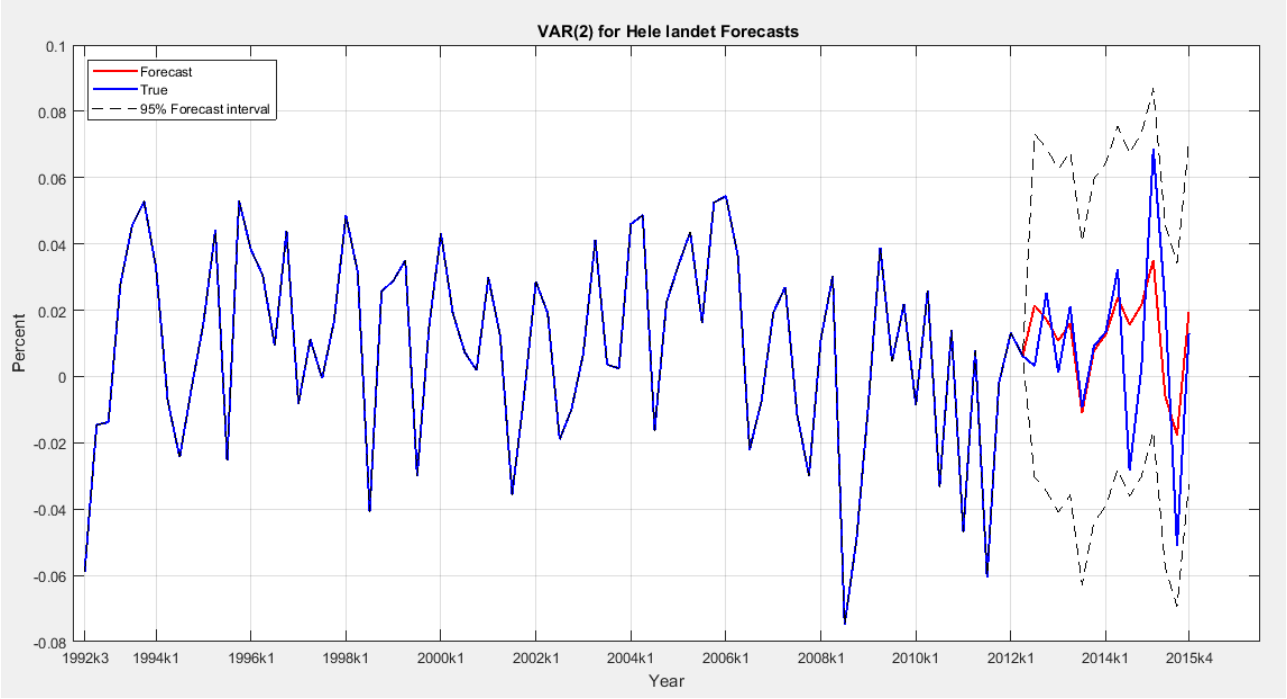
Til dette vil der blive benyttet en række tests til at vurdere, om VAR modellerne er bedre end deres modpart AR(1) modellerne, som her vil blive brugt som benchmark. Da det ikke er muligt at vurdere

forecasts på fremtidigt data, er der til dette projekt blevet estimeret VAR modeller på 80 observationer ud af de i alt 94 observationer. De resterende 14 observationer vil blive brugt til *out of sample* forecast. Ved *out of sample* forecast kan de forecastede værdier blive sat op imod de observerede værdier, og derved kan det vurderes, hvor præcise de forecastede værdier er på de faktiske værdier. Der vil til forklaring af evalueringsprocessen hovedsageligt blive taget udgangspunkt i modellerne for hele Danmark.

Først vil der blive set på 2 grafiske illustrationer af det observerede data og de forecastede værdier for henholdsvis VAR(2) modellen for hele Danmark, og AR(1) modellen for hele Danmark. Her bliver ligeledes indsat et 95 procents konfidensinterval for forecastene. Konfidensintervallet beregnes ved at tillægge den estimerede værdi, $\hat{y}_t$, og tillægge de to 95 procents konfidensinterval værdier til forecastene standardafvigelse $\hat{\sigma}$[40]:

$$\hat{y}_t \pm 1.96\hat{\sigma}$$

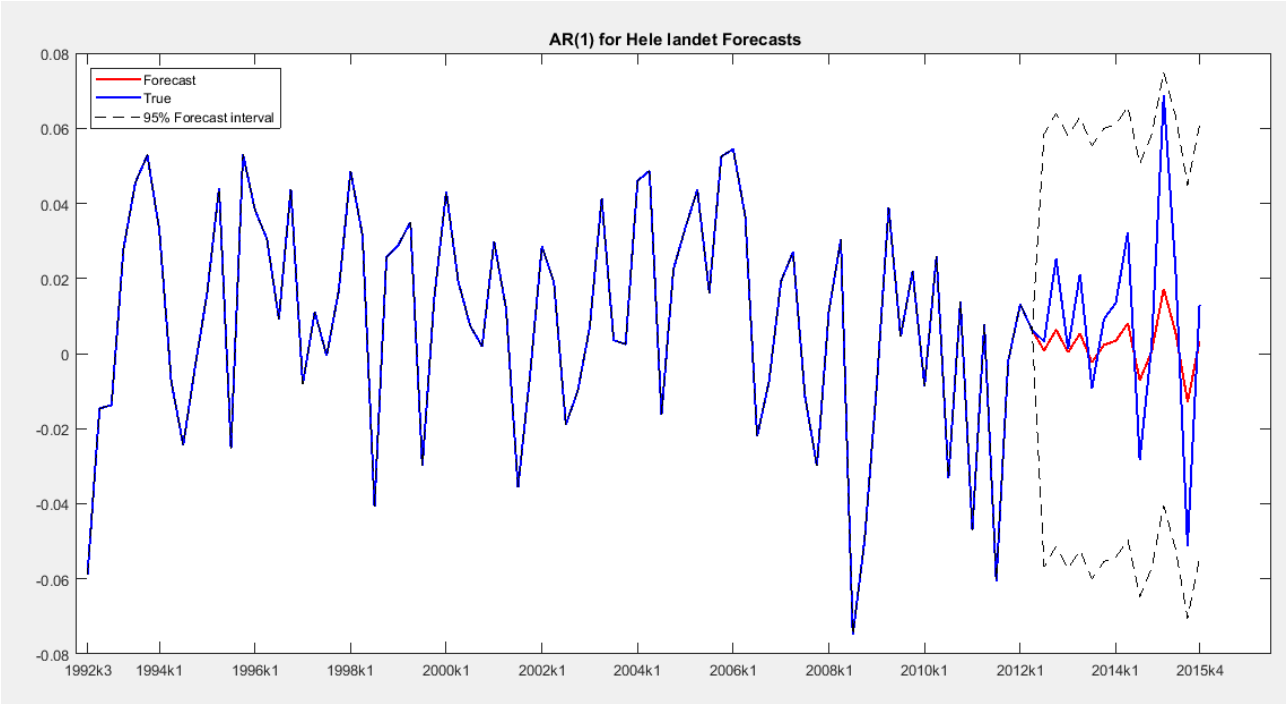
Figur (12.1): VAR(2) Udviklingen i huspriserne for hele Danmark



Som det kan ses i figur (12.1) følger de forecastede værdier fra VAR modellen de observerede værdier rimeligt godt. Med undtagelse af et par enkelte forecasts i starten af perioden, følger forecastene udviklingen i de observerede værdier. Særligt i perioden 2013k2 til 2014k2 fanger VAR modellens forecastede værdier udviklingen i perioden næsten perfekt. Dette fortsætter dog ikke, da den volatilitet, der kommer i efterfølgende periode 2014k3 og frem, ikke bliver fanget af forecastene.

[40] Prediction intervals, Otexts

Figur (12.2): AR(1) Udviklingen i huspriserne for hele Danmark



Kilde: Dansk statistisk og egne beregninger.

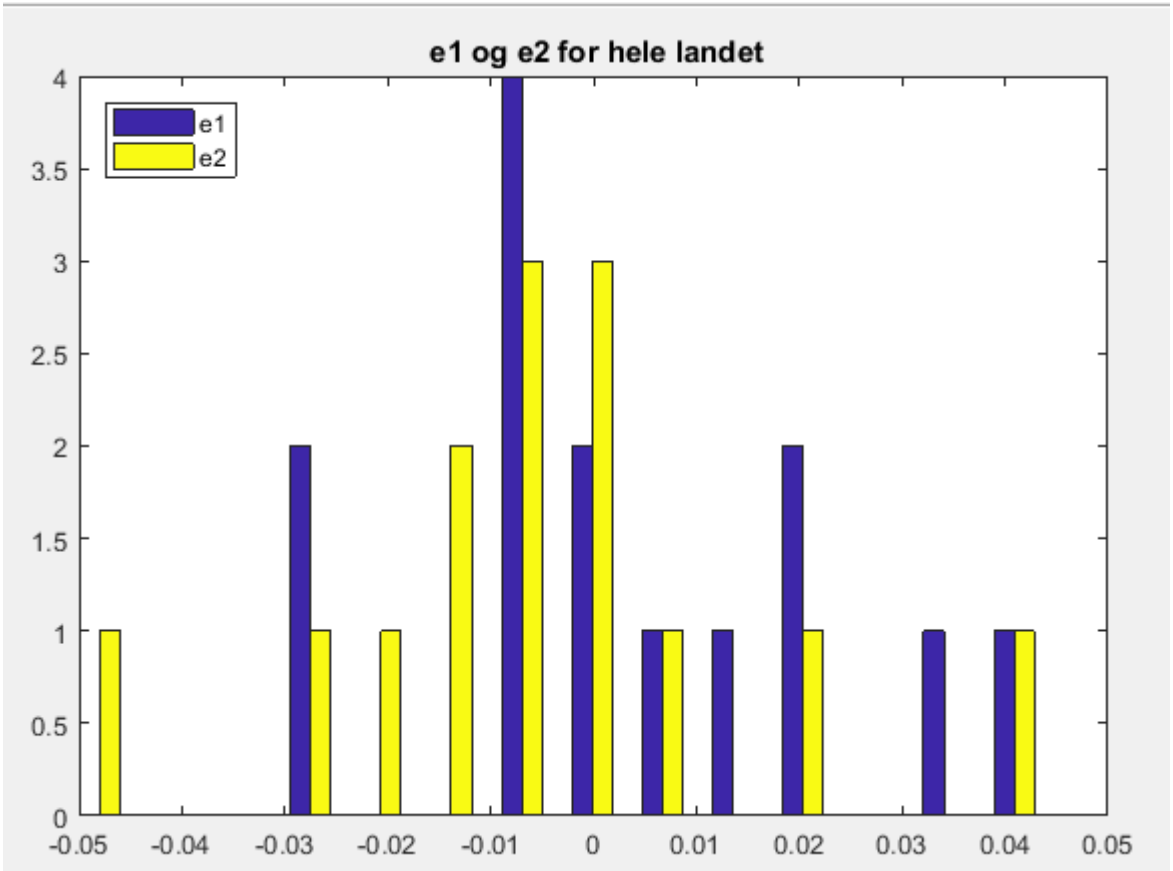
AR(1) modellen for hele Danmark i figur (12.2) giver ikke lige så præcise forecasts som VAR(2)-modellen. Skønt AR(1) modellen fanger udviklingen i de observerede værdier, så rammer forecastene langt fra størrelsen på udsvingene i ejendomsprisudviklingen.

Test af residualer

Inden der bliver udført test for forecastenes præcision, vil der først blive sat fokus på residualerne. Her er antagelsen at forskellen på de observerede værdier og de forecastede værdier, de såkaldte residualer, følger en normalfordeling med en middelværdi på nul, og en varians på σ^2, samt der ikke er autokorrelation i residualerne. Dette kaldes også for hvid støj. [41]

[41] PROPERTIES OF FORECASTS. Enders, Walter, side 85

Figur (12.3): residualerne for VAR(2)-modellen, *e1*, og residualerne for AR(1)-modellen, *e2*



I figur (12.3) er der blevet plottet et histogram af residualer for *e1* og *e2*, for at se om de følger en normalfordeling. Her er der dog for få antal af residualer, hvilket ikke gør det muligt at konkludere om de følger en normalfordeling.

Da det lave antal af residualer gør, at det ikke er muligt, at bruge histogrammet til at vurdere om residualer følger en normaldeling, bliver der udført en *Ljung-Box* test, for at teste om der er autokorrelation i residualerne. I *Ljung-Box* testen er H_0-hypotesen, at der ikke er autokorrelation i residualerne, og da dette ikke kan forkastses ved testen af *e1* og *e2*, antages det, at residualerne er hvid støj.

Der er blevet benyttet *Ljung-Box* tests på alle residualerne for alle modellerne i projektet. Her har det ikke været muligt at forkaste H_0-hypotesen, og dermed antages det at alle residualerne for alle modellerne er hvid støj.

MSPE

Da det ikke er muligt at se en decideret forskel på forecastene i figur (12.1) og (12.2), kan det være mere optimalt at benytte *mean square prediction error (MSPE)*. Her ses på de gennemsnitlige kvadrerede forecasts residualer, hvilket gøres ved at tage summen af alle residualerne, og dividere dette med antallet af forecasts. Formlen for *MSPE* er:

$$MSPE = \frac{1}{H}\sum_{i=1}^{H} e_{ji}^2$$

Her er *H* antallet af *one step ahead* forecasts og e_{ji} er forskellen mellem de observerede værdier og de forecastede værdier.

Hvis der ses på forecastene fra de to modeller for hele Danmark, har forecastene fra VAR modellen for hele Danmark en MSPE på 0,000425, mens forecastene for AR modellen for hele Danmark har 0,000452. Skønt forskellen er marginal, er VAR modellen for hele Danmark bedre end AR modellen for hele Danmark, da MSPE her er lavere. I den nedenstående tabel (12.1) er MSPE for alle modellerne vist.

Tabel 12.1: MSPE for alle modellerne

	MSPE for VAR(2)	MSPE for AR(1)
Hele Danmark	0,000425	0,000452
Hovedstaden	0,000478	0,000428
Midtjylland	0,000924	0,000665
Nordjylland	0,0018	0,0016
Sjælland	0,0011	0,0010
Sydjylland	0,0012	0,000922

MSPE-værdierne i tabel (12.1) er beregnet i Matlab. Som det kan ses i tabel (12.1) er det kun VAR modellen for hele Danmark, der har en MSPE, som er lavere end MSPE for AR(1) modellen for hele Danmark. Da de resterende MSPE værdier for VAR modellerne tilhørende regionerne er højere end MSPE værdierne for AR(1) modellerne, betyder dette, at for de enkelte regioner er AR(1) modellerne bedre til forecaste de fremtidige værdier for udviklingen i huspriserne.

Diebold Mariano

I 1995 udgav Francis X. Diebold og Roberto S. Mariano *comparing Predictive accuracy*, hvori de fremlagde en såkaldt DM-test. Denne DM-test benyttes til at vurdere, om der er forskel i præcisionen af forecastene, som er baseret på henholdsvis den estimerede VAR model og den estimerede AR model. [42]

Der ligges ud med at kigge på de 2 forecasts residualer, som her defineres, ved forskellen mellem den forecastede værdi på tidspunkt *t*, $\hat{y}_{it}$, og den observerede værdi på tidspunkt *t*, y_t. Her dannes der 2 residualer for forecasts model *j*:

$$e_{jt} = \hat{y}_{jt} - y_t \qquad j = 1{,}2$$

På baggrund af forecastenes residualer dannes en *tabsfunktion*, *g*, som er de kvadrerede residualer for de enkelte modeller. Tabsfunktionen *g* vil være nul, hvis forecastene er 100 procent præcise. Ligeledes kan *g* ikke tage en negativ værdi:

$$g(e_{jt}) = e_{jt}^2 \qquad j = 1{,}2$$

Herefter beregnes d_t, som værende forskellen mellem de 2 $g(e_{jt})$:

$$d_t = g(e_{1t}) - g(e_{2t})$$

Hvis d_t her er lig med 0, må størrelsen på de 2 tabsfunktioner være ens, og derfor er ikke forskel i størrelsen af de 2 tabsfunktioner. Kun hvis d_t er lig med nul for alle værdier af *t*, vil modellernes forecasts være lige præcise.
Derved dannes nulhypotesen på baggrund af forventningen til d_t. Hvis den forventede værdi til d_t er lig med nul, er der ikke forskel på de 2 forecasts præcision:

$$H_0: E(d_t) = 0$$

Alternativ hypotesen er, at der er forskel på de 2 modellers præcision. Herved vil den forventede værdi til d_t ikke være lig med nul[43]:

$$H_0: E(d_t) \neq 0$$

For at finde den forventede værdi til d_t, beregnes den gennemsnitsnitlige værdi af alle d_t:

$$\bar{d} = \frac{1}{H}\sum_{i=1}^{H}[g(e_{1t}) - g(e_{2t})]$$

[42] Comparing Predictive Accuracy of two Forecasts: The Diebold-Mariano Test side 2-3
[43] Comparing Predictive Accuracy of two Forecasts: The Diebold-Mariano Test side 2-3

Hvor *H* er antallet af forecast.

Hvis det er muligt at finde variansen til $\bar{d}$, kan der opsættes en ratio:

$$DM = \bar{d}/\sqrt{var(\bar{d})} \qquad (12.1)$$

Givet at de kvadrede forecasts residualer, d_t, er normalt fordelt med en varians på σ^2, er den estimerede varians[44]:

$$var(\bar{d}) = \sigma^2/(H-1)$$

Derved kan DM-teststørrelsen[45] beregnes således:

$$DM = \bar{d}/\sqrt{(\frac{\sigma^2}{[H-1]}])}$$

DM-teststørrelsen følger en t-statistik. Derfor for at finde de kritiske værdier for DM-statistikken, findes de kritiske værdier for en t-statistik, med H-1 frihedsgrader.

I den nedenstående tabel (12.2) er alle de beregnede DM-teststørrelserne for hele Danmark og de fem regioner. DM- teststørrelserne er blevet beregnet i Matlab ved hjælp af funktionen *dmtest*. Denne funktion kan findes i bilaget ved navnet *dmtest*.

Tabel (12.2): Diebold Mariano test statistikker for hele Danmark og de fem regioner

	DM
Hele Danmark	-0,1367
Hovedstaden	0,2476
Midtjylland	0,8306
Nordjylland	0,5760
Sjælland	0,1548
Sydjylland	1.1604

DM-teststørrelserne testes imod en t-statistik med 13 frihedsgrader og et signifikantniveau på 5 procent. Her er det kritiske niveau for t-statistikken 2,160[46], hvilket ingen af de overstående DM-teststørrelser ligger over. Derfor kan H_0-hypotesen om, at der ikke er forskel på de 2 modellers præcision, ikke forkastes.

[44] PROPERTIES OF FORECASTS. Enders, Walter, side 86
[45] PROPERTIES OF FORECASTS. Enders, Walter, side 86

Skønt MSPE er forskellige mellem alle VAR og AR(1) modellerne, så fortæller DM-testen, at det ikke kan forkastes at forecastene for VAR modeller og AR(1) modeller har den samme præcision. Dermed havde der været flere forecasts-observationer, kan det ikke entydigt siges, at VAR modellen for hele Danmark er bedre end AR(1) modellen for hele Danmark, og modsat med de enkelte regioner, hvor det ikke definitivt kan siges at AR(1) modeller er bedre end de genererede VAR modeller.

13 Robustheden af forecast

Forventningerne til forecast-analyser afhænger meget af, hvor langt frem i tiden der ønskes at forecastes. De estimerede forecast, i afsnittet om evaluering af forecasts, er udregnet for en kort periode.
Det ønskes i dette afsnit at belyse problematikkerne ved at anvende forecast, til eksempelvis meget korte forecast perioder i forhold til lange forecast perioder. I afsnittet evaluering af forecast kunne det konkluderes ud fra Diebold Mariano testen, at der ikke var en signifikant forskel mellem benchmark modellen og VAR(2) modellen. Begge modeller var derfor lige relevant til at forecast ændringen i boligpriserne for hele Danmark og regionerne. I dette afsnit bliver perioden for forecast udvidet, således at det er muligt at teste om benchmark modellen i forhold til VAR(2) modellen er bedre til langsigtede forecast.

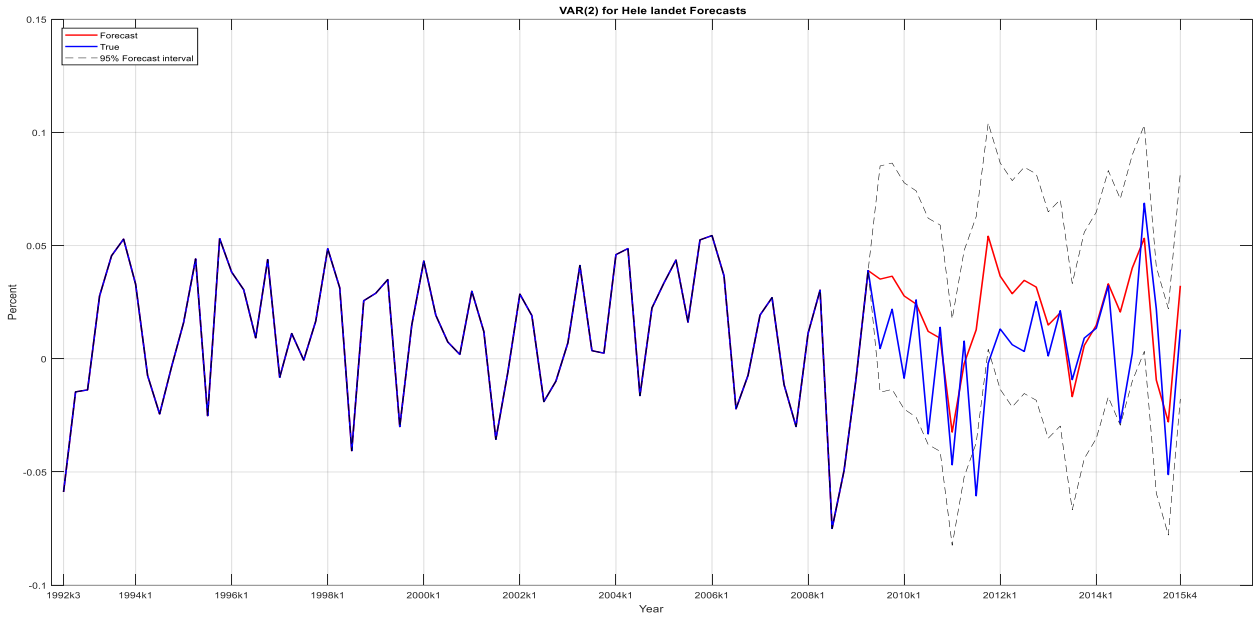
VAR(2) forecast output i afsnit (12) er estimeret på en reduceret datasæt med 80 observationer, ud af det totale datasæt på 94 observationer. Det vil sige, at der er 14 observerede værdier, der ikke anvendes til model estimation, men som tidligere nævnt, benyttes de undladte observationer til at holde op mod de estimerede forecast. Dette tillader, at residualerne for forecastene kan findes, og herefter kan vurderes hvilken model, der forecaster bedst. De udregnede *MSPE* værdier i tabel (12.1) resulterede i at VAR(2) modellen for hele Danmark var den bedste, mens for regionerne var det den autoregressive benchmark model, som dominerede. I Diebold-Mariano testen i tabel (12.2) kunne det konkluderes, at der ikke var en signifikant forskel mellem forecast af AR(1) modellen og VAR(2) modellen. Begge modeller præsterede derfor på samme niveau.
De 14 *steps ahead forecast* der først er blevet estimeret, som er gennemgået i afsnit (12), og er vist grafisk i figur (12.1) for VAR(2) og figur (12.2) for AR(1), følger det observerede data forholdsvis pænt. Det vil derfor være interessant at undersøge, om dette er tilfældigt, at model forecastene er så præcise. For at undersøge dette, vil det reducerede datasæt reduceres endnu engang, hvorefter der estimeres en ny VAR model og ny AR model på det nye reducerede datasæt. Tidsperioden for det nye reducerede datasæt er blevet sat til lige efter at finanskrisen er indtruffet. Det kan ses i afsnittet for databeskrivelsen, at omkring 2007-2008 har

[46] Values of the t-distribution (two-tailed), Medcalc

finanskrisen ført til fald i ejendomspriserne for enfamiliehuse. Det vil derfor blive analyseret om AR modellen og VAR modellen vil være i stand til at fange den nedadgående udvikling. De nye estimerede modeller vil nu blive brugt til at forecaste for en længere periode, end de tidligere gennemgået forecast. Det mest interessante er her, om modellerne vil være i stand til at fange den nedadgående udvikling, som er forsaget af finanskrisen i 2008, samt om den mere parametriserede VAR model vil give et mere præcis forecast i forhold til benchmark modellen.

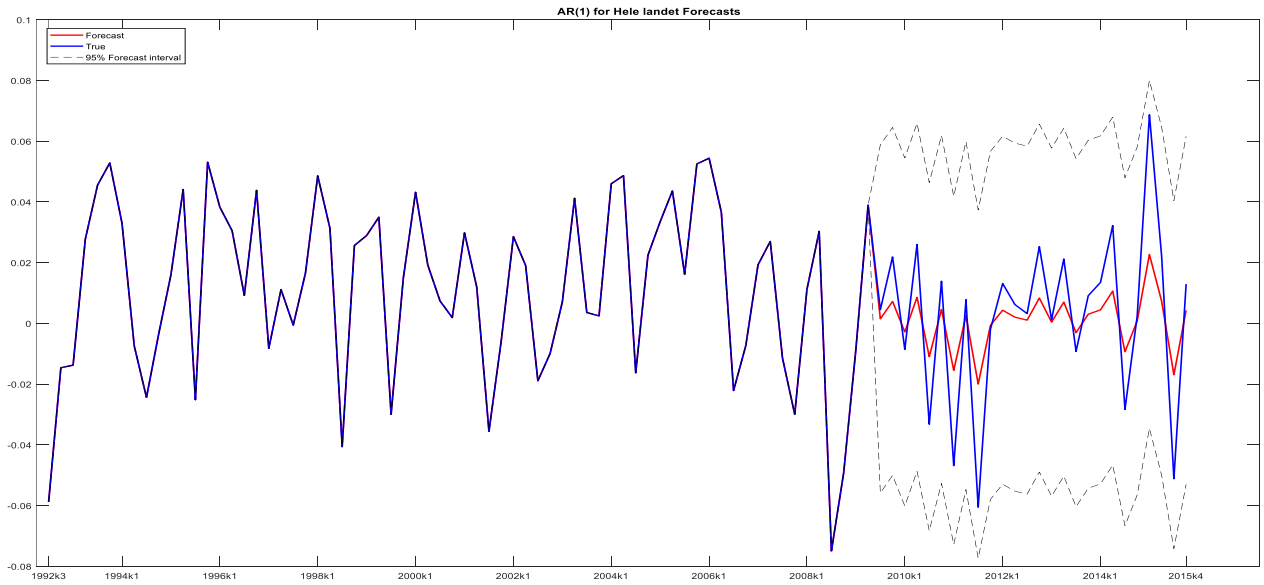
Benchmark modellen er stadig sat til at være en AR(1) model, som er estimeret på 68 observationer. BIC og AIC vælger stadig en vektor autoregressiv model med to lags, for det reducerede datasæt på 68 observationer. Vi tager udgangspunkt i forecast estimaterne for hele Danmark, hvilket vil blive grafisk illustreret i de to nedenstående figurer.

Figur (13.1): VAR(2) plot af forecast og observerede værdier



Figur (13.1) er der vist med den blå graf, de observerede ejendomspriserne for enfamiliehuse. Den røde graf angiver de forecastede værdier for udviklingen i ejendomspriserne. Det kan ses i figur (13.1) at forecastene lige efter finanskrisen for huspriserne følger den nedadgående ændring i de faktiske husprisændringer. Omkring 2011 sker der dog et radikalt skift i forecast værdierne, de stiger i stedet for at falde. Forecast residualerne kan allerede nu antages at være høje efter 2011, hvilket vil påvirke den overordnede præcision af VAR(2) modellen. Ved indgangen til 2013 retter forecastene sig ind igen, og følger de observerede værdier. Dette foregår dog kun indtil udgangen af 2014, hvor der igen sker en stigning i de forecastede værdier i forhold til de observerede. Henimod slutningen af tidsperioden fanger VAR(2) modellen ændringerne i de observerede værdier igen.

Figur (13.2): AR(1) plot af forecast og observerede værdier



I figur (13.2) er de observerede værdier plottet, samt de estimerede forecast værdier for benchmark AR(1) modellen. AR(1) modellens forecasts for enfamiliehuse følger det observerede data pænt i hele perioden. Det observerede data er meget volatilt i perioden 2010 til 2012, og det kan ses at AR(1) modellen fanger udviklingen godt, den overprisfastsætter dog priserne. I starten af 2014 fanger AR(1) modellen stadig de rigtige ændringer i ejendomspriserne, det kan dog ses, at der stadig er en stor forskel mellem de observerede værdier og de estimerede, hvilket vil forårsage høje residualer.

Sammenlignes VAR(2) modellen med vores benchmark AR(1) model, kan det ses, at AR(1) modellen bedre fanger ændringerne i perioden 2011 til 2013. Henimod slutningen af 2014 begynder VAR(2) modellen at fange ændringerne i ejendomspriserne bedre end AR(1) modellen. Det kan derfor ikke grafisk konkluderes, at den ene model forecaster bedre end den anden. Derfor er det nødvendigt at benyttet *mean square prediction error,* således at der bedre kan dannes et overblik, om VAR(2) eller benchmark modellen bedst forecaster udviklingen i ejendomspriserne for hele Danmark.

Mean square prediction error er gennemgået i afsnit (12) *evaluering af forecast*. *MSPE* anvendes igen til at kontrollere hvilken af de to estimerede modeller, der har de laveste residualer. Den ovenstående grafiske gennemgang viste, at de to modeller havde forskellige perioder, hvor den ene forecastede bedre end den anden. *MSPE* udregnes derfor igen for de nye langsigtede forecast:

$$MSPE = \frac{1}{H}\sum_{i=1}^{H} e_{1i}^2$$

Tabel (13.1): MSPE for VAR og AR hele Danmark

MSPE	VAR(2)	AR(2)
Hele Danmark	0,000834	0,000344

Det kan ses i tabel (13.1) at *MSPE* værdien for benchmark modellen er mindre end *MSPE* værdien for VAR(2) modellen. Sammenlignes tabel (12.1) fra afsnittet evaluering af forecast, ændres konklusion om at VAR(2) modellen for hele Danmark var den bedste, til at benchmark AR(1) modellen for hele Danmark er den bedste. *MSPE* for de danske regioner er illustreret i nedenstående tabel (13.2).

Tabel (13.2): MSPE for VAR og AR for de danske regioner.

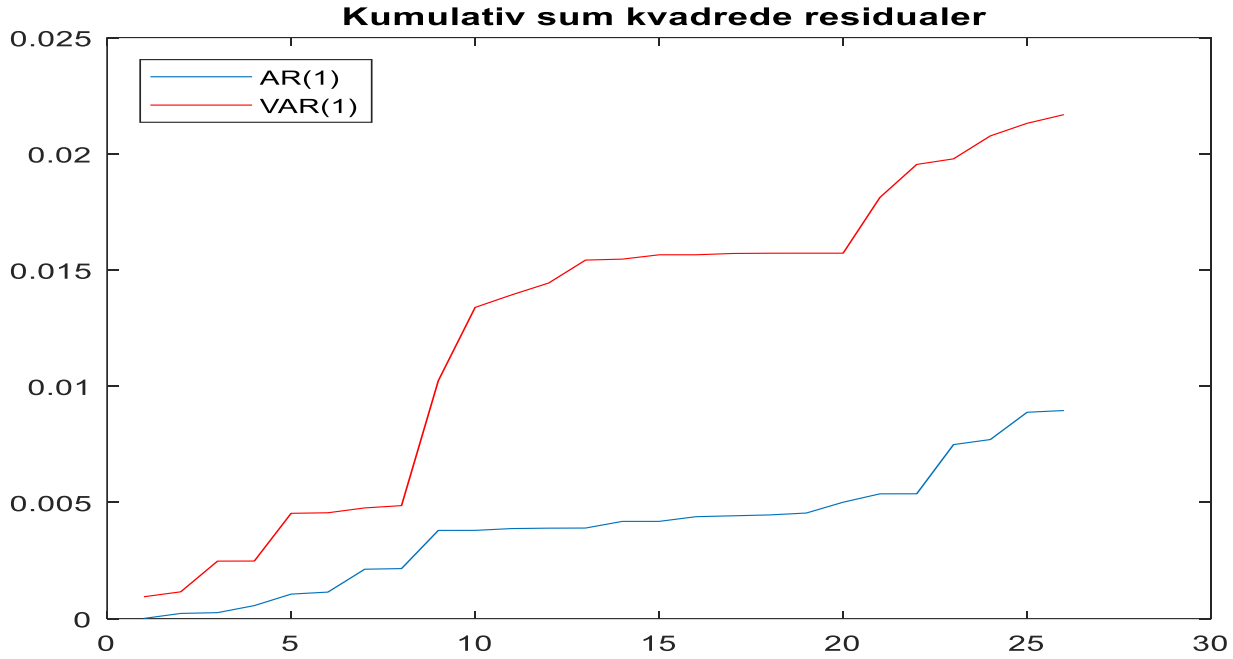
MSPE	VAR(2)	AR(1)
Hovedstaden	0,001069	0,000486
Midtjylland	0,000948	0,000522
Nordjylland	0,002004	0,001393
Sjælland	0,001107	0,000855
Sydjylland	0,001129	0,000594

AR(1) modellen for hele Danmark er nu den bedste til at forecaste for en lang tidsperiode. I tabel (12.1) fra evaluering af forecast afsnittet blev det konkluderet, at benchmark modellen var bedre til at forecaste. Det samme er tilfældet for den længere forecast periode på 26 *steps ahead*, hvilket kan ses i tabel (13.2), da gennemsnittet for de kvadrerede residualer for AR(1) modellerne for alle regionerne er lavere end for VAR(2) modellen.

Goyal og Welch

Alternativt til *MSPE* undersøgelsen af hvilken model der er bedst, kan en kumulativ grafisk illustration anvendes. I nedenstående figur (13.3) er den kumulative sum af kvadrede residualer plottet. Det kan udledes fra de første 8 *steps ahead* forecasts, at begge modeller følger hinanden pænt, dog er residualerne for AR(1) modellen mindre end for VAR(2). Efter de 8 første forecasts kan det ses, at den radikale ændring fra figur (13.1) indtræder. De kumulative kvadrede residualer kan videre udvides, ved at bruge Goyal og Welch artikel fra 2008.

Figur (13.3): Kumulativ sum af kvadrede residualer



Goyal og Welchs artikel fra 2008 undersøger variabler, der er blevet forslået af akademisk litteratur for at have en god performance i forecast analyse[47]. I deres analyse estimerer de flere forskellige modeller. Herefter tager de modellernes *squared residuals* for forecastene og akkumulerer dem over tid, således at de står tilbage med de kumulative kvadrede residualer for hver enkelt estimeret model. Goyal og Welch (2008) gør herefter brug af de kumulative kvadrede residaler fra en historisk lineær regression, fratrækker dem fra deres oprindelige kumulative kvadrede residualer[48]. Goyal og Welchs kumulative forecast fejl findes ved anvende ligning (13.1):

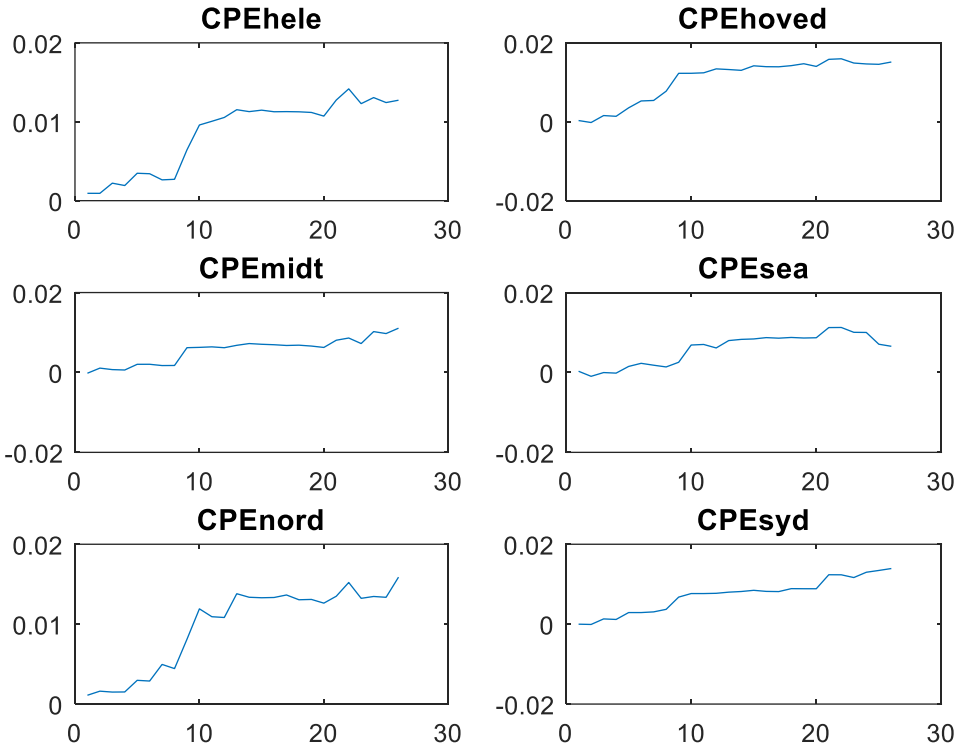
$$CPE = \sum_{i=1}^{I}\left((y_i - \hat{y}_i^a)^2 - \left(y_i - \hat{y}_i^b\right)^2\right) \qquad (13.1)$$

Goyal og Welch's fremgang måde til at evaluere deres forecast ved brug af den kumulative sum fordeling, vil overføres til benchmark modellen og VAR modellen for dette projekt. y_i angiver den observerede værdi for huspriserne, disse er tilgængelige pga. observationerne allerede er kendt.

$\hat{y}_i^a$ er i ligning (13.1) de estimerede forecasts for alternativ modellen, hvilket vil sige VAR(2) modellen. $\hat{y}_i^b$ er i ligning (13.1) de estimerede forecasts for benchmark modellen, hvilket refererer til AR(1) modellen.

[47] *A Comprehensive Look at The Empirical Performance of Equity Premium Prediction,* Welch, Ivo og Amit Goyal, *side 1465*
[48] *A Comprehensive Look at The Empirical Performance of Equity Premium Prediction,* Welch, Ivo og Amit Goyal, *side 1465*

Figur (13.4): Goyal and Welch CPE



I figur (13.4) er Goyal og Welchs princip påført. AR(1) modellen som er anvendt til benchmark er fratrukket alternativ modellen VAR(2). Der kan derfor udledes, hvor meget VAR(2) modellens residualer ændrer sig i forhold til AR(1) modellen. Vi tager udgangspunkt i CPE for hele Danmark. I figur (13.1) og figur (13.2) så det ud som om, at VAR(2) modellen var bedre til at forecaste de første 8 steps ahead forecast, men det viser sig ikke at være sandt, da figur (13.4) tydeligt illustrerer at residualerne ligger over benchmark modellen. Regionerne Midt-, og Sydjylland samt Region Sjælland har pæne residualer der ligger tæt på benchmark modellen. Det kan også ses, at modellerne for disse regioner ikke har det samme radikale skift i ejendomspriserne, som f.eks. VAR(2) modellen for hele Danmark lider af. Fra CPE i figur (13.4) kan det se ud som om, at der er en signifikant forskel mellem alternativ og benchmark modellen, i forhold til hvilken model der forecaster bedst. For at kunne analysere om dette er sandt, anvendes Diebold og Marianos test til at endelig kunne konkludere om AR(1) er bedre til at forecaste end VAR(2) modellen.

Diebold Mariano

I afsnittet om evaluering af forecast bliver Diebold og Marianos test anvendt til at analyser, hvorvidt der er en signifikant forskel mellem forecastene fra de estimerede AR(1) og VAR(2) modeller. Det samme princip bliver tilført i dette afsnit, hvor forecast perioden er sat op til 26 *steps ahead*. H_0-hypotesen for Diebold Mariano testen er, at der ikke er forskel i modellernes forecast præcision.

Tabel (13.3): DM-statistik

DM	t-value
Hele Danmark	2,0450
Hovedstaden	2,4140
Midtjylland	1,7650
Nordjylland	2,1327
Sjælland	0,9272
Sydjylland	2,7424

DM	Critical value[49]
0.05%	2,060
0.10%	1,708

I tabel (13.3) er t-værdierne for Diebold Mariano testen angivet, samt de tilhørende kritiske værdier ved 5 og 10 procents signifikantniveau. DM testen udregner t-værdien for hele Danmark til at være 2,0450, hvor den tilhørende kritiske værdi er 2,060 ved et 5 procents signifikantniveau. Det er ikke muligt at forkaste H_0-hypotesen at den ene model skulle være mere præcis end den anden. Ved et 10 procents signifikantniveau er den kritiske værdi 1,708, hvilket gør det muligt at forkaste H_0-hypotesen. Benchmark modellen er derfor bedre til forecast end VAR(2).

Den samme konklusion gør sig gældende ved Region- Hovedstaden, Nordjylland og Sydjylland. Her kan H_0-hypotesen forkastes ved et 5 procents signifikantniveau. Benchmark modellerne er signifikant mere præcise i deres forecasts for disse regioner.

Region Midtjyllands t-værdi er udregnet til 1,7650, og ved et 5 procents signifikantniveau kan H_0-hypotesen ikke forkastes, men ved et 10 procents signifikantniveau forkastes H_0-hypotesen. VAR(2) modellen for Midtjylland antages derfor, at have den samme præcision som benchmark modellen. Region Sjælland er den eneste model, der tydeligt ikke er signifikant dårligere i dens præcision end benchmark modellen. H_0-hypotesen om at begge modeller har samme præcision, kan hverken forkastes ved 1 procents signifikantniveau eller ved et 5 procents signifikantniveau.

[49] Values of the t-distribution. MEDCALC

I afsnittet (12) evaluering af forecast blev der fundet frem til, at både VAR(2) modellen og AR(1) modellen havde den samme præcision i forecasts af udviklingen i huspriserne, fordi DM-værdierne ikke kunne forkaste H_0-hypotesen for hele Danmark samt regionerne. Det var nødvendigt at ændre forecast perioden fra 14 steps til 26 steps for at tjekke robusthed i forecastene. Det observerede datasæt blev reduceret, således at observationerne blev afskåret lige efter finanskrisen var indtruffet.
Sammenligningen mellem de to forecast perioder viste tydeligt, at VAR(2) modellen for hele Danmark ikke var i stand til at fange den negative udvikling i huspriserne. Især efter 8 forecast perioder blev forecast præcisionen forringet. I figur (13.4) med de plottede Goyal og Welch kumulative sum af residualerne for hele Danmark var det tydeligt at se, at forecastene efter periode 8 blev radikalt forringet. Den eneste VAR model der stadig er robust i forecastene efter justeringen af forecast intervallet, er VAR(2) modellen for Region Sjælland. VAR modellen for Region Sjælland kunne både ved 14 steps og 26 steps måle sig med benchmark modellen, da H_0-hypotesen i begge Diebold Mariano tests ikke kunne forkastes.

14 Analyse af de udvalgte forecast modeller

Dette analyse afsnit indeholder en gennemgang af modellerne, der anvendes til at besvare problemformuleringen, som er blevet nøje udvalgt. Modellerne vil blive gennemgået, og plots af forecastene for de tilhørende modeller vil blive illustreret. De grafiske plots anvendes til, at analysere på hvordan forventningen til ændringerne i huspriserne i hele Danmark og regionerne kommer til at bevæge sig i forecast perioden. Til sidst vil der komme en diskussion, som tager udgangspunkt i informationen fra forecastene og holder denne information op mod hvad der virkelig er sket. Ligeledes vil der komme en fortolkning af de konsekvenser husprisstigninger kan føre med sig. Det vil også blive diskuteret, hvorfor projektets VAR modeller har givet så dårlige forecasts.

Afsnittet (12) *Evaluering af forecast* starter med at beskrive udviklingen i de estimerede forecasts, hvor forecastene er blevet estimeret med et 14 *steps ahead* forecasts interval. Fra figur (12.1), der illustrer VAR(2) for hele Danmark og figur (12.2), som illustrer benchmark AR(1) modellen for hele Danmark, bliver det udledt at begge modellers forecast ser fornuftige ud. Dog er VAR(2) den præcisions dominerende model.
VAR(2) modellens dominans bliver endvidere understøttet af *mean squared prediction error. MSPE* for VAR(2) modellen for hele Danmark er blevet udregnet til 0,000425, hvilket er mindre end benchmark modellens *MSPE* på 0,000452. De regionale VAR modellers MSPE er ringere end de tilhørende AR modellers MSPE. Regionernes Benchmark- og VAR modellernes forecast ligger sig relativt fornuftigt op af de observerede, hvilket besværliggøre det at konkludere om en af modellerne er signifikant bedre end den

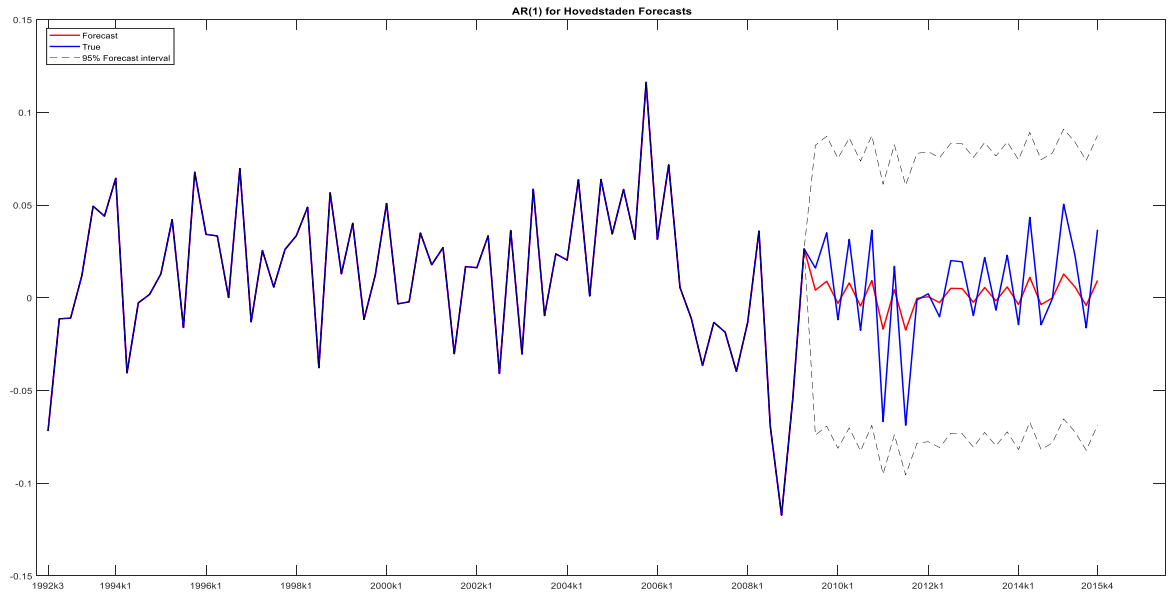
anden.
Diebold og Mariano har udviklet en teori, der gør det muligt at teste, om hvorvidt der er en signifikant forskel mellem to forecasts fra to forskellige modeller. Teorien bliver anvendt i forecast evaluerings afsnittet (12) og selve testen er illustreret i tabel (12.2). DM-test værdierne bliver forkastet ved et 5 procents signifikantniveau, med den tilhørende kritiske værdi 2,160. Det er ikke muligt at forkaste H_0-hypotesen for DM testene i tabel (12.2). Da H_0-hypotesen ikke kan forkastes i DM-testen, betyder det at VAR- og benchmark modellerne begge har samme præcision i deres forecast. Da der ikke kan drages en konklusion fra de første forecast, gøres der brug af et nyt forecast interval på 26. De nye forecasts estimeres for at undersøge modellernes robusthed, når det observerede data har en abnorm udvikling. I afsnittet Robustheden af forecast udledes det, at den eneste model, der ikke er signifikant dårligere end benchmark modellen, er VAR(2) modellen for Region Sjælland. Region Sjællands DM-test har en t-værdi på 0,9272, og kan derfor ikke forkaste H_0-hypotesen ved 5 procents signifikantniveau med tilhørende kritiskværdi 2,060, og ved 10 procents signifikantniveau med tilhørende kritiskværdi 1,708.

Forecast evaluerings afsnittet og Robustheden af forecast afsnittet kan nu anvendes til at udlede de mest optimale modeller til brug af forecasting af ændring i ejendomspriserne i alle regionerne.
Modellen for hele Danmark, der bliver anvendt til forecast, bliver sat til at være en AR(1) model. Dette er gjort på baggrund af, at VAR(2) modellen var signifikant dårligere end AR(1) modellen i dens forecast præcision. Regionerne Nordjylland, Midtjylland, Sydjylland og Hovedstaden havde samme konklusion som den for hele Danmark. I Robustheden af forecast afsnittet blev det udledt af VAR(2) modellerne for disse regioner var signifikant dårligere i deres præcision end benchmark modellerne, så derfor anvendes benchmark AR(1) modellen til forecast. Den eneste VAR(2) model, der bliver anvendt i analysen, er den for Region Sjælland. VAR(2) modellen for Region Sjælland er taget med, da det ikke var mulig at forkaste H_0-hypotesen i DM testen. VAR(2) modellen kan derfor antages at have samme præcision i forecastene som benchmark modellen.

De grafiske illustrationer af forecast samt de tilhørende observerede værdier bliver gennemgået i nedenstående. AR(1) modellen for hele Danmark er vist i Robustheden af forecast afsnittet, og kan derfor ses i figur (13.2). Benchmark AR(1) modellen for hele Danmark har problemer med at fange de ekstreme ændringer i huspriserne, der forekommer lige efter finanskrisen. Modellen fanger dog, at der vil være en negativ ændring i huspriserne i de kommende år efter finanskrisen. 4 år efter finanskrisen forecaster benchmark modellen, at huspriserne vil se en positiv ændring igen, dog langt fra på samme niveau der kunne observeres før finanskrisen. Henimod slutningen af forecast perioden 2015k1, forventes det at huspriserne vil se den højeste positive ændring, hvilket også har været tilfældet, hvis forecastet

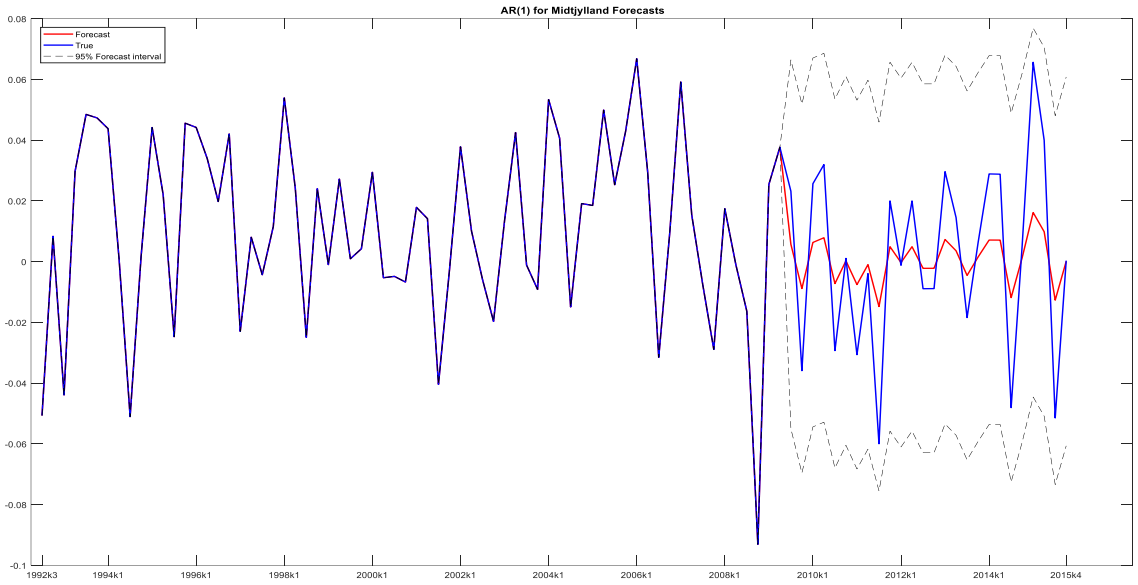
sammenlignes med den observerede ændring i huspriserne for denne periode. Den faktiske ændring har været større end forecastet. Forecast modellerne for regionerne vil nu blive analyseret på om hvorvidt de har samme forecasts evne, som forecastene for hele Danmark.

Figur (14.1): AR(1) Forecast for Region Hovedstaden



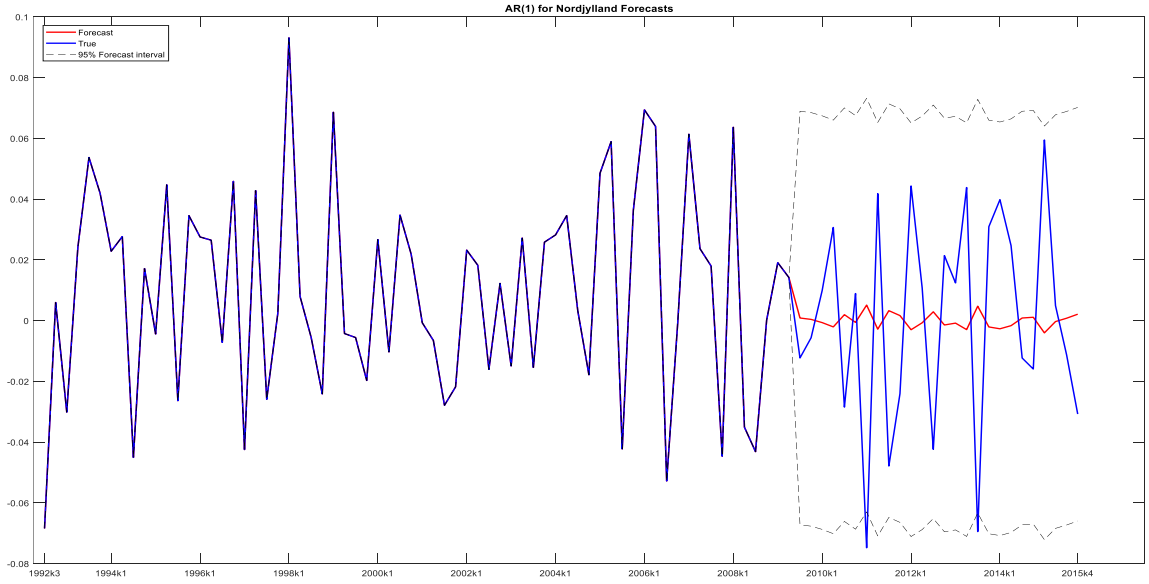
Forecastene for Region Hovedstaden kan ses i figur (14.1) samt de tilhørende observerede værdier. Enfamiliehuse i Region Hovedstaden har set ekstrem negative ændringer i priserne på tidspunktet omkring finanskrisen. AR(1) modellens forecaster lige efter finanskrisen i 2008. AR(1) modellen fortæller os at de fremtidige huspriser for enfamiliehuse i Region Hovedstaden, vil se en stagnation i de første 4 år efter finanskrisen. Hvis de første forecasts sammenlignes med de faktiske værdier, kan det ses at forecast modellen fanger den rigtige udvikling i enfamiliehuspriserne. Huspriserne falder dog mere end modellen forventer, de skal gøre. Udviklingen, 4 år efter finanskrisen, forventes det ud fra forecastene at huspriserne vil begynde at se positive ændringer. I start 2015 til slut 2015 forecaster AR(1) modellen at ejendomspriserne vil se høje ændringer. De forecastede ændringer for denne periode, kan ses at være sande, hvis forecastene sammenholdes med de faktiske observerede værdier. Dog er AR(1) modellen ikke helt i stand til at fange det høje ændring i enfamiliehuspriserne.

Figur (14.2): AR(1) Forecast for Region Midtjylland



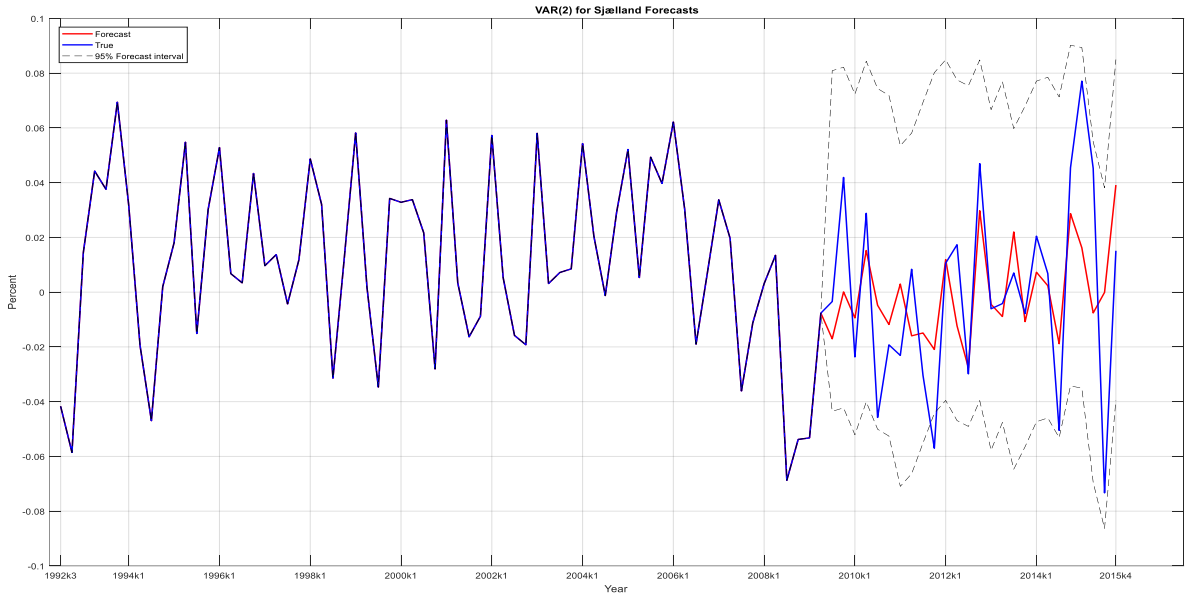
Figur (14.2) illustrer udviklingen af huspriserne for enfamiliehus, samt forecast for de sidste 26 observationer. AR(1) modellen for Region Midtjylland forecaster de korrekte ændringer, dog sammenlignes forecastene med de faktiske værdier, er der stadig en relativ stor forskel mellem de to. Huspriserne for enfamiliehus forventes at have en negativ ændring i perioden efter finanskrisen. Ses der på de forecastede værdier, kan det ses, at dette er sandt. Der forekommer dog en positiv ændring i priserne i år 2010. Perioden fra omkring 2015 til slutningen af forecast perioden, forecastes relativ små positive ændringer i huspriserne. Ultimo 2015 kan det ses at huspriserne forecastes til at have en negativ ændring, hvilket også gøre sig gældende for de faktiske observerede værdier.

Figur (14.3): AR(1) Forecast for Region Nordjylland



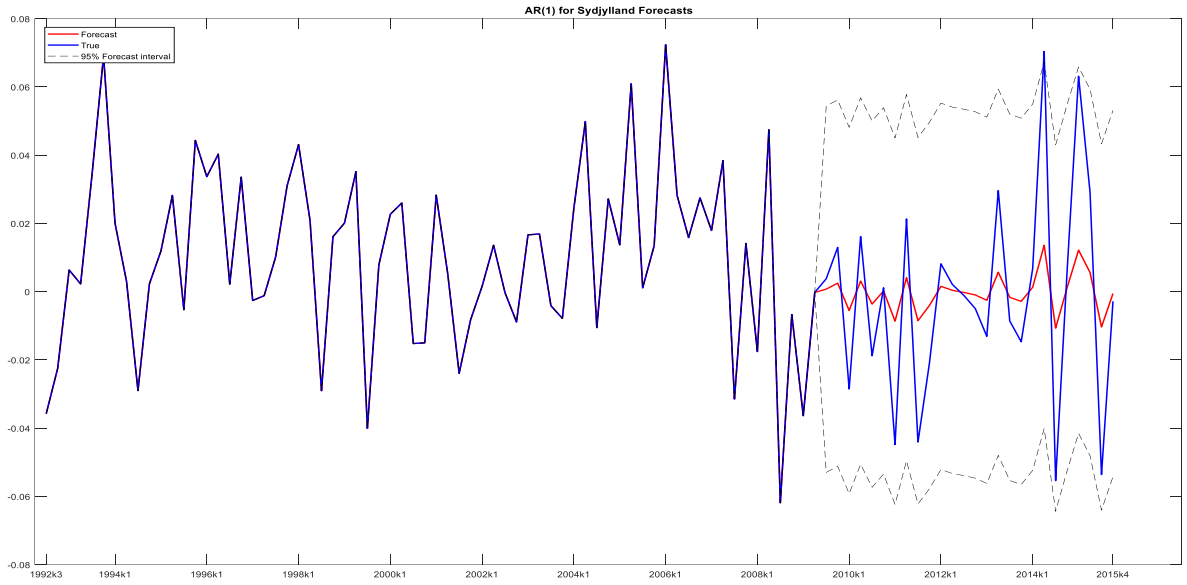
AR(1) modellen for Region Nordjylland har meget monotone ændringer i ejendomsprisernes forecast. Modellen er ikke i stand til at fange de observerede værdier. Huspriserne i Region Nordjylland har ikke se store negative ændringer i husprisudviklingen. Dermed er det dog ikke sagt, at der ikke kan ses en nedafgående ændring i priserne lige efter finanskrisen. Både forecast og de faktiske værdier ligger og svinger lige omkring nul.

Figur (14.4): VAR(2) Forecast for Region Sjælland



Forecast modellen, der er blevet udvalgt til at forecaste husprisændringerne for Region Sjælland, er en VAR(2) model. Det observerede data samt forecast fra VAR(2) modellen er plottet i figur (14.4). Enfamiliehuspriserne i Region Sjælland har set store negative ændringer lige efter finanskrisen. Forecastene fra regionsmodellen estimerer ændringerne de første par år efter finanskrisen, til at være hovedsaligt negative. Efter 2012k1 sker der en negativ ændring, hvorefter forecast værdierne forventer at huspriserne vil se en stigning. Henimod slutningen af forecast perioden, forecaster VAR(2) modellen at huspriserne vil se en høj positiv ændring. Sammenlignes de sidste forecast med de faktiske observerede værdier, kan det ses at de mere eller mindre følger hinanden. Dog for den sidste periode 2015k4 forecaster modellen en højere ændring end den faktiske.

Figur (14.5): AR(1) Forecast for Region Sydjylland



Region Sydjylland anvender en AR(1) model til at forecaste. De observerede værdier samt forecastene er plottet i figur (14.5). Forecastene er meget monotone, og svinger kun lige omkring nul. De negative ændringer, som er forsaget af finanskrisen, er kun lige fanget af AR(1) modellen. De store udsving, der kan ses i figur (14.5) for det observerede data, er langt fra forecastet af AR(1) modellen. Mod slutningen af forecast perioden angiver de forecastede værdier, at der vil komme en positiv ændring i ejendomspriserne for enfamiliehuse. Det sidste forecast i 2015k4 fortæller, at huspriserne vil begynde at se positive ændringer. Hvis det sidste forecast sammenlignes med den faktiske observerede værdi, kan det ses at enfamiliehuse begynder at få positive prisstigninger.

Forecast modellerne for hele Danmark og de tilhørende regioner, har ikke helt fanget de faktiske udsving i huspriserne. Forecastene fanger dog den rigtige udvikling, der sker i ejendomspriserne for enfamiliehuse. Hvis der kun fokuseres på regionerne, kan det ses at udviklingen i huspriserne i Region Hovedstaden har lidt mest efter finanskrisen. Dog er det også i Region Hovedstaden at de største prisstigninger forecastes. Det samme gøre sig gældende for Region Sjælland. Regionen har ikke set helt så store positive ændringer i huspriserne som Region Hovedstaden, men de forecastede værdier angiver, at henimod slutningen af observationsperioden vil der forekomme ejendomsprisstigninger. Dette viste sig også at være sandt, når de forecastede værdier bliver sammenholdt med de faktiske observerede værdier.
Region Nordjylland er den region, der har været mindst påvirket af finanskrisen, samtidig kan det også udledes fra forecastene, at der heller ikke vil være store ændringer i ejendomspriserne for enfamiliehuse. Hvis de forecastede værdier sammenholdes med de observerede for Region Nordjylland, kan de ses at der har været udsving i huspriserne. Ændringerne er dog langt fra ligeså store, som det er observerede for Region Hovedstaden og Region Sjælland.

15 Konklusion

I projektet blev den følgende problemformulering, belyst og besvaret:

Er det muligt ved hjælp af en vektor autoregressiv model, at generere forecasts for udviklingen i de danske boligpriser, både på landsplan og regionalt, som er bedre end en simpel parametersparsom autoregressiv model?

For at besvare denne problemformulering, blev der estimeret 6 AR(1) modeller og 6 VAR(2) modellerne. Af modellerne, blev der indledningsvist genereret *14 out of sample one step ahead* forecasts. Her blev der ved hjælp af Diebold Mariano testen, udledt at der ikke var en signifikant forskel mellem modellernes forecast præcision. Modellernes forecast blev derfor antaget at have ens præcision. For at undersøge om dette er sandt, eller blot et tilfælde blev forecast perioden udvidet til at inkludere 26 *steps ahead*. Robustheds analysen blev fortaget i afsnittet robustheden af forecast. Det blev fra afsnittet konkluderet, at det kun var tilfældigt at VAR(2) modellerne og AR(1) modellerne med *14 out of sample one step ahead* forecasts havde samme præcision. Den eneste VAR(2) model, som ikke var signifikant dårligere end benchmark modellen, var VAR(2) modellen for Region Sjælland. Det kan herfra konkluderes at de vektor autoregressive modeller har svært ved at måle sig med en simpel autoregressive model, der anvender ét lag. Det var kun en ud af de 6 modeller der blev estimeret, der som sagt ikke var signifikant dårligere end benchmark modellerne. Benchmark modellerne for hele Danmark og regionerne Hovedstaden, Nordjylland, Midtjylland og Sydjylland samt VAR modellen for Region Sjælland blev anvendt at analyserer på forecast udviklingen af ejendomspriserne. Fra analysen kan det konkluderes, at der fra de forecastede observationer kan ses en forskel mellem huspriserne i regionerne. AR(1) modellen for Region Hovedstaden forecaster, at der vil komme stigninger i ejendomspriserne henimod slutningen af forecast perioden. Sammenlignes dette resultat med AR(1) modellen for Region Nordjylland, kan det tydeligt ses at Hovedstaden oplever meget højere positive ændringer i huspriserne end Region Nordjylland. Region Nordjylland er den region, der er den mindste ejendomspris ændring i. Det kan også konkluderes fra afsnittet Analyse af de udvalgte forecast modeller, at der vil komme en stigning i de danske huspriser, hvor det hovedsageligt er Region Hovedstaden og Region Sjælland der trækker denne stigning. De forecastede stigninger i huspriserne, kan have konsekvenser for den fremtidige danske økonomi, da det kan være muligt, at dette er begyndelsen på en ny overophedning af det danske boligmarked.

16 Appendiks

Dette afsnit kommer til at omhandle den kodning, der ligger bag projektets udregninger. Her vil der ikke i blive forklaret i dybden, hvad de enkelte koder gør, men der vil derimod komme en gennemgang af hvordan koderne forløber. Resultaterne, som der er kommet af kodningen, vil ikke vist, da de relevante resultater allerede kan findes i projektet.

R-studio

Som forklaret tidligere under afsnittet *Interpolering*, benyttes der programmet *R-studio* til at udføre interpolering af data, hvor der mangler observerede datapunkter. Filer og koder, som er blevet benyttet til denne interpolering, kan findes i mappen *Interpoleringsbilag*. I mappen er der 3 *R*-filer med navnene *indkomst, skat* og *indeks*, og i hver af disse foregår kodningen til interpoleringen af den enkelte variabel, som filnavnet refererer til. Den første linje kode, som der køres ved hver interpolering, er dataindlæsning med kommandoen *read.csv2*. Her skal der gøres opmærksom på, at denne kode referer til en fil, der er gemt lokalt på computeren, og hvis ønskes der at eftertjekke interpoleringen på en anden computer, er det nødvendigt at ændre dette, således at der bliver refereret til den nye filplacering. Herefter indlæses pakken "imputeTS", der benyttes til at udføre den faktiske interpolering. Hvis denne pakken ikke er installeret, er det nødvendigt at installere den ved eftertjek.

Da der under variablerne *skat* og *indkomst* er 6 tidsrækker, bliver der gennemført interpolering på hver af de enkelte tidsrækker, ved hjælp af *na.interpolation.* Her får hver tidsrække et variabel-navn, der refererer til den fremtidige model, som dataet vil indgå i. For eksempel refererer variabel-navnet *hele* under koderne i *skat.r* til den tidsrække, der vil blive brugt for variablen *skat* i VAR modellen for husprisudviklingen i hele Danmark. Ligeledes anvendes variabel-navnet *midt* under koderne i *indkomst.r* til den tidsrække, der vil blive brugt for variablen *indkomst* i VAR modellen for husprisudviklingen i Region Midtjylland. Da der kun er behov for én tidsrække af indekset til deflateringen, vil der for indekset kun blive dannet en tidsrække i koderne i *Indeks.r.* Alle variablerne i de enkelte r-filer bliver sat sammen til en tabel kaldet *fuld* ved hjælp af kommandoen *cbind*, og bliver herefter gemt som en csv-fil med kommandoen *write.csv*.

Matlab

I Matlab foregår den største del af kodningsarbejdet. I mappen *Matlab* findes der 4 Matlab-filer, som indeholder koderne, der er blevet benyttet til beregningerne i projektet. Den første Matlab-fil, der benyttes, er filen *ADFtest.m*. I denne fil er der blevet udført test for stationaritet i dataet, som der er blevet gennemgået under afsnittet *Dickey-Fuller*. Herefter kommer filen *Forecast1.m*, som indeholder koderne til generering af VAR modellerne, AR modellerne og 14 steps ahead forecastene heraf. I denne fil er der

ligeledes koderne til evaluering af forecastene, som der er gennemgået i afsnittet *evaluering af forecast*. Den næste fil er *Forecast2.m*, som indeholder næsten identiske koderne, som i *Forecast1.m*. Her er dog den forskel, at der her blevet dannet 26 steps ahead forecasts, på baggrund af modeller med et mindre antal observationer. Den sidste fil, der benyttes, er filen *dmtest.m,* hvilket indeholder funktionen *dmtest*. Endvidere kan det data, som der benyttes i projektet, findes i mappen *data*. Nedenfor vil koderne i de enkelte Matlab-filer blive gennemgået.

ADFtest.m

I afsnittet hvor data bliver undersøgt for stationaritet, anvendes teorien fra Dickey og Fuller gennemgået i afsnit (8). Udregningen af de tre ligninger AR-(8.5), ARD-(8.6) og TS-(8.7) bliver fortaget i Matlab scripted *ADFtest*.m.

I den første sektion af Matlab koden, efter at data er blevet indlæst, anvendes et loop med funktionen *adftest.* Funktion tillader at estimerer de tre forskellige ligninger Dickey og Fuller, det antages at data vil følge. Fra regressionen udtrækkes de nødvendige SSR værdier for at kunne estimere de enkelte ϕ_i værdier, dette bliver gjort for alle regionerne og hele Danmark. Modellen, der er blevet udvalgt igennem ϕ_i værdierne, bliver anvendt i et ny loop hvor den optimale lag-længde for at den passende model kan bestemmes. Lag-længden der bruges som initial, skal selv justeres. For huspriserne er initial lag-længden sat til 6. Fra estimeringen udtrækkes p-værdien for alle de estimerede koefficienter, således det sidste lag kan analyseres for om værende signifikant eller ej. De udvalgte lag-længder for hele Danmark og regionerne bruges til Dickey og Fullers test for stationaritet. T-værdierne trækkes ud og gemmes i vektoren *tvalue[variable]*, samt de tilhørende kritiske værdier, som bliver gemt i vektoren *cvalue[variable]*. De næste sektioner gør nøjagtig det samme, bare for de resterende variabler.

Forecast1.m

Forecasts1.m er opdelt i en række af sektioner. I den første sektion ligges der ud med indlæsning af dataet fra mappen *data*, hvilket bliver gjort ved hjælp kommandoen *xlsread.* Her bliver renten divideret med hundrede for at få den i procent. Herefter følger databehandlingen af de enkelte variable, som der er blevet gennemgået i afsnittet *Databeskrivelse*. Her benyttes kommandoen *diff* til at finde differensen af variablerne, men kommandoen *log* benyttes til at finde logaritmen til variablerne.

I den næste sektion sker dannelse af VAR- og AR modellerne for udviklingen i huspriserne i hele Danmark. Da variablerne for indkomst, skat og huspriser indeholder tidsrækker til alle 6 VAR modeller, ligges der ud med at definere hvilke tidrækker, der skal med i VAR modellen for hele Danmark. Disse bliver samlet i en matrix, kaldet *yhele*, hvor efter der bliver frataget de sidste 14 observationer i matrixen, og dannet en ny

matrix kaldet *Yhele2*. På baggrund af data i *Yhele2*, dannes der i et loop 7 VAR modeller med 1 til 7 lags. VAR modeller dannes ved hjælp af kommandoen *varm*. Herefter anvendes *AIC* og *BIC* på begge modeller.

Som beskrevet tidligere for VAR modellerne, viser AIC og BIC at den bedste model har 2 lags. Derfor dannes der igen en model*, mdlhele*, ved hjælp af *varm*, hvoraf modellen *hele* estimeres ved hjælp af funktion *estimate*. Her benyttes *summarize* for at se estimaterne af *Hele*.

Da forecastene er *one step ahead* forecasts er det blevet valgt at danne disse forecasts i et loop. Først dannes der en vektor, *B,* som består af talrække fra 1 til 14. Herefter dannes forecastene ved brug af funktionen *forecast*, hvor der benyttes dataobservationerne fra 1 til 80+*i* fra *Yhele.* Her følger *i* vektoren *B*, således der estimeres på baggrund af alt det tidligere data op til tidspunkt 80+*i*. De forecastede værdier defineres midlertidigt som *p*, inden den næste omgang i loopet overskriver *p*. I loopet gemmes specifik de forecastede værdier for udviklingen i huspriserne i variablen *N,* mens alle værdier af *p* gemmes i variablen *M*. Ligeledes gemmes *prediction mean square error* i variablen *PMSE.*

Herefter benyttes værdier fra *M* og *PMSE* til at danne et 95 procents konfidensinterval. Da *M* i loopet er blevet sammensat som en vektor, indledes der her med at omdanne denne til en 4x14 matrix, kaldet *Y*. Her dannes funktionen *exctractMSE,* som er en funktion af *diag*, der danner en matrix, med de givne værdier diagonalt i matrixen. Denne funktion sættes ind i en *cellfun*-funktion, for at påfører funktion på værdierne fundet i *M,* og derved findes de kvadrerede residualer. Herefter findes standardfejlene, *SE*, ved at tage de kvadrede residualer fra *MSE*. Her dannes 2 variabler med navnene *interval1* og *interval2*, ved at tage 1,96 gange standardfejlene, *SE,* og pålægge dem matrixen *Y.* Her er *interval1* den nedre del af konfidensintervallet:

$$interval1 = Y - 1{,}96 * SE$$

Og *interval2* er den øvre del af konfidensintervallet:

$$interval2 = Y + 1{,}96 * SE$$

Da det kun er udviklingen i huspriserne, som er vigtige for projektet, tages der kun den første kolonne i *interval1* og *interval2* med i plottet.

For at plotte forecastene sammen med de observerede værdier fra *Yhele,* dannes der 2 variabler, *y1* og *y2*. Her er *y1* alle de observerede værdier for udviklingen i huspriserne, og *y2* er kun de observerede værdier for udviklingen i huspriserne frem til observation nummer 80. Til *y2* tillægges de forecastede værdier fra *N*, som blev dannet i det tidligere loop, hvilket danner variablen *x*. Herudover dannes 2 nye variabler, *h1* og

h2, som er *y2* tillagt *interval1* og *interval2*, således VAR modellens konfidensinterval ligeledes plottes. De resterende koder er ene og alene tilegnet grafiske aspekter i plottet.

Herefter findes residualerne til evalueringen af forecastene. Dette gøres ved hjælp af et loop, hvor der igen benyttes vektoren *B.* Her følger *i* vektoren *B*, og den *i*'te estimerede værdi af forecastene, *N,* bliver fratrukket den observerede værdi for udviklingen i huspriserne ved observation nummer 80+i. Derved findes differensen i mellem de forecastede værdier og de observerede værdier for udviklingen i huspriserne. Residualerne gemmes midlertidig som *e,* og inden de bliver overskrevet, gemmes de i variablen *e1*.

Efter nu at have estimeret VAR modellen, dannet forecasts og gemt residualerne til forecast evalueringen, bevæges der videre til AR(1) modellen for udviklingen i huspriserne for hele Danmark. Her kunne det have været relevant at danne en ny sektion, men da det var ønsket at beholde VAR modellen og AR modellen for hele Danmark samlet, blev dette fravalgt.

Til dannelse af AR modellen benyttes funktionen *ar*, hvorved modellen *m* fremkommer. Her benyttes igen dataet *huspriserhele,* som blev dannet til VAR modellen. Da de forecastede værdier i AR modellen *one step ahead forecasts* er det blevet valgt at danne disse forecasts i et loop. Her benyttes igen vektoren *B.* Herefter dannes forecastene ved brug af funktionen *forecast*, hvor der benyttes dataobservationerne fra 1 til 80+*i* fra *Yhele,* hvor *i* følger vektoren *B*, således der estimeres på baggrund af alt det tidligere data op til tidspunkt 80+*i*. Hertil benyttes modellen *m*. De forecastede værdier defineres midlertidigt som *p*, og bliver fast gemt som *N,* inden den næste omgang i loopet overskriver *p*. Her bliver ligeledes gemt *yf_sd*, som er modellen *m*'s standardfejl.

Standardfejlene benyttes herefter til finde konfidensintervallet for modellen *m*, ved at tillægge og fratrække dem de forecastede værdier for huspriserne, *N*, under variablen *ForecastInt.* Her dannes den nedre del af konfidensintervallet *interval1* ved at tage den første kolonne i *ForecastInt,* mens den øvre del, i*nterval2*, dannes ved at tage den anden kolonne i *ForecastInt.*

For at plotte AR modellen og dens forecasts dannes der 2 variabler, *y1* og *y2*. Her er *y1* igen alle de observerede værdier for udviklingen i huspriserne, og *y2* de observerede værdier for udviklingen i huspriserne frem til observation nummer 80. *y2* bliver tillagt forecastene fundet i det overstående loop, *N,* hvilket danner variablen *x*. Det samme sker ved dannes af *h1* og *h2*, som er *y2* tillagt *interval1* og *interval2*. De resterende koder er igen tilegnet grafiske aspekter i plottet.

Herefter findes der igen residualerne af AR modellen til evalueringen af forecastene. Dette gøres her ved hjælp af et lignende loop, til det loop hvor residualerne fra VAR modellen findes. Her følger *i* vektoren *B*, og

den *i*'te estimerede værdi af forecastene, *N,* bliver fratrukket den observerede værdi for udviklingen i huspriserne ved observation nummer 80+i. Residualerne gemmes midlertidig som *e,* og inden de bliver overskrevet, gemmes de i variablen *e2*.

Evaluering af forecast

Sektionen efter sektionen med dannelse af VAR modellen og AR modellen for hele Danmark indeholder evalueringen af forecastene, som dannes på baggrund af de fundne residualer *e1* og *e2*. Her lægges der ud med at plotte residualerne i et histogram, for at tjekke om de følger en normalfordeling. Herefter udregnes *mean square prediction error*, *MSPE,* som der gemmes under variablen *MSPEhele.* Der bliver ligeledes testet for autokorrelation i residualerne igennem en *Ljung-Box* test med funktion *lbqtest*, for at kunne afgøre om residualerne er hvid støj.

For at beregne Diebold-Mariano test-værdierne benyttes funktionen *dmtest*, som vil blive redegjort under *dmtest.m.* Her er det nødvendigt at sætte *h=1*, da funktionen kræver en initial værdi for alle variablerne for at starte.

Sektionen med dannelse af VAR- og AR modeller gentages 5 gange med baggrund i forskelligt data, således at der bliver dannet modeller og forecasts for hver af de fem regioner. Ligeledes gentages sektionen med forecast evaluering, så der også bliver dannet MSPE-værdier og DM-tests-størrelser for hver af de fem regioner.

I den sidste sektion sker der en opsamling af de kumulative kvadrerede summer af residualerne, kaldet *CPEalle*. Her dannes plot for alle de kumulative kvadrerede summer af residualerne.

Forecast2.m

I Matlab-filen forecast2.m benyttes der den samme metodologi, som der blev benyttet i Forecast1.m. Her er dog den ene forskel, i stedet for at benytte 80 observationer til at estimere VAR- og AR modeller, så benyttes i stedet de første 68 observationer. Således bliver der frigjort 26 observationer til at teste modellernes evne til at forecaste *out of sample*, frem for de 14 observationer der var tilrådelighed i *Forecast1*.m Dette gøres for at teste for robustheden i forecastene, og referer derfor til afsnittet Robustheden af forecast.

Dmtest.m

Matlab-filen *dmtest.m* indeholder funktionen, der benyttes til at udregne Diebold-Mariano testværdierne. Her gøres der opmærksom på, at denne funktion er skrevet af Semin Ibisevic[50]. De første linjer af koden indeholder fejlmeddelelser i tilfælde af at der er for få observationer, eller at vektorerne for dataet ikke har samme længde. Herefter bliver *T* defineret som antallet af residualer, der er med i testen, og *d* bliver defineret som værende differensen imellem residualerne for VAR modellen *e1* og residualerne for AR modellen *e2*. *h* tæller hvor mange *steps ahead* forecast der er. Her findes variansen af *d* ved hjælp af funktion ***var***, hvilket bliver defineret som *gamma0*. Herefter følger et loop, hvor i der tages højde for autokorrelation i residualerne, i tilfælde af at *h* er større end 1. Da der i dette projekt kun bliver benyttet *one step ahead* forecasts, vil *h* altid være lig 1, og derfor benyttes dette loop ikke. Derfor sættes variansen til at være *varD=gamma0*. Til sidst beregnes værdien for *DM,* ved at følge ligning 12.1 fra afsnittet om forecast evaluering:

$$DM = \overline{dMean}/\sqrt{varD/[T]}$$

[50] Diebold-Mariano Test Statistic, Mathworks.

Litteraturliste

Bøger

International Economics: Theory and Policy, 10th Edition. Krugman, Paul m.fl, Side 393-394. 10. udg. Pearson, 2013. (Bog)

MODELS WITH TREND. Enders, Walter : I: APPLIED ECONOMETRIC TIME SERIES. 4. udg. WILEY, 2014. side 181-255. (Afsnit i bog)

PROPERTIES OF FORECASTS. Enders, Walter: I: APPLIED ECONOMETRIC TIME SERIES. 4. udg. WILEY, 2014. side 80-87. (Afsnit i bog)

SAMPLE AUTOCORRELATIONS OF STATIONARY SERIES. Enders, Walter: I: APPLIED ECONOMETRIC TIME SERIES. 4. udg. WILEY, 2014. side 69-70. (Afsnit i bog)

SUPPLEMENTARY MANUAL TO ACCOMPANY. Enders, Walter : I: APPLIED ECONOMETRIC TIME SERIES. 4. udg. WILEY, 2014. side 120-120. Internetadresse: http://time-series.net/yahoo_site_admin/assets/docs/SupplementaryManual_all.117125921.pdf Besøgt d. 05.05.2017 (Afsnit i bog)

Internet:

Comparing Predictive Accuracy of two Forecasts: The Diebold-Mariano Test. Udgivet af Umberto Triacca. Internetadresse: http://www.phdeconomics.sssup.it/documents/Lesson19.pdf - Besøgt d. 3.5.2017 (Internet)

Cubic Spline Interpolation. Udgivet af Sky McKinley and Megan Levine. Internetadresse: http://pages.intnet.mu/cueboy/education/notes/numerical/cubicsplineinterpol.pdf - Besøgt d. 18.05.2017 (Internet)

Diebold-Mariano Test Statistic. Udgivet af Mathworks. Internetadresse:https://se.mathworks.com/matlabcentral/fileexchange/33979-diebold-mariano-test-statistic - Besøgt d. 14.4.2017 (Internet)

Ejendomsværdiskat og ejendomsskat (grundskyld). Udgivet af Skatteministeriet. Internetadresse: http://www.skm.dk/aktuelt/temaer/boligskat-og-de-offentlige-ejendomsvurderinger/ejendomsvaerdiskat-og-ejendomsskat-grundskyld - Besøgt d. 16.3.2017 (Internet)

FORBRUGERPRISINDEKS. Udgivet af Danmarks Statistik. Internetadresse: http://www.dst.dk/da/Statistik/emner/priser-og-forbrug/forbrugerpriser/forbrugerprisindeks - Besøgt d. 17.4.2017 (Internet)

HUSHOLDNINGERNES FORMUE OG GÆLD. Udgivet af Danmarks Nationalbank. Internetadresse: http://www.nationalbanken.dk/da/publikationer/tema/Sider/Husholdningernes-formue-og-g%C3%A6ld.aspx - Besøgt d. 4.5.2017 (Internet)

Interpolation Methods. Udgivet af Orcaflex. Internetadresse: https://www.orcina.com/SoftwareProducts/OrcaFlex/Documentation/Help/Content/html/InterpolationMethods.htm - Besøgt d. 24.5.2017 (Internet)

Obligationsrenter. Udgivet af Realkreditrådet. Internetadresse: http://finansdanmark.dk/toerre-tal/boligstatistik/obligationsrenter/ - Besøgt d. 13.03.2017 (Internet)

Prediction intervals. Udgivet af Otexts. Internetadresse: https://www.otexts.org/fpp/2/7 - Besøgt d. 6.5.2017 (Internet)

Values of the t-distribution (two-tailed). Udgivet af Medcalc. Internetadresse: https://www.medcalc.org/manual/t-distribution.php - Besøgt d. 14.4.2017 (Internet)

Artikler

A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Welch, Ivo og Amit Goyal, I: UNIL, 2008, s. 1465-1465. Internetadresse: http://www.hec.unil.ch/agoyal/docs/Predictability_RFS.pdf Besøgt d. 23.5.2017 (Artikel)

AR(1) TIME SERIES PROCESS. I: Econometrics 7590, Horvath, Zsuzsanna og Ryan Johnston, s. 3-8. Internetadresse: https://www.math.utah.edu/~zhorvath/ar1.pdf Besøgt d. 10.5.2017 (Artikel)

Fitting the term structure of interest rates: the practical implementation of cubic spline methodology. I: Pienaar , Rod og Moorad Choudhry, yieldcurve, 2005, s. 4-8 (Artikel)

HOUSE PRICE BUBBLES AND THE ADVANTAGES OF STABILISING HOUSING TAXATION. Klein, Asbjørn m.fl. I: Nationalbanken, 2016, s. 48-49. Internetadresse: http://www.nationalbanken.dk/en/publications/Pages/2016/09/House-price-bubbles-and-the-advantages-of-stabilising-housing-taxation.aspx Besøgt d. 4.4.2017 (Artikel)

Husholdningernes balancer og gæld – et internationalt landestudie. Isaksen, Jacob og Paul Lassenius Kramp, I: Danmarks Nationalbank, 2011, s. 44-44. Internetadresse: https://www.nationalbanken.dk/da/publikationer/Documents/2011/12/husholdningernes%20balancer.pdf Besøgt d. 20.5.2017 (Artikel)

MONA - en kvartalsmodel af dansk økonomi. Danmarks Nationalbank, I: Mona, 2003, s. 43-43. Internetadresse: https://www.nationalbanken.dk/da/publikationer/Documents/2003/11/mona_web.pdf Besøgt d. 5.3.2017 (Artikel)

Pengepolitik i Danmark. Danmarks Nationalbank, I: Nationalbanken, 2009, s. 9-9. Internetadresse: http://www.nationalbanken.dk/da/publikationer/Documents/2009/10/pengepol_3udg_dk_web.pdf Besøgt d. 14.4.2017 (Artikel)

REGIONALE ASPEKTER PÅ BOLIGMARKEDET. Hviid, Simon Juul, I: Nationalbanken, 2016, s. 48-51. Internetadresse: http://www.nationalbanken.dk/da/publikationer/Sider/2016/12/Regionale-aspekter-p%C3%A5-boligmarkedet.aspx Besøgt d. 15.3.2017 (Artikel)

Årsberetning . Den Europæiske Centralbank, I: Nationalbanken, 2016, s. 43-43. Internetadresse: https://www.nationalbanken.dk/da/publikationer/andre_institutioner/Documents/ar2016.da.pdf Besøgt d. 12.4.2017 (Artikel)