

Arbejdsløshed - Ugeeksamen i Forecasting

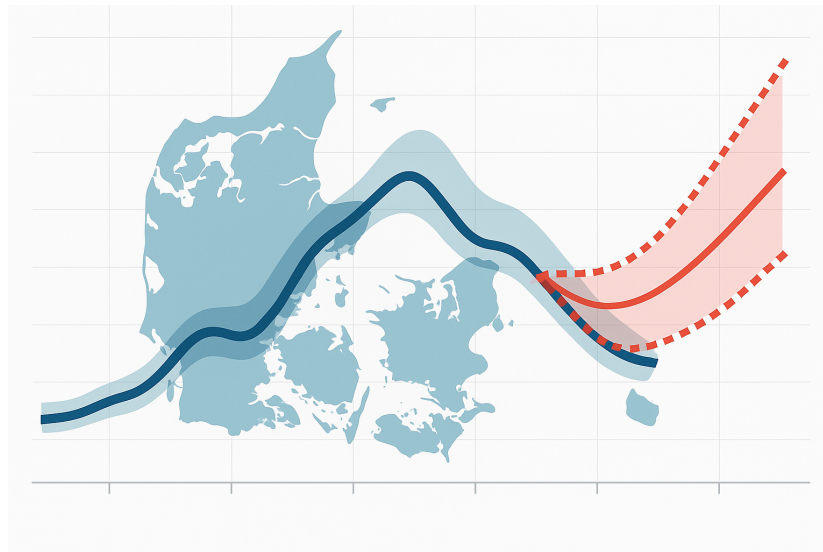
Christine Hegelund

Jing Wei

Marcus Nielsen

2025-06-13

Forecasting Eksamen



Antal tegn (inkl. mellemrum): 97023

Vejledere:

Bjarne Taulo Sørensen

Indholdsfortegnelse

Tabel over figurer	3
1 Introduktion	2
2 Problemformulering	2
2.1 Afgrænsning	3
2.1.1 Anvendelse af AI-værktøj	3
2.2 Definitioner/forkortelser	3
2.3 Struktur	4
3 Data og forberedelse	4
4 Eksplorativ dataanalyse (EDA)	4
4.1 Visualisering af udvikling og mønstre	5
4.1.1 Udvikling over tid	5
4.1.2 Sæsonmønstre	7
4.1.3 Regionale og kønsmæssige forskelle	10
4.2 Deskriptive statistikker	12
4.3 STL-dekomposition	13
5 Modelvalg	15
5.1 Valgte modeltyper	15
5.2 Modellering i praksis	16
6 Modelvalidering	17
6.1 Trænings- og testsplit	17
6.2 Evalueringsmetrikker: RMSE og MAPE	18
6.3 Valg af bedste model pr. serie	18
6.4 Visualisering af forecast vs. faktiske værdier	19
6.5 Time series cross-validation	19
6.6 Residualanalyse og hvid støj	20

7	Forecasting	22
7.1	Estimering og forecast på 2020	22
7.2	Visualisering med prædiktionsintervaller	22
7.3	Centrale prognoseværdier	24
8	Sammenligning og fortolkning	24
8.1	Sammenligning på tværs af serier	24
8.2	Tendenser mellem regioner og køn	24
9	Konklusion	24
10	Kildeliste	25
11	Bilagsoversigt	26

Tabel over figurer

Indholdsfortegnelse

1 Introduktion

Arbejdsløshedstal er en central indikator for et lands økonomiske tilstand og udvikling. De påvirker både den enkelte borger og den nationale økonomi og indgår som en væsentlig faktor i politiske og økonomiske beslutningsprocesser. I takt med øget datatilgængelighed og forbedrede statistiske værktøjer er det blevet muligt at analysere og forudsige sådanne udviklingstræk med større præcision.

Denne opgave har til formål at modellere og forudsige udviklingen i arbejdsløsheden i Danmark, fordelt på køn og region, med afsæt i månedlige data for perioden januar 2007 til december 2019. Datasættet består af ti tidsserier – én for hver kombination af fem regioner og to køn – hvilket muliggør en detaljeret og differentieret analyse af både strukturelle og sæsonbetingede mønstre i ledigheden.

Metodisk anvender analysen klassiske teknikker til tidsserieanalyse med særligt fokus på modellerne ARIMA, ETS og en simpel benchmarkmodel (Seasonal Naive). For hver serie estimeres de tre modeltyper, og deres performance evalueres ved hjælp af forecast-metrikker som RMSE og MAPE. Derudover inddrages elementer som STL-dekomposition, transformation af data (log og Box-Cox), residualanalyse og test for hvid støj (Ljung-Box).

Formålet er ikke blot at fremskrive arbejdsløsheden for året 2020, men også at vurdere styrker og begrænsninger ved klassiske modeller til tidsserieanalyse. Med dette udgangspunkt opstår spørgsmålet, hvordan sådanne modeller bedst kan anvendes til at analysere og forudsige arbejdsløsheden i Danmark, fordelt på region og køn.

2 Problemformulering

Hvordan kan klassiske tidsseriemodeller anvendes til at analysere og forudsige udviklingen i arbejdsløsheden i Danmark, fordelt på region og køn?

For at besvare dette spørgsmål undersøges følgende delspørgsmål:

1. Hvordan varierer arbejdsløshedens udvikling og sæsonmønstre på tværs af regioner og køn, og hvordan kan disse identificeres gennem eksplorativ dataanalyse og dekomposition?
2. Hvordan performer modellerne ARIMA, ETS og Seasonal Naive i forhold til hinanden, når det gælder præcision, residualstruktur og prognoseegenskaber?
3. Hvilke modeller er bedst egnede til at fremskrive arbejdsløshedstallene for 2020, og hvordan varierer resultater og usikkerhed på tværs af serierne?

2.1 Afgrænsning

Analysen er udelukkende baseret på de udleverede data og har fokus på metodisk klarhed, reproducerbarhed og faglig konsistens. Eksterne forklarende faktorer som COVID-19, konjunkturændringer og politiske tiltag inddrages ikke.

Formålet er ikke at give strategiske eller forretningsmæssige anbefalinger, men at demonstrere korrekt og velbegrundet anvendelse af klassiske forecasting-metoder. Der lægges vægt på modellering, evaluering og sammenligning af tids-seriemodeller i en struktureret analytisk ramme.

2.1.1 Anvendelse af AI-værktøj

I udarbejdelsen af denne opgave er ChatGPT (GPT-4o) anvendt som et støtte-værktøj. Værktøjet har primært været anvendt til idéudvikling, sproglig sparring, forbedring af formuleringers klarhed, grammatisk gennemgang samt støtte ved udformning af enkelte kodestumper. Anvendelsen har udelukkende omfattet sproglige, strukturelle og tekniske elementer. Alt analytisk, fortolkende og konkluderende indhold er selvstændigt udarbejdet af gruppens medlemmer.

2.2 Definitioner/forkortelser

I dette afsnit afklares centrale begreber og forkortelser, der anvendes gennem opgaven, for at sikre en ensartet forståelse.

- **ARIMA:** AutoRegressive Integrated Moving Average – klassisk tidsseriemodel med trend og autokorrelation
- **ETS:** Exponential Smoothing State Space Model – glatningsbaseret model til tidsserier med trend og sæson
- **SNAÏVE:** Seasonal Naive – benchmarkmodel, hvor forecast er lig med seneste sæsonværdi
- **RMSE:** Root Mean Squared Error – fejlmål hvor store afvigelser vægtes højt
- **MAPE:** Mean Absolute Percentage Error – fejl angivet i procent
- **STL:** Seasonal-Trend decomposition using Loess – metode til dekomponering af tidsserier
- **CV:** Cross-validation – validering af modeller via rullende trænings-/testvinduer

- **Ljung-Box:** Test for autokorrelation i modelresidualer (white noise)

2.3 Struktur

Opgaven er struktureret i overensstemmelse med en klassisk tilgang til analyse af tidsserier og følger fire hovedfaser.

Først gennemføres en eksplorativ dataanalyse (EDA), hvor serierne visualiseres og undersøges for tendenser og sæsonmønstre. Dette sker ved hjælp af grafer, deskriptive statistikker og STL-dekomposition. Data transformeres ved behov. Dernæst estimeres tre modeltyper for hver serie: ARIMA, ETS og Seasonal Naive. Disse modeller valideres gennem krydsvalidering og residualanalyse. På baggrund af modelperformance foretages fremskrivninger af arbejdsløsheden for 2020 med tilhørende prædiktionsintervaller. Afslutningsvis sammenlignes modellerne og arbejdsløshedsniveauerne på tværs af regioner og køn, og der konkluderes på analysens hovedspørgsmål.

3 Data og forberedelse

Analysen er baseret på et datasæt med månedlige arbejdsløshedstal i Danmark fra januar 2007 til december 2019. Data er opdelt på køn (mænd og kvinder) og region (de fem danske regioner), hvilket resulterer i ti separate tidsserier.

Datasættet er udleveret i forbehandlet format som en tsibble med korrekt angivne indeks- og nøglevariabler.

```
# A tsibble: 6 x 4 [1M]
# Key:      kon, region [1]
   kon      region      yearmonth svalue
  <fct>   <fct>         <mth>   <dbl>
1 Kvinder Region Hovedstaden 2007 Jan    2.26
2 Kvinder Region Hovedstaden 2007 Feb    2.19
3 Kvinder Region Hovedstaden 2007 Mar    2.09
4 Kvinder Region Hovedstaden 2007 Apr    1.97
5 Kvinder Region Hovedstaden 2007 May    1.99
6 Kvinder Region Hovedstaden 2007 Jun    1.93
```

4 Eksplorativ dataanalyse (EDA)

Dette afsnit har til formål at skabe et overblik over datasættets struktur og variation forud for modelvalget. Gennem visualiseringer, deskriptive statistikker og dekomponering undersøges arbejdsløshedens udvikling på tværs af regioner

og køn fra 2007 til 2019. Analysen afdækker tendenser, sæsonmønstre og forskelle i niveau og variation. STL-dekomposition anvendes til at isolere trend, sæson og residualer, mens datatransformation vurderes for at sikre modellérbarhed. Resultaterne danner det metodiske grundlag for den efterfølgende modellering.

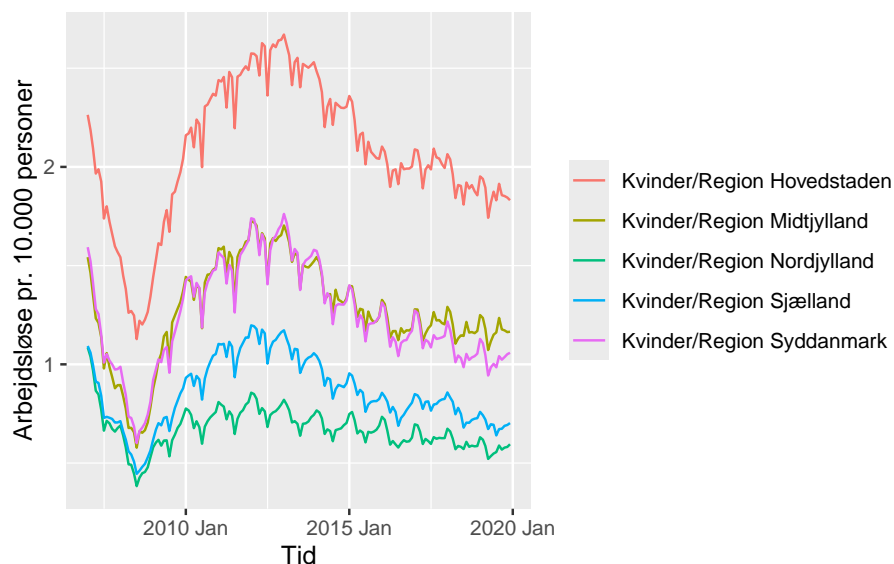
4.1 Visualisering af udvikling og mønstre

Formålet med dette afsnit er at identificere centrale mønstre og karakteristika i arbejdsløshedsdataene – herunder trends, sæsonvariationer, regionale forskelle og forskelle mellem køn. Visualiseringerne danner grundlag for valg af passende modeller til hver tidsserie.

4.1.1 Udvikling over tid

I de følgende figurer undersøges udviklingen i arbejdsløsheden over tid. Til at begynde med anvendes den oprindelige skala for at give et lettilgængeligt overblik (Figur 1 og 2). Fra Figur 3 og frem benyttes en log-transformation for at stabilisere variationen i serierne, især blandt mænd, og for at sikre sammenlignelighed i videre modellering og forecast.

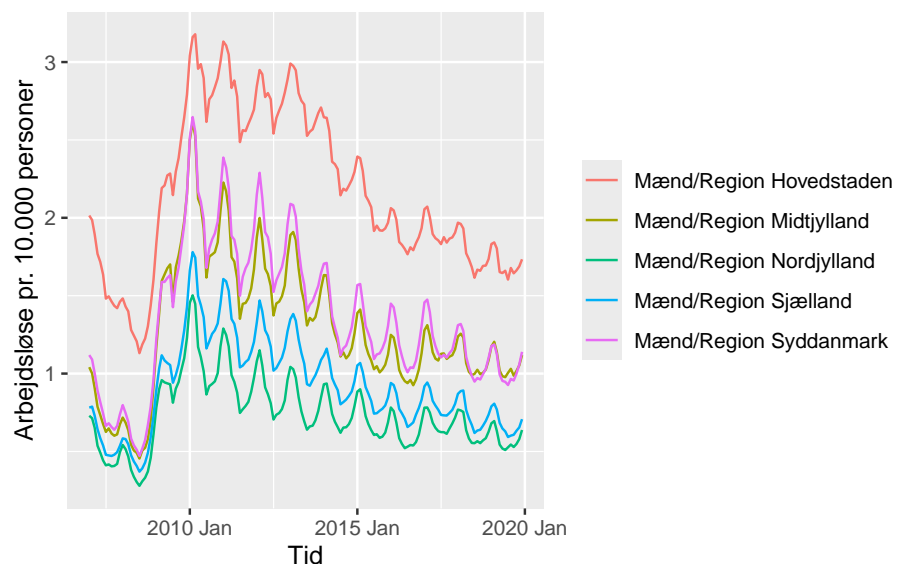
Figur 1: Udvikling i kvinders arbejdsløshed pr. region



Figuren viser den månedlige arbejdsløshed blandt kvinder i de fem danske regioner fra 2007 til 2019. Der fremgår et tydeligt sæsonmønster med højere ledighed i vintermånederne og lavere i sommerperioden. Niveautet varierer regionalt, hvor

Region Hovedstaden typisk ligger højest og Nordjylland lavest. Samlet ses en faldende tendens efter finanskrisen.

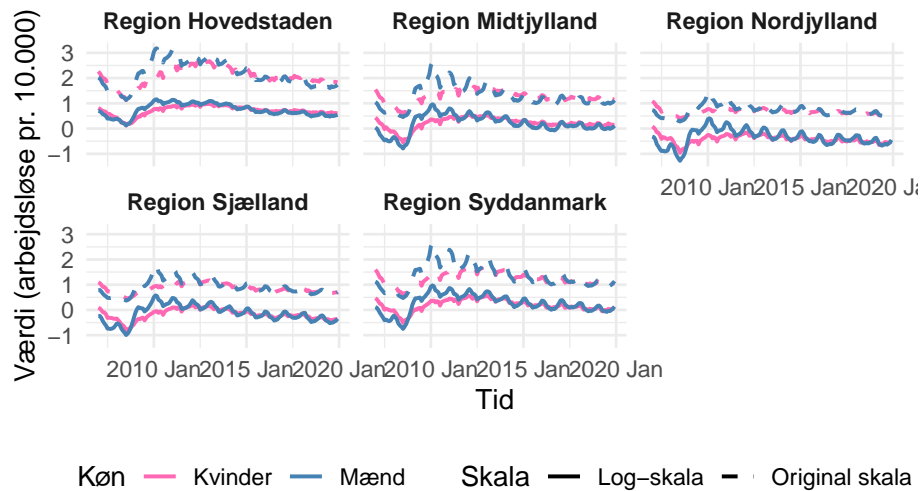
Figur 2: Udvikling i mænds arbejdsløshed pr. region



Sammenlignet med kvinder ses større udsving og et generelt højere niveau i mænds arbejdsløshed. Sæsonmønstrene er mere markante, og variationen mellem regionerne er tydeligere – især i Region Syddanmark og Midtjylland, hvor krisen i 2008-2009 medfører et kraftigt midlertidigt løft i ledigheden.

Figur 3: Udvikling i arbejdsløshed hos pr. region og køn (log-skala)

Sammenligning af original og log-skala

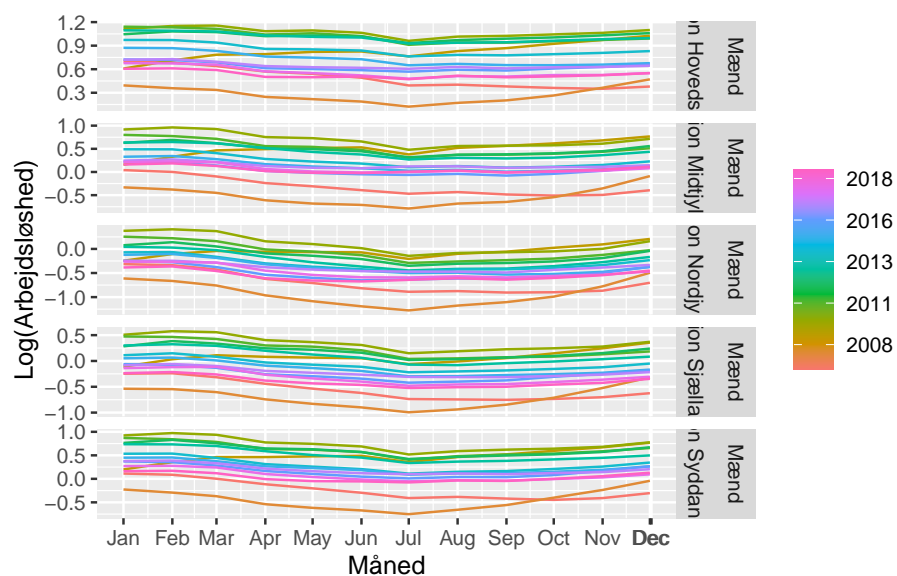


Figur 3 viser, hvordan arbejdsløsheden har udviklet sig for mænd og kvinder i de fem danske regioner – både på original og log-transformeret skala. De stiplede linjer repræsenterer de oprindelige værdier, mens de fuldt optrukne linjer viser samme udvikling på log-skala. Log-transformationen gør det nemmere at få øje på mønstre og forskelle, især i serier med store udsving eller høj varians. Hvor den oprindelige skala fremhæver de absolutte niveauer, giver log-skalaen et mere balanceret overblik over den relative udvikling. Den log-transformerede fremstilling danner derfor grundlag for den videre analyse.

4.1.2 Sæsonmønstre

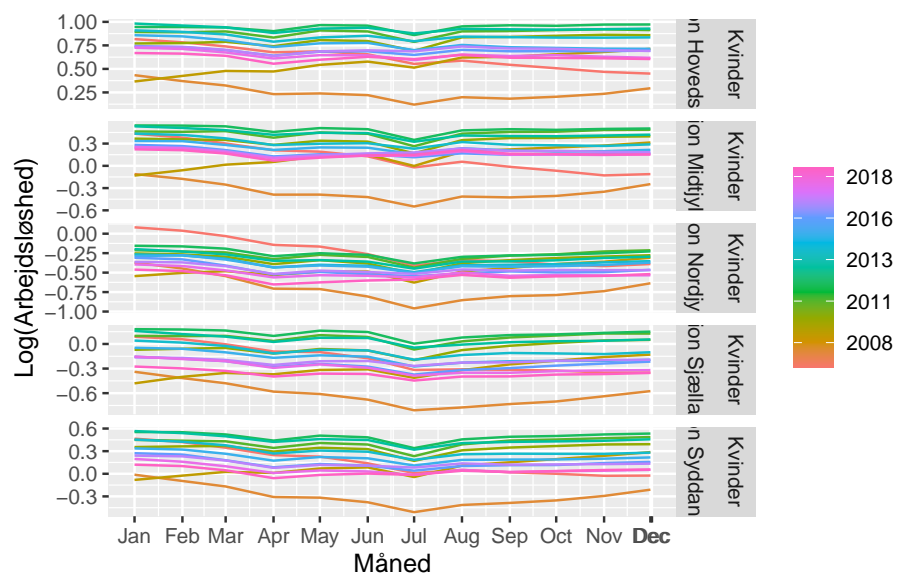
Med log-transformerede data som fundament undersøges nu sæsonmønstrene i arbejdsløsheden nærmere. Gennem sæsonplots og subserieplots vurderes, hvordan arbejdsløsheden typisk varierer over året – og hvordan dette adskiller sig mellem mænd og kvinder samt på tværs af regioner.

Figur 4: Sæsonmønster for mænd pr. region



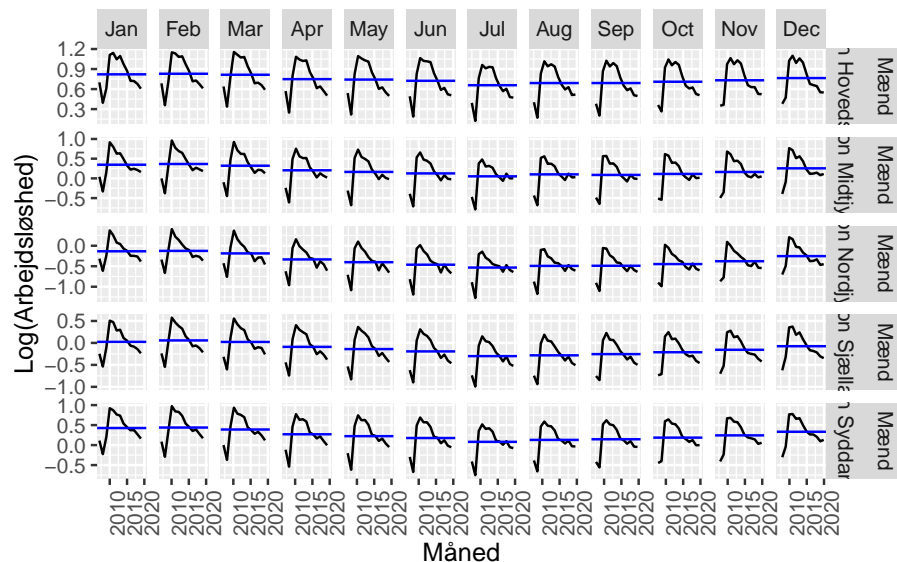
Figur 4 viser sæsonmønstre i mænds arbejdsløshed fordelt på regioner. Hver farvet linje repræsenterer et år, og linjernes forløb illustrerer, hvordan arbejdsløsheden typisk er højest i årets begyndelse (januar-marts) og lavest i sommermånederne (juni-august). Mønsteret er relativt konsistent på tværs af år og regioner, hvilket understøtter tilstedeværelsen af en stabil sæsonkomponent i mænds ledighed. De højeste niveauer ses i Region Hovedstaden, mens Region Sjælland og Region Nordjylland ligger lavere. Bemærk især det markante fald fra marts til juni, som gentager sig på tværs af årstal.

Figur 5: Sæsonmønster for kvinder pr. region



Sammenlignet med mænd er sæsonudsvingene blandt kvinder mindre udtalte, og udviklingen fremstår mere jævn og stabil på tværs af årene. Niveauforskellene mellem regionerne er bevaret, men udsvingene over måneder er mere afdæmpede. Det indikerer en lavere volatilitet i kvinders arbejdsløshed og en mere ensartet sæsonprofil.

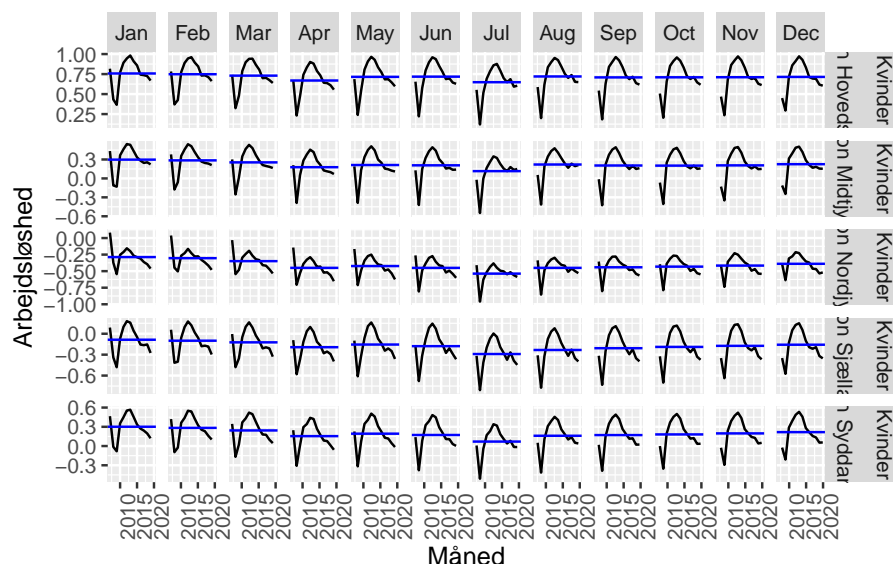
Figur 6: Subserieplot for mænd pr. måned og region



Figur 6 viser udviklingen i mænds arbejdsløshed fordelt på måneder og region-

ner, hvor hver facetteret celle repræsenterer en måned. Den sorte linje viser udviklingen over tid, mens den blå streg repræsenterer gennemsnittet. Det ses, at arbejdsløsheden konsekvent toppes i årets første måneder og falder hen over sommeren. Desuden er udsvingene mellem år særligt store i vinterhalvåret, hvilket understreger en høj sæsonafhængighed og volatilitet i mænds ledighedsniveau. Mønstret er gennemgående i alle regioner, om end niveauet varierer.

Figur 7: Subserieplot for kvinder pr. måned og region



Sammenlignet med mænd viser kvinder tilsvarende sæsonmønstre med højere ledighed i årets begyndelse og lavere i sommermånederne. Udsvingene er dog mindre markante, og variationen mellem år er mindre tydelig. Mønstret fremstår mere stabilt og jævnt fordelt på tværs af regioner, hvilket tyder på lavere volatilitet i kvindernes arbejdsløshed.

Udover sæsonmæssige udsving varierer arbejdsløsheden betydeligt på tværs af både geografi og køn. Dette undersøges nærmere i de følgende visualiseringer.

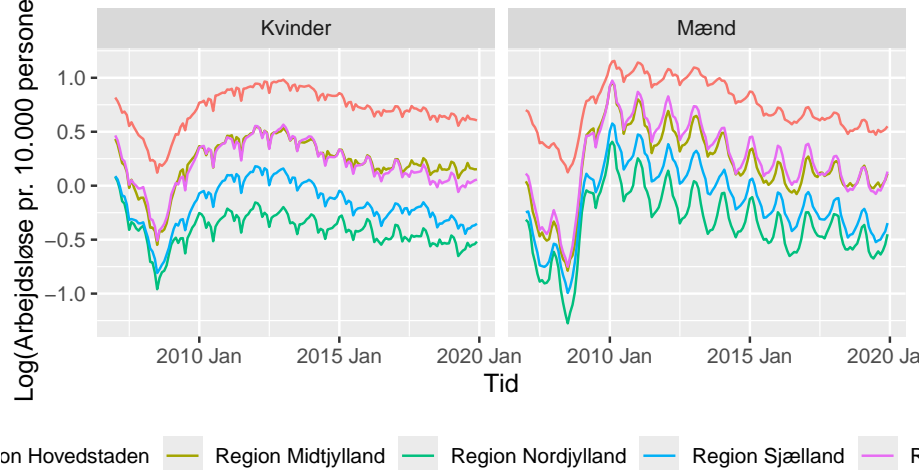
4.1.3 Regionale og kønsmæssige forskelle

I dette afsnit rettes fokus mod samspillet mellem køn og region. Der undersøges, hvordan forskelle i arbejdsløshedsniveau og -mønstre varierer både mellem regioner og mellem mænd og kvinder. Visualiseringerne præsenteres i forskellige facetteringer for at fremhæve de vigtigste forskelle.

Figur 8: Arbejdsløshed fordelt på regioner – opdelt efter køn

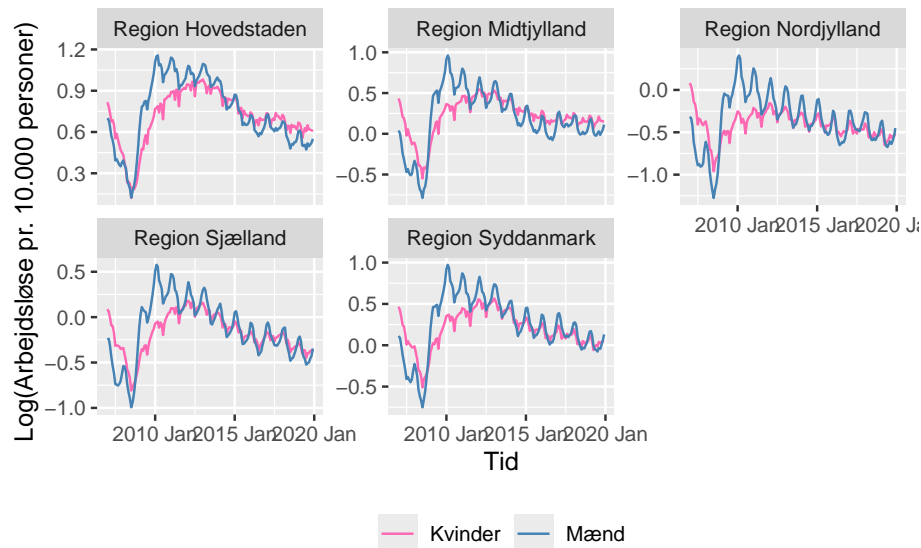
Figur 8

Arbejdsløshed fordelt på regioner – opdelt efter køn (log-skala)



Figur 8 viser den log-transformerede udvikling i arbejdsløsheden blandt mænd og kvinder, fordelt på regioner. Hver farve repræsenterer en region, og mænd og kvinder vises i hver sin fane, så det er lettere at sammenligne på tværs. Log-skalaen gør forskellene mellem regionerne mere overskuelige og dæmper de største udsving. Niveauet er stadig højest i Region Hovedstaden og lavest i Nordjylland. Der er tydelige sæsonbevægelser, særligt blandt mænd, hvor udsvingene fortsat er størst. For kvinder fremstår udviklingen mere jævn.

Figur 9: Arbejdsløshed fordelt på køn – opdelt efter region



Hvor Figur 8 fokuserede på at sammenligne regioner inden for hvert køn, giver Figur 9 et omvendt perspektiv: her sammenlignes køn i hver af de fem danske regioner. Mænd har generelt højere arbejdsløshed og mere markante udsving end kvinder – særligt i perioden omkring finanskrisen. Forskellene mellem kønnene er forholdsvis stabile på tværs af regioner, men mønstret bekræfter, at mænds ledighed både toppe og varierer mere over tid. Kvindernes ledighed er lavere og udvikler sig mere glidende. Figuren understøtter dermed beslutningen om at analysere serierne individuelt. Samlet set viser figurene, at arbejdsløsheden varierer tydeligt mellem både regioner og køn, hvilket bekræfter relevansen af at behandle hver tidsserie separat i den videre analyse. Som næste skridt suppleres med deskriptive statistikker, der opsummerer forskelle i niveau og udsving.

4.2 Deskriptive statistikker

For at supplere graferne anvendes deskriptive statistikker som gennemsnit, spredning, minimum og maksimum. De opsummerer niveau og variation i arbejdsløsheden pr. region og køn (Hyndman & Athanasopoulos, 2021). Statistikkerne giver et hurtigt overblik over forskelle mellem grupper og hjælper med at identificere serier med høj volatilitet. Det bidrager til at vurdere, hvor der kan være behov for særlig opmærksomhed i den videre modellering.

Tabel 1: Deskriptiv statistik per region og køn

region	kon	Gennemsnit	Standardafvigelse	Minimum	Maksimum
--------	-----	------------	-------------------	---------	----------

Region Hovedstaden	Kvinder	2.07	0.36	1.13	2.67
Region Hovedstaden	Mænd	2.17	0.52	1.13	3.18
Region Midtjylland	Kvinder	1.27	0.25	0.58	1.72
Region Midtjylland	Mænd	1.29	0.44	0.45	2.62
Region Nordjylland	Kvinder	0.67	0.10	0.38	1.09
Region Nordjylland	Mænd	0.74	0.23	0.28	1.50
Region Sjælland	Kvinder	0.86	0.17	0.45	1.20
Region Sjælland	Mænd	0.92	0.30	0.37	1.78
Region Syddanmark	Kvinder	1.24	0.26	0.60	1.76
Region Syddanmark	Mænd	1.37	0.47	0.47	2.65

Tabel 1 opsummerer arbejdsløsheden på tværs af regioner og køn. Gennemsnittet viser det generelle niveau, mens standardafvigelsen giver et mål for variation over tid. Mænd har generelt højere gennemsnit og større udsving end kvinder, særligt i Region Hovedstaden og Syddanmark. Omvendt ses lavere og mere stabile niveauer i Nordjylland, især blandt kvinder. Minimums- og maksimumsværdier understøtter indtrykket af høj volatilitet i visse serier – hvilket kan få betydning for modelvalget i næste fase.

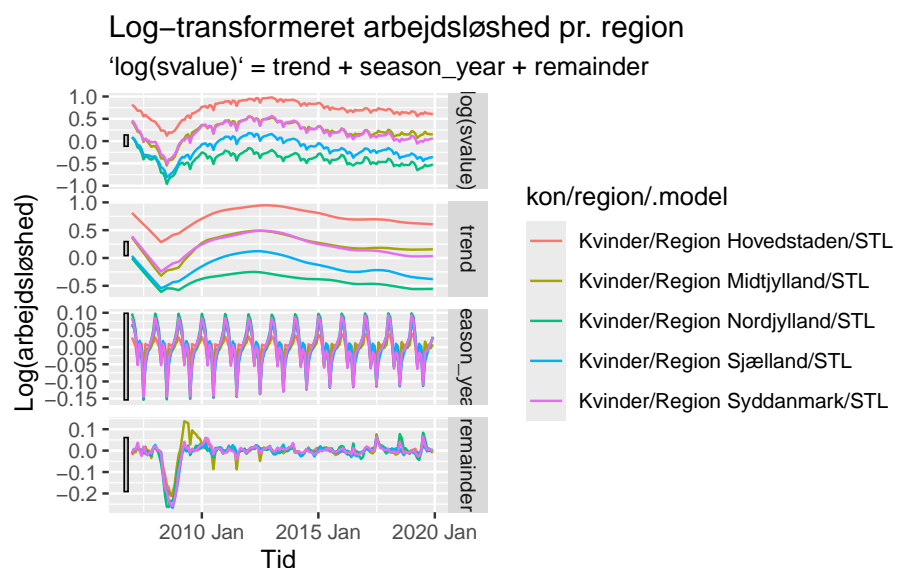
4.3 STL-dekomposition

For at opnå en bedre forståelse af de strukturelle komponenter i arbejdsløshedsserierne anvendes STL-dekomposition (Seasonal-Trend decomposition using Lo-ess). Metoden adskiller tidsserier i tre komponenter: trend, sæson og remainder (Hyndman & Athanasopoulos, 2021). Det giver indblik i, hvor meget af variationen der kan forklares af langsigtede bevægelser, gentagne sæsonmønstre eller kortsigtede, uregelmæssige udsving.

Da særligt mændenes serier udviser høje niveauer og større variation, er alle serier log-transformeret forud for dekompositionen. Denne transformation stabiliserer variansen og gør det lettere at sammenligne komponenterne på tværs af regioner og køn.

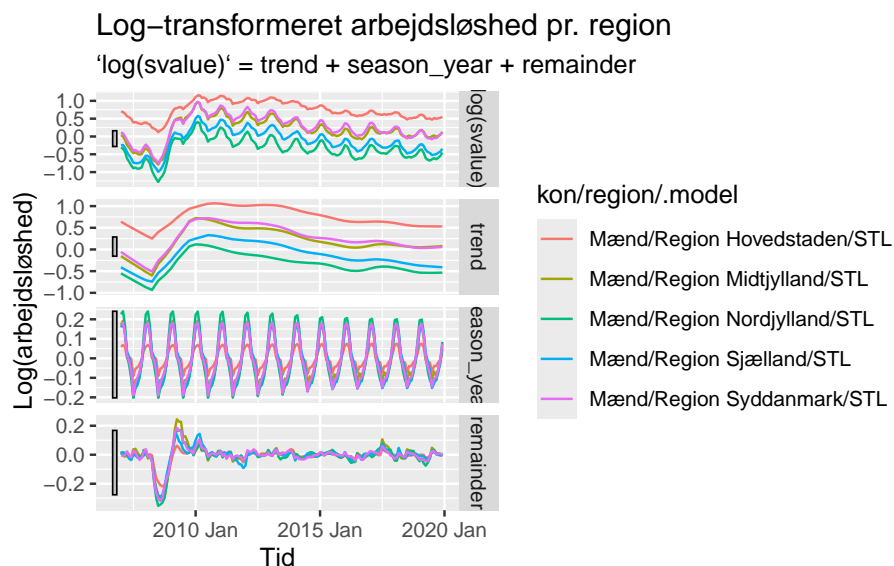
STL anvendes her på de log-transformerede serier for henholdsvis kvinder og mænd i hver region. Resultaterne præsenteres i figur 10 og 11.

Figur 10: STL-dekomposition for kvinder – log-transformeret arbejdsløshed pr. region



Figur 10 viser STL-dekompositionen for kvinder i de fem danske regioner. Der ses tydelige sæsonmønstre med tilbagevendende lavpunkter i sommermånederne og højere ledighed i vinterhalvåret. Trends varierer mellem regionerne: Region Hovedstaden udviser generelt højere ledighed, mens Nordjylland og Sjælland ligger lavere. Residualerne ligger stabilt omkring nul, hvilket indikerer, at modellen formår at fange de dominerende strukturer i data.

Figur 11: STL-dekomposition for mænd – log-transformeret arbejdsløshed pr. region



Figur 11 viser tilsvarende dekomposition for mænd. Her ses en tilsvarende stærk sæsonkomponent, men med mere markante udsving og højere ledighedsniveauer – især efter finanskrisen. Trends følger samme overordnede forløb som for kvinder, men afvigelserne er større. Dette bekræfter tidligere observationer om højere volatilitet i mænds ledighed. Residualkomponenten er mere varierende, hvilket kan indikere, at nogle udsving ikke fanges fuldt ud af modellen.

Disse observationer fra STL-dekompositionen bekræfter, at både trend og sæsonmønstre varierer betydeligt på tværs af serierne. Det understreger behovet for modeller, der kan tilpasses hver series struktur, hvilket vil blive præsenteret i det følgende afsnit.

5 Modelvalg

I dette afsnit estimeres og sammenlignes tre klassiske modeltyper, ARIMA, ETS og benchmarkmodellen SNaive, for hver tidsserie (region \times køn). Hver modeltype har forskellige antagelser og fordele og kan fange forskellige karakteristika i data, såsom trend, sæson og autokorrelation. Modelvalget foretages automatisk for hver serie via fable-pakken, og resultaterne anvendes i den efterfølgende validering og forecast.

5.1 Valgte modeltyper

ARIMA (AutoRegressive Integrated Moving Average)

ARIMA-modeller anvendes til at modellere tidsserier med autokorrelation og

ikke-stationaritet. De kombinerer autoregressive (AR) led, differens (I) for at opnå stationaritet, og glidende gennemsnit (MA) led til at modellere fejl. Ved at inkludere sæsonkomponenter (SARIMA) kan modellen håndtere årligt gentagende mønstre. ARIMA er særligt velegnet til serier med langsigtede trends og strukturelle skift, hvor tidligere værdier har stor indflydelse på den nuværende tilstand (jf. Hyndman & Athanasopoulos, kap. 9).

ETS (Exponential Smoothing State Space Model)

ETS-modeller benytter eksponentiel glatning til at vægte nyere observationer højere og beskriver tidsserier ud fra tre elementer: fejl, trend og sæson. Hver komponent kan være additiv eller multiplicativ afhængigt af dataens karakter. ETS egner sig godt til serier med tydelige strukturer og sæsonrytmer, især når der ikke er behov for at modellere autokorrelation direkte (jf. kap. 8).

SNAÏVE (Seasonal Naive)

SNaive er en enkel benchmarkmodel, der baserer hvert forecast på den tilsvarende værdi fra samme sæson i den foregående periode. Modellen fanger sæsonmønstre, men tager ikke højde for trend eller afhængighed i data. Den anvendes til at vurdere, om mere komplekse modeller giver en mærkbart bedre forecast-præcision (jf. kap. 3 og 8).

5.2 Modellering i praksis

Ved hjælp af fable-pakken estimeres de tre modeller automatisk for hver af de ti tidsserier på baggrund af data fra 2007 til 2019. Herefter genereres forecasts for 2020 med tilhørende prædiktionsintervaller. Prognoserne visualiseres, så man kan sammenligne modellernes adfærd på tværs af regioner og køn.

Tabel 2: Automatisk valgte modelstrukturer for hver serie (region \times køn)

kon	region	.model	AICc
Kvinder	Region Hovedstaden	ARIMA	-519.59277
Kvinder	Region Hovedstaden	ETS	-106.04957
Kvinder	Region Hovedstaden	SNaive	NA
Kvinder	Region Midtjylland	ARIMA	-559.62392
Kvinder	Region Midtjylland	ETS	-186.73947
Kvinder	Region Midtjylland	SNaive	NA
Kvinder	Region Nordjylland	ARIMA	-794.48456
Kvinder	Region Nordjylland	ETS	-420.44010
Kvinder	Region Nordjylland	SNaive	NA
Kvinder	Region Sjælland	ARIMA	-714.46022
Kvinder	Region Sjælland	ETS	-355.53652
Kvinder	Region Sjælland	SNaive	NA
Kvinder	Region Syddanmark	ARIMA	-615.90506

Kvinder	Region Syddanmark	ETS	-194.25088
Kvinder	Region Syddanmark	SNaive	NA
Mænd	Region Hovedstaden	ARIMA	-451.76084
Mænd	Region Hovedstaden	ETS	-94.42293
Mænd	Region Hovedstaden	SNaive	NA
Mænd	Region Midtjylland	ARIMA	-365.22171
Mænd	Region Midtjylland	ETS	24.67626
Mænd	Region Midtjylland	SNaive	NA
Mænd	Region Nordjylland	ARIMA	-536.74821
Mænd	Region Nordjylland	ETS	-202.87509
Mænd	Region Nordjylland	SNaive	NA
Mænd	Region Sjælland	ARIMA	-524.52351
Mænd	Region Sjælland	ETS	-271.66140
Mænd	Region Sjælland	SNaive	NA
Mænd	Region Syddanmark	ARIMA	-387.29279
Mænd	Region Syddanmark	ETS	-55.07382
Mænd	Region Syddanmark	SNaive	NA

Modelvalget er gennemført for hver enkelt tidsserie ved automatisk estimering af ARIMA-, ETS- og SNaive-modeller med fable-pakken. Udvalgelsen er sket på baggrund af AICc, som sikrer en afbalanceret vurdering af modelkompleksitet og tilpasningsevne – særligt vigtigt for relativt korte tidsserier. Resultaterne i Tabel 2 viser, at modellerne varierer betydeligt på tværs af køn og regioner, hvilket understreger nødvendigheden af en differentieret strategi, hvor hver tidsserie behandles individuelt.

SNaive-modellen er anvendt som benchmark og danner referencepunkt for den efterfølgende evaluering. Den egentlige vurdering af modellernes forecast-præcision foretages først i næste afsnit, hvor modellerne sammenlignes på baggrund af tidsserie-krydsvalidering og metrikker som RMSE og MAPE.

6 Modelvalidering

For at opnå en robust vurdering kombineres testsplit og time series cross-validation. Testsplittet simulerer en realistisk forecast-situation, mens cross-validation måler præcisionen over flere tidspunkter og reducerer risikoen for tilfældige variationer.

6.1 Trænings- og testsplit

Datasættet opdeles i en træningsperiode (2007–2018) og en testperiode (2019). Modellerne estimeres på træningsdata og genererer forecasts for teståret. Præ-

cisionen vurderes ved sammenligning med de faktiske observationer fra 2019.

6.2 Evalueringsmetrikker: RMSE og MAPE

For at vurdere, hvor godt modellerne forudsiger arbejdsløsheden, sammenlignes deres forecasts for 2019 med de faktiske observationer. Der anvendes to metrikker:

- RMSE måler den gennemsnitlige afvigelse og lægger vægt på store fejl.
- MAPE viser den gennemsnitlige procentvise fejl og muliggør sammenligning på tværs af serier med forskellig måleenhed.

Tabellen nedenfor viser RMSE og MAPE for alle modeller og serier. Værdierne bruges som grundlag for at vurdere, hvilke modeller der præsterer bedst.

```
# A tibble: 30 x 5
  kon      region      .model  RMSE  MAPE
  <fct>   <fct>      <chr>   <dbl> <dbl>
1 Kvinder Region Hovedstaden ARIMA  0.0340 1.53
2 Kvinder Region Hovedstaden SNaive 0.0662 3.02
3 Kvinder Region Hovedstaden ETS     0.0932 4.17
4 Mænd    Region Hovedstaden SNaive 0.0784 3.31
5 Mænd    Region Hovedstaden ETS     0.0786 3.57
6 Mænd    Region Hovedstaden ARIMA  0.0948 5.08
7 Kvinder Region Midtjylland ARIMA  0.0135 0.999
8 Kvinder Region Midtjylland SNaive 0.0215 1.53
9 Kvinder Region Midtjylland ETS     0.0334 2.60
10 Mænd   Region Midtjylland SNaive 0.0416 2.98
# i 20 more rows
```

6.3 Valg af bedste model pr. serie

For hver kombination af region og køn vælges den model, der opnår lavest RMSE i testperioden. Dette valg sikrer, at den videre analyse baseres på modeller med høj prædiktiv nøjagtighed.

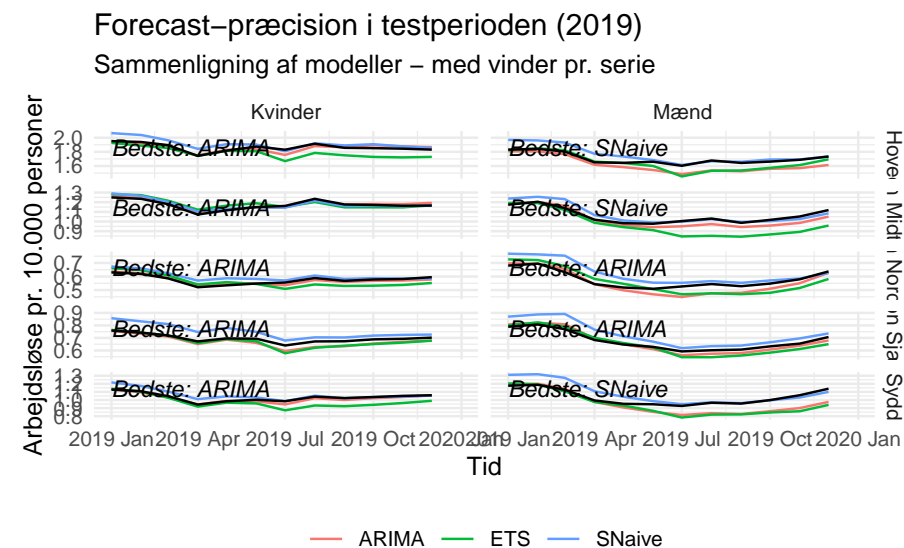
```
# A tibble: 10 x 5
  kon      region      .model  RMSE  MAPE
  <fct>   <fct>      <chr>   <dbl> <dbl>
1 Kvinder Region Hovedstaden ARIMA  0.0340 1.53
2 Kvinder Region Midtjylland ARIMA  0.0135 0.999
3 Kvinder Region Nordjylland ARIMA  0.0119 1.83
```

4	Kvinder	Region Sjælland	ARIMA	0.0282	3.61
5	Kvinder	Region Syddanmark	ARIMA	0.0174	1.36
6	Mænd	Region Hovedstaden	SNaive	0.0784	3.31
7	Mænd	Region Midtjylland	SNaive	0.0416	2.98
8	Mænd	Region Nordjylland	ARIMA	0.0391	5.71
9	Mænd	Region Sjælland	ARIMA	0.0236	3.22
10	Mænd	Region Syddanmark	SNaive	0.0826	5.85

Som det fremgår af tabellen, vælges ARIMA-modellen for samtlige serier for kvinder. For mænd er billedet mere varieret: ARIMA vinder i Nordjylland og Sjælland, mens SNaive udgør bedste model i de øvrige regioner. Det viser, at præcisionen afhænger både af den regionale kontekst og datamønstre knyttet til køn.

6.4 Visualisering af forecast vs. faktiske værdier

Forecast og faktisk udvikling kan visualiseres for alle 10 tidsserier:



Den identificerede model med lavest forecastfejl anvendes i den videre analyse, da den vurderes mest robust til at forudsige fremtidige observationer.

6.5 Time series cross-validation

For at supplere den traditionelle opdeling i trænings- og testperiode anvendes time series cross-validation med rullende vindue. Dette gør det muligt at evaluere

forecast-præcisionen på tværs af flere tidspunkter og dermed sikre mere robust modelvurdering.

Datasættet strækkes med et startvindue på 60 måneder og forecast-horisont på 1 måned. For hver udvidet vindue estimeres modellerne (ARIMA, ETS og SNaive), og forecast-fejl beregnes.

Gennemsnitlig performance på tværs af alle tidsserier:

```
# A tibble: 3 x 3
  .model  RMSE  MAPE
  <chr>   <dbl> <dbl>
1 ARIMA  0.0329  2.00
2 ETS    0.0389  2.41
3 SNaive 0.115   7.69
```

ARIMA opnår de laveste gennemsnitlige RMSE- og MAPE-værdier og præsterer bedst på tværs af serier.

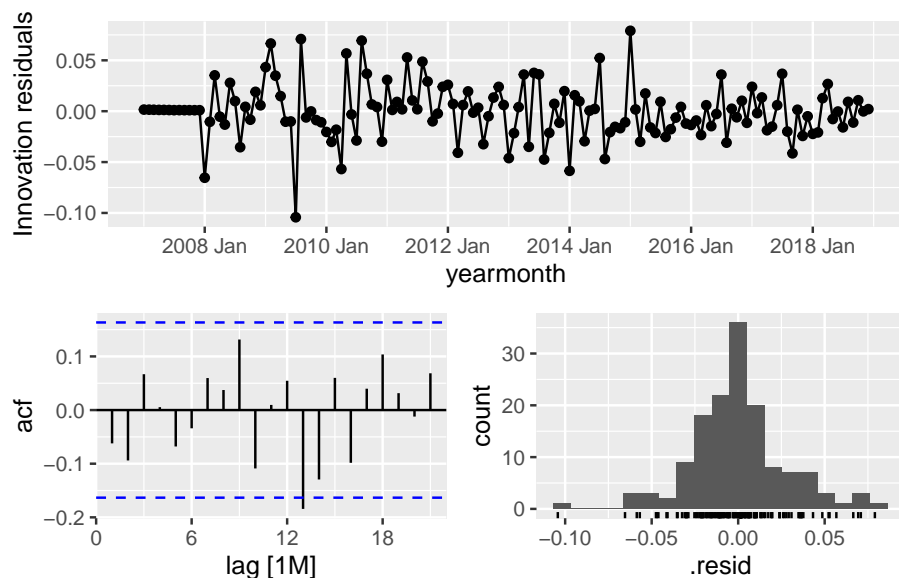
Den bedste model pr. serie (lavest RMSE):

```
# A tibble: 10 x 5
  kon      region      .model  RMSE  MAPE
  <fct>   <fct>      <chr>   <dbl> <dbl>
1 Kvinder Region Hovedstaden ARIMA  0.0376  1.30
2 Kvinder Region Midtjylland ARIMA  0.0250  1.44
3 Kvinder Region Nordjylland ETS    0.0185  2.27
4 Kvinder Region Sjælland   ARIMA  0.0170  1.38
5 Kvinder Region Syddanmark  ARIMA  0.0236  1.38
6 Mænd    Region Hovedstaden ARIMA  0.0465  1.52
7 Mænd    Region Midtjylland ARIMA  0.0422  2.41
8 Mænd    Region Nordjylland ARIMA  0.0281  2.95
9 Mænd    Region Sjælland   ARIMA  0.0308  2.16
10 Mænd   Region Syddanmark  ARIMA  0.0503  2.42
```

Resultatet viser, at ARIMA vælges for størstedelen af serierne, hvilket bekræfter den højere præcision i forhold til benchmark og ETS-modeller.

6.6 Residualanalyse og hvid støj

For at undersøge, om den valgte model fanger strukturen i data tilstrækkeligt, kan residualerne fra forecastmodellen inspiceres. Figuren nedenfor ses et eksempel på residualanalyse for ARIMA-modellen estimeret på kvinder i Region Syddanmark:



Figuren indeholder tre elementer: et residualplot over tid, et ACF-plot (autokorrelationsfunktion) og et histogram. Residualerne synes at være tilfældigt fordelt omkring nul, uden synlige systematiske mønstre. ACF-plottet viser ingen signifikant autokorrelation, og histogrammet er centreret og symmetrisk. Samlet indikerer dette, at modellen er veltilpasset og ikke efterlader systematisk information i fejlleddene. Dette mønster genfindes i de øvrige ARIMA-modeller og understøttes af de statistiske test nedenfor.

Som supplement til den visuelle inspektion anvendes Ljung-Box-testen til at undersøge, om residualerne fra modellerne er ukorrelerede – altså om de udgør hvid støj.

```
# A tibble: 30 x 5
  kon      region      .model lb_stat lb_pvalue
<fct> <fct>      <chr>   <dbl>   <dbl>
1 Kvinder Region Hovedstaden ARIMA    30.6 1.67e- 1
2 Kvinder Region Hovedstaden ETS     99.6 3.54e-11
3 Kvinder Region Hovedstaden SNaive  699.  0
4 Kvinder Region Midtjylland ARIMA    24.4 4.41e- 1
5 Kvinder Region Midtjylland ETS     70.9 1.62e- 6
6 Kvinder Region Midtjylland SNaive  627.  0
7 Kvinder Region Nordjylland ARIMA    26.2 3.45e- 1
8 Kvinder Region Nordjylland ETS    100. 2.53e-11
9 Kvinder Region Nordjylland SNaive  614.  0
10 Kvinder Region Sjælland  ARIMA    26.5 3.29e- 1
# i 20 more rows
```

For ARIMA-modellerne er p-værdierne generelt ikke-signifikante ($p > 0.05$),

hvilket indikerer ukorrelerede residualer og dermed god modeltilpasning. Omvendt viser ETS- og SNaive-modeller i flere tilfælde tegn på strukturer i residualerne. Det styrker indtrykket af, at ARIMA er den mest velegnede modeltype i dette datasæt.

På tværs af både testsplit og cross-validation opnår ARIMA den højeste forecast-præcision og opfylder modelantagelserne bedst. Den anvendes derfor som primær model i den videre analyse.

7 Forecasting

På baggrund af modelvalideringen foretages nu egentlige forecasts for året 2020. For hver af de ti tidsserier anvendes den model, der klarede sig bedst under valideringen. Modellerne estimeres på hele perioden 2007–2019 for at sikre maksimal informationsudnyttelse i den endelige træning.

7.1 Estimering og forecast på 2020

Data frem til december 2019 bruges som træningsgrundlag. For hver serie estimeres ARIMA, ETS og SNaive, hvorefter det forecast vælges, som matcher den bedste modeltype (fra valideringen):

```
data_train <- data |> filter_index(. ~ "2019 Dec")

models_final <- data_train |>
  model(
    ARIMA = ARIMA(svalue),
    ETS = ETS(svalue),
    SNaive = SNAIVE(svalue)
  )

forecast_2020 <- models_final |> forecast(h = "12 months")

forecast_best <- forecast_2020 |>
  inner_join(vinder_modeller, by = c("kon", "region", ".model"))
```

7.2 Visualisering med prædiktionsintervaller

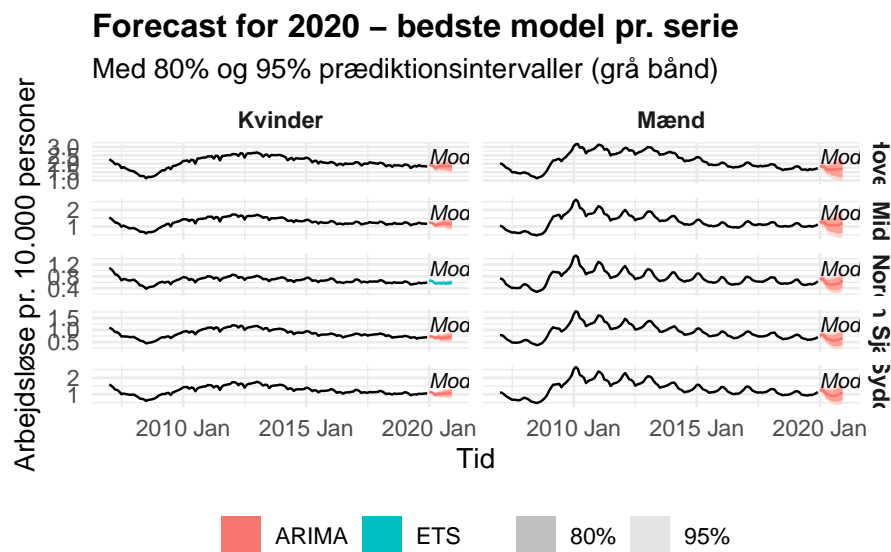
Forecasts for 2020 visualiseres for alle serier med prædiktionsintervaller på 80% og 95%. Den bedste model for hver serie fremgår af farven på forecast-linjen og angives desuden som label i hvert facet.

```

labels <- vinder_modeller |>
  mutate(label = paste0("Model: ", .model))

forecast_best |>
  autoplot(data) +
  geom_text(data = labels,
            aes(x = as.Date("2020-01-01"), y = Inf, label = label),
            inherit.aes = FALSE,
            vjust = 1.5, hjust = 0, size = 3.2, fontface = "italic") +
  facet_grid(region ~ kon, scales = "free_y") +
  labs(
    title = "Forecast for 2020 - bedste model pr. serie",
    subtitle = "Med 80% og 95% prædiktionsintervaller (grå bånd)",
    y = "Arbejdsløse pr. 10.000 personer",
    x = "Tid"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    legend.title = element_blank(),
    plot.title = element_text(face = "bold"),
    strip.text = element_text(face = "bold")
  )

```



Forecast vises med farvet linje (model) og grå bånd for prædiktionsintervaller. De to niveauer (80% og 95%) angiver forskellig grad af usikkerhed.

7.3 Centrale prognoseværdier

```
forecast_best |>
  group_by(kon, region) |>
  summarise(prognose2020 = mean(.mean, na.rm = TRUE))
```

```
# A tsibble: 120 x 4 [1M]
# Key:      kon, region [10]
# Groups:   kon [2]
   kon      region      yearmonth prognose2020
  <fct>   <fct>         <mth>         <dbl>
1 Kvinder Region Hovedstaden 2020 Jan          1.92
2 Kvinder Region Hovedstaden 2020 Feb          1.92
3 Kvinder Region Hovedstaden 2020 Mar          1.87
4 Kvinder Region Hovedstaden 2020 Apr          1.73
5 Kvinder Region Hovedstaden 2020 May          1.80
6 Kvinder Region Hovedstaden 2020 Jun          1.86
7 Kvinder Region Hovedstaden 2020 Jul          1.82
8 Kvinder Region Hovedstaden 2020 Aug          1.91
9 Kvinder Region Hovedstaden 2020 Sep          1.86
10 Kvinder Region Hovedstaden 2020 Oct          1.86
# i 110 more rows
```

Med udgangspunkt i validerede modeller er der nu genereret forecasts for 2020. I næste afsnit analyseres og sammenlignes de centrale resultater på tværs af køn og regioner.

8 Sammenligning og fortolkning

8.1 Sammenligning på tværs af serier

8.2 Tendenser mellem regioner og køn

9 Konklusion

10 Kildeliste

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>

11 Bilagsoversigt

- Bilag 1: Udliveredede Powerpointpræsentationer v. Michael Freundlich: Chefkonsulent, Erhvervsservice og facility