

Arbejdsløshed - Ugeeksamen i Forecasting

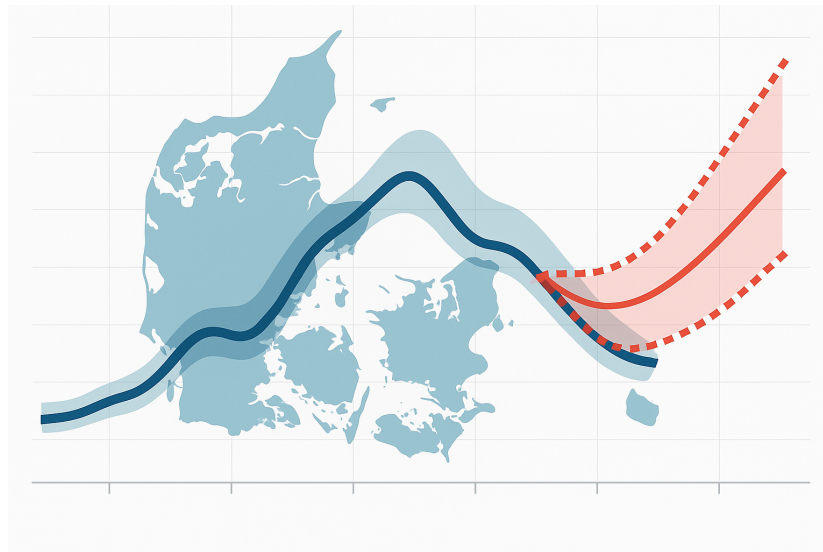
Christine Hegelund

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	2
2.1.1 Anvendelse af AI-værktøj	3
2.2 Definitioner/forkortelser	3
2.3 Struktur	3
3 Data og forberedelse	3
4 Eksplorativ dataanalyse (EDA)	4
4.1 Visualisering af udvikling og mønstre	4
4.1.1 Udvikling over tid	4
4.1.2 Sæsonmønstre	6
4.1.3 Regionale og kønsmæssige forskelle	8
4.2 Deskriptive statistikker	9
4.2.1 Features	10
4.3 STL-dekomposition	11
5 Modelvalg	13
5.1 Valgte modeltyper	13
5.2 Modellering i praksis	14
6 Modelvalidering	15
6.1 Trænings- og testsplit	16
6.2 Evalueringsmetrikker: RMSE og MAPE	16
6.3 Valg af bedste model pr. serie	16
6.4 Visualisering af forecast vs. faktiske værdier	17
6.5 Time series cross-validation	18
6.6 Residualanalyse og hvid støj	18

7	Forecasting	20
7.1	Estimering og forecast på 2020	20
7.2	Visualisering med prædiktionsintervaller	20
7.3	Centrale prognoseværdier	22
8	Sammenligning og fortolkning	22
8.1	Sammenligning på tværs af serier	22
8.2	Tendenser mellem regioner og køn	22
9	Konklusion	22
10	Kildeliste	23
11	Bilagsoversigt	24

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 undersøger udviklingen i arbejdsløsheden i Danmark i perioden 2007 til 2019 – opdelt på køn og region. Datagrundlaget består af ti tidsserier, én for hver kombination af fem regioner og to køn. Det giver mulighed for at analysere både regionale forskelle og kønsspecifikke mønstre i arbejdsløsheden.

Med udgangspunkt i klassiske tidsseriemodeller, ARIMA, ETS og en simpel benchmarkmodel (Seasonal Naive), undersøges, hvordan arbejdsløsheden har udviklet sig, og hvordan den kan fremskrives. Undervejs vurderes modellernes egnethed med fokus på præcision og residualstruktur, og der anvendes teknikker som STL-dekomposition og transformation af data. Målet er ikke blot at forudsige udviklingen i 2020, men også at vurdere, hvor godt klassiske modeller formår at håndtere forskellene mellem regioner 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 usikkerheden på tværs af serier?

2.1 Afgrænsning

Analysen bygger udelukkende på de udleverede arbejdsløshedsdata for perioden januar 2007 til december 2019 og inddrager ikke eksterne forklarende faktorer som COVID-19, konjunkturudsving eller politiske tiltag. Formålet er ikke at

forklare årsager til ledighed, men at undersøge, hvordan klassiske forecasting-modeller kan anvendes metodisk og reproducerbart. Fokus ligger på statistisk modellering og sammenligning af modelpræcision – ikke på strategiske eller politiske anbefalinger.

2.1.1 Anvendelse af AI-værktøj

ChatGPT (GPT-4o) er anvendt som støtteværktøj i forbindelse med idéudvikling, sproglig formulering og udformning af enkelte kodestumper. Værktøjet er kun brugt til tekniske og sproglige formål – al analyse, fortolkning og konklusion er udarbejdet selvstændigt af gruppens medlemmer.

2.2 Definitioner/forkortelser

Nedenfor er en oversigt over centrale begreber og forkortelser, som anvendes i opgaven:

- **ARIMA:** AutoRegressive Integrated Moving Average
- **ETS:** Exponential Smoothing State Space Model
- **SNAIVE:** Seasonal Naive
- **RMSE:** Root Mean Squared Error
- **MAPE:** Mean Absolute Percentage Error
- **STL:** Seasonal-Trend decomposition using Loess
- **CV:** Cross-validation

2.3 Struktur

Opgaven er opbygget i fem hovedfaser, som følger en klassisk tilgang til analyse af tidsserier. Analysen er opdelt i fem overordnede faser. Først gennemføres en eksplorativ dataanalyse, hvor arbejdsløsheden undersøges for tendenser, sæsonmønstre og forskelle mellem køn og region. Derefter estimeres tre modeller – ARIMA, ETS og Seasonal Naive – for hver tidsserie. I tredje fase valideres modellerne med krydsvalidering og residualanalyse, og deres præcision vurderes. På den baggrund fremskrives arbejdsløsheden for 2020. Til sidst sammenlignes resultaterne, og der konkluderes på tværs af modeller, regioner og køn.

3 Data og forberedelse

Analysen bygger på månedlige arbejdsløshedstal fra Danmark i perioden januar 2007 til december 2019. Data er opdelt efter køn og region, hvilket giver ti separate tidsserier. Datasættet er udleveret i et forbehandlet tsibble-format med tydeligt definerede indeks- og nøglevariabler, og nedenfor indlæses datasættet:

```
# 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)

Formålet med dette afsnit er at skabe et overblik over datasættets struktur forud for modelleringen. Ved hjælp af visualiseringer, deskriptive statistikker og STL-dekomposition undersøges arbejdsløshedens udvikling på tværs af regioner og køn i perioden 2007–2019. Analysen afdækker overordnede tendenser, sæsonmønstre og forskelle i niveau og variation. For at sikre god modellering vurderes behovet for datatransformation, og tidsrækkerne dekomponeres i trend, sæson og residualer. Resultaterne danner det metodiske fundament for det videre modelarbejde.

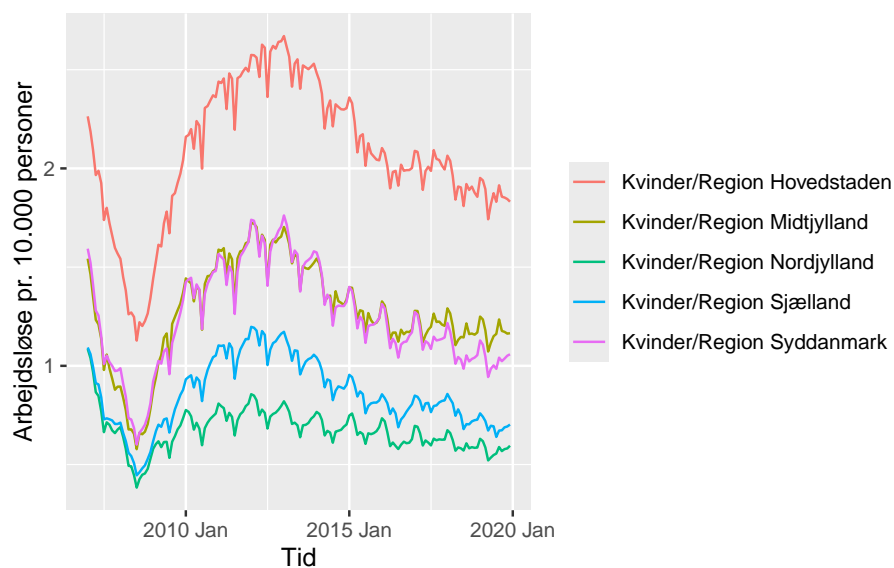
4.1 Visualisering af udvikling og mønstre

I dette afsnit undersøges, hvordan arbejdsløsheden har udviklet sig over tid – med fokus på tendenser, sæsonvariationer og forskelle mellem regioner og køn. Formålet er at identificere mønstre, som kan guide valg af modeller i den videre analyse.

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



Ovenstående figur 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. Niveaueet varierer regionalt, hvor Region Hovedstaden typisk ligger højest og Nordjylland lavest. Samlet ses en faldende tendens efter finanskrisen.

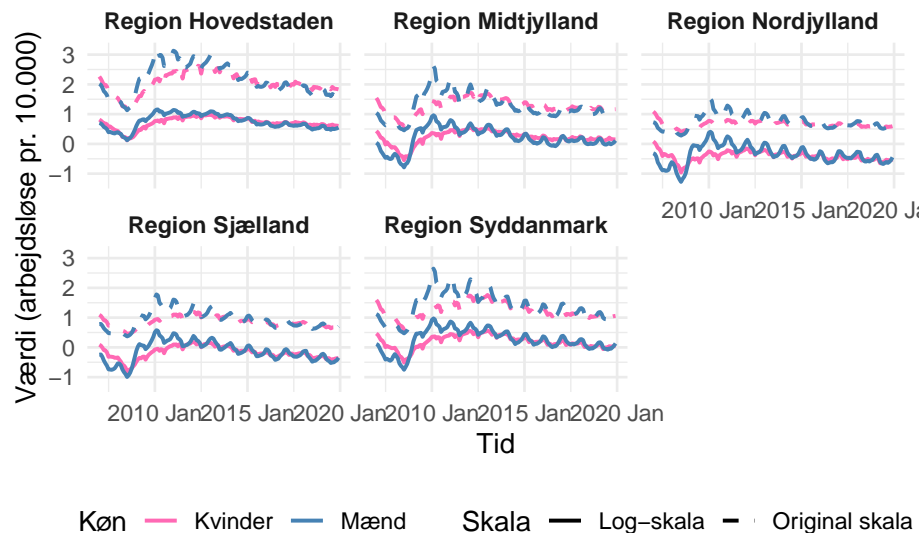
Figur 2: Udvikling i mænds arbejdsløshed pr. region



Ud fra Figur 1 og 2 ses, at mænds arbejdsløshed generelt har større udsving

og mere strukturelle skift end kvinders. Dette bliver tydeligere, når serierne ses både i deres oprindelige og log-transformerede form, hvilket ses i figur 3,

Figur 3: Sammenligning af original og log-skala



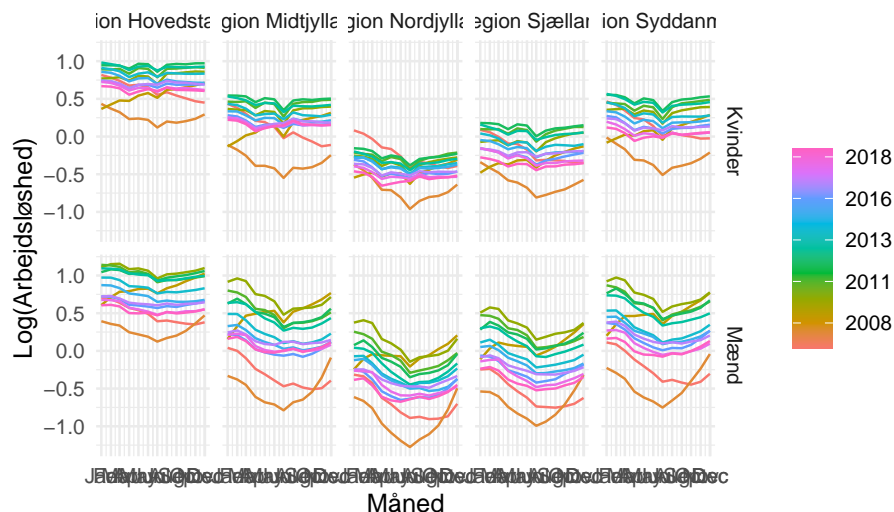
Figur 3 viser, hvordan log-transformeringen dæmper udsvingene, særligt i serier med højt niveau og stor variation, som det ses blandt mænd. Denne transformation gør det lettere at sammenligne udviklingen på tværs af regioner og køn og sikrer en mere stabil varians i det videre analysearbejde. Derfor anvendes den log-transformerede version fremadrettet i analysen.

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.

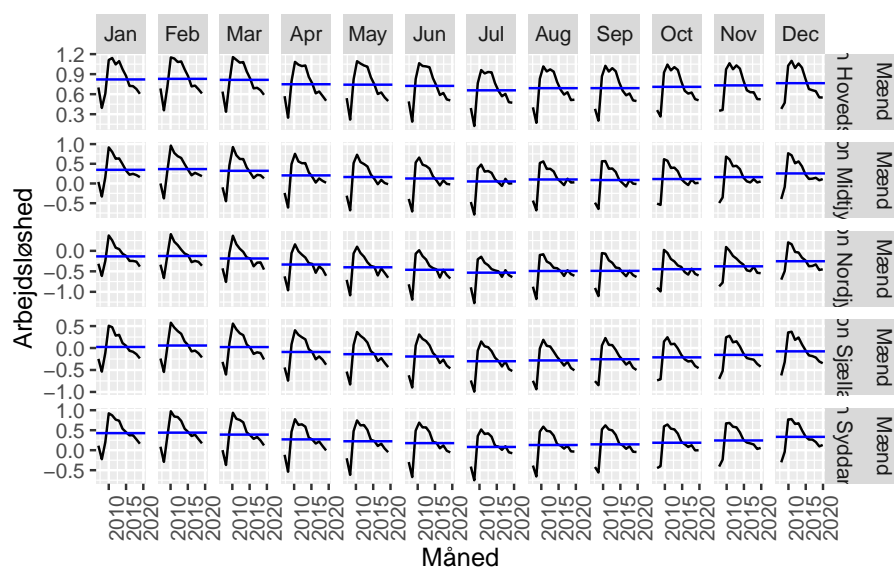
Figur 4: Sæsonmønstre i arbejdsløsheden pr. region og køn (log-skala)

Figur 4: Sæsonmønstre i arbejdsløsheden pr. region og køn



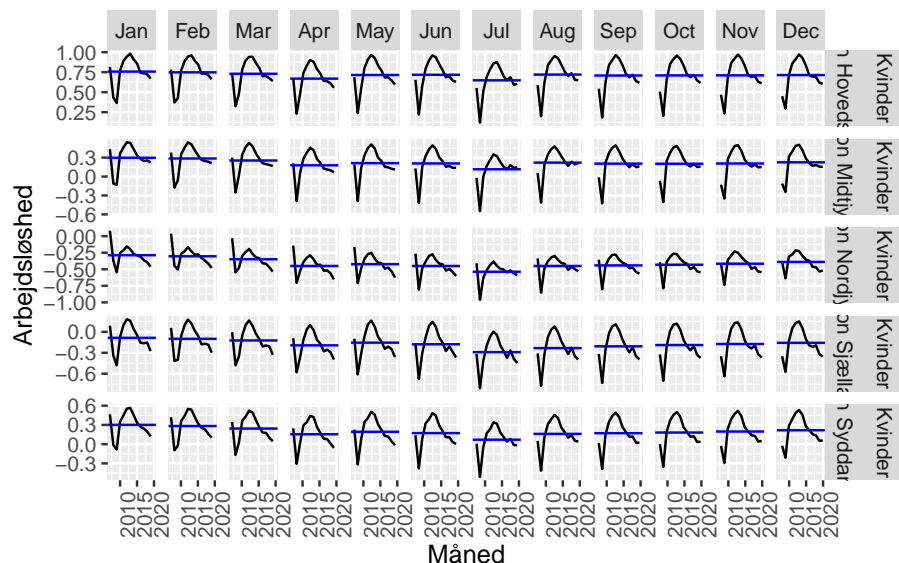
Figur 4 viser sæsonmønstre i arbejdsløsheden pr. region og køn (log-transformeret). Hver farve repræsenterer et år, og linjerne viser en tydelig sæsonrytme med høj ledighed i årets begyndelse og lav i sommermånederne. Mænd udviser generelt større sæsonudsving end kvinder. Niveaut er højest i Region Hovedstaden og lavest i Nordjylland og Sjælland. Negative log-værdier forekommer, hvor arbejdsløsheden ligger under 1 pr. 10.000.

Figur 5: Subserieplot for mænd pr. måned og region



Figur 5 viser subserieplottet for arbejdsløshedens udvikling for mænd pr. måned og region. Hver celle viser forløbet for en specifik måned over årene, med sorte linjer for de enkelte år og blå linjer for gennemsnit. Ledigheden topper typisk i årets første måneder og falder hen over sommeren. Mænd udviser generelt større sæsonudsving og højere variation mellem år – især i vinterhalvåret.

Figur 6: 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.

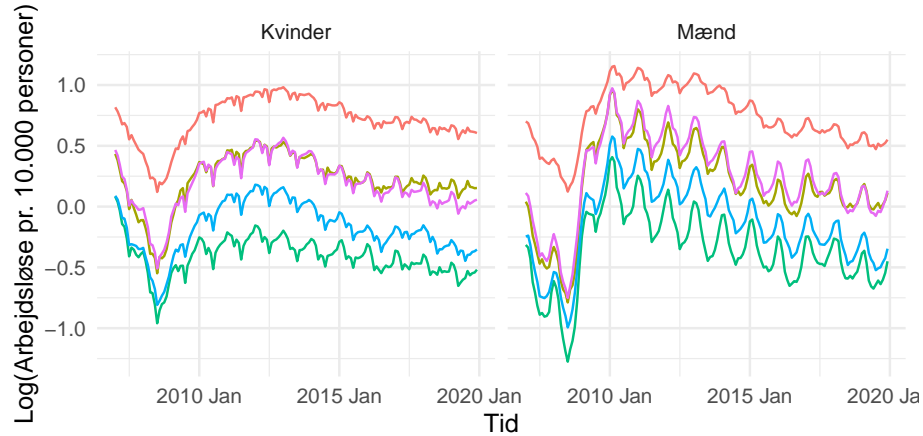
Udover sæsonmæssige udsving varierer arbejdsløsheden betydeligt på tværs af både geografi og køn. Dette undersøges nærmere i de følgende visualiseringer.

4.1.3 Regionale og kønsmæssige forskelle

I dette afsnit undersøges samspillet mellem køn og region i arbejdsløshedens udvikling. Målet er at vurdere, hvordan niveau og variation adskiller sig på tværs af grupper over tid.

Figur 7: Arbejdsløshed fordelt på regioner – opdelt efter køn (log-skala)

Figur 7: Udvikling i arbejdsløsheden (log-skala), fordelt på reg



on Hovedstaden — Region Midtjylland — Region Nordjylland — Region Sjælland — F

Figur 7 viser, hvordan arbejdsløsheden varierer mellem regioner for hvert køn og gør det muligt at sammenligne regionale niveauer og mønstre. Kønsforskellene inden for hver region blev tidligere visualiseret i Figur 3, hvor udviklingen for mænd og kvinder blev sammenlignet direkte. Tilsammen giver figurerne et samlet overblik over både geografiske og kønsmæssige forskelle i arbejdsløsheden.

For at underbygge disse visuelle indsigter med konkrete tal, præsenteres i næste afsnit deskriptive statistikker. De giver en numerisk opsummering af centrale mål som gennemsnit, variation og ekstreme værdier og bidrager til at identificere særligt stabile eller volatile grupper, der kan kræve særlig opmærksomhed i den videre analyse.

4.2 Deskriptive statistikker

De foregående visualiseringer viste klare forskelle i arbejdsløshedens niveau, variation og sæsonmønstre på tværs af regioner og køn (Figur 1–7). For at supplere disse observationer med en kvantitativ opsummering præsenteres i Tabel 1 centrale deskriptive mål: gennemsnit, standardafvigelse, minimum og maksimum. Tabellen giver et hurtigt overblik over niveau og variation og fremhæver fx Nordjylland som en stabil region – særligt blandt kvinder, mens andre kombinationer viser højere udsving.

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

4.2.1 Features

Selvom deskriptive statistikker giver et første indtryk af variation og niveau, indfanger de ikke strukturelle træk som trend, sæsonmønstre og stationaritet. For at kvantificere disse egenskaber beregnes en række såkaldte features ved hjælp af `features()`-funktionen fra `feasts`-pakken (Hyndman & Athanasopoulos, 2021, kap. 4).

Tabel 2 opsummerer udvalgte funktioner, der kvantificerer centrale træk ved serierne og supplerer de visuelle indsigter fra Figur 1–7.

```
# A tibble: 10 x 8
  kon      region acf1 seasonal_strength_year var_tiled_mean shift_level_max
  <fct>   <fct>   <dbl>          <dbl>          <dbl>          <dbl>
1 Kvinder Region H~ 0.970          0.652          0.972          0.381
2 Kvinder Region M~ 0.957          0.720          0.920          0.535
3 Kvinder Region N~ 0.890          0.873          0.726          0.465
4 Kvinder Region S~ 0.961          0.814          0.917          0.429
5 Kvinder Region S~ 0.955          0.833          0.911          0.454
6 Mænd    Region H~ 0.983          0.810          0.977          0.580
7 Mænd    Region M~ 0.975          0.840          0.925          1.05
8 Mænd    Region N~ 0.959          0.908          0.802          0.893
9 Mænd    Region S~ 0.973          0.918          0.896          0.804
10 Mænd    Region S~ 0.975          0.884          0.914          0.957
# i 2 more variables: ndiffs <int>, trend_strength <dbl>
```

Tabel 2 præsenterer en række kvantitative tidsseriefunktioner, der understøtter og uddyber de visuelle observationer fra Figur 1–7. Nedenfor gennemgås hovedpointerne opdelt i fem centrale kategorier: afhængighed, sæsonmønstre, variation og skift, stationaritet og trend.

Afhængighed

`acf1` måler korttidsafhængighed og viser høje værdier på tværs af regioner (typisk 0.96–0.98). Dette indikerer stærk autokorrelation, hvilket stemmer overens med de jævne og forudsigelige bevægelser set i Figur 1–3 og 7.

Sæsonmønstre

`sasaonal_strength_year` viser, at mænd generelt har mere udtalte sæsonmønstre end kvinder. Værdien er fx 0.91 i Region Nordjylland og 0.92 i Sjælland blandt mænd – grupper, der også i Figur 4–6 fremstod med klare sæsonudsving. Kvinder i Region Hovedstaden har en lavere værdi på 0.65, hvilket afspejler mere stabile mønstre.

Variation og skift

`var_tiled_mean` er lavest blandt kvinder i Region Nordjylland (0.73), hvilket indikerer begrænset variation over tid – i tråd med de stabile kurver i Figur 3 og 7. Derimod er `shift_level_max` højest for mænd i Region Midtjylland (1.05), hvilket peger på pludselige ændringer, som fx under finanskrisen, og ses tydeligt i samme figurer.

Stationaritet

`ndiffs` angiver behovet for differensiering for at opnå stationaritet. De fleste serier kræver én differens, men enkelte, fx kvinder i Nordjylland og mænd i Midtjylland og Nordjylland, har værdi 0. Dette bekræfter deres stabile niveau, som også fremgår af Figur 3 og 5.

Trend

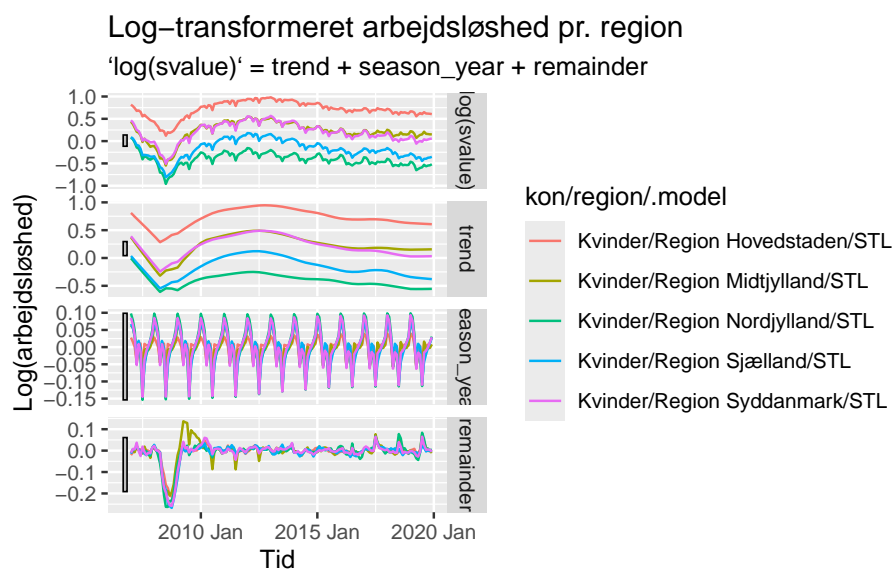
`trend_strength` er generelt høj (typisk 0.97–0.99) og viser tydelige trends i alle serier. Dette er i tråd med de markante udviklingsforløb i arbejdsløsheden over tid, især i Figur 1–3, hvor fx Region Hovedstaden skiller sig ud.

Samlet set supplerer disse features de visuelle observationer med numerisk dokumentation og danner et vigtigt udgangspunkt for videre modellering.

4.3 STL-dekomposition

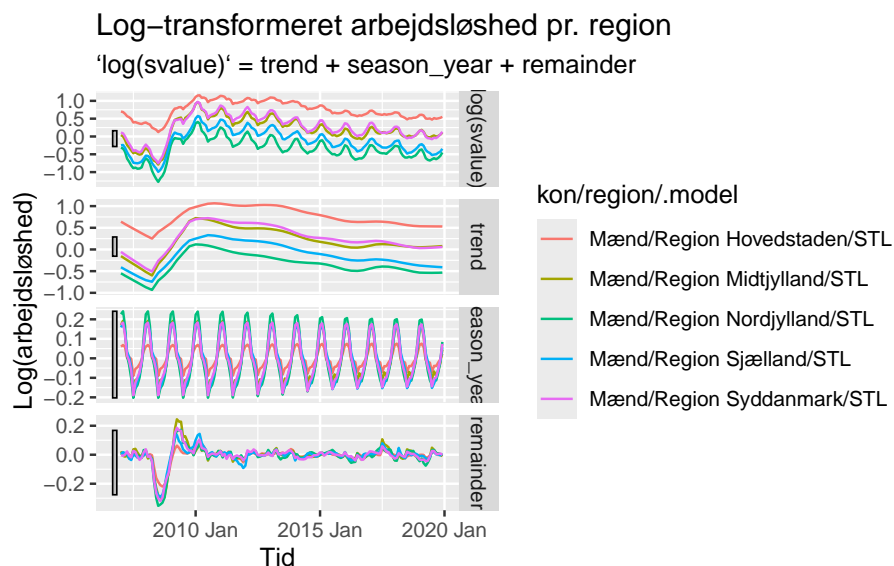
STL-dekomposition anvendes for at undersøge de strukturelle komponenter i arbejdsløshedsserierne. Metoden opdeler tidsserierne i tre dele: trend, sæson og remainder, hvilket giver indblik i, hvor stor en del af variationen der skyldes henholdsvis langsigtet udvikling, tilbagevendende mønstre eller kortsigtede udsving (Hyndman og Athanasopoulos, 2021). Dekompositionen er udført separat for mænd og kvinder i hver region, og resultaterne vises 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 (Hyndman & Athanasopoulos, 2021, 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 (Hyndman & Athanasopoulos, 2021, kap. 8).

SNAIVE (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 (Hyndman & Athanasopoulos, 2021, 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. Præcisionen 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.

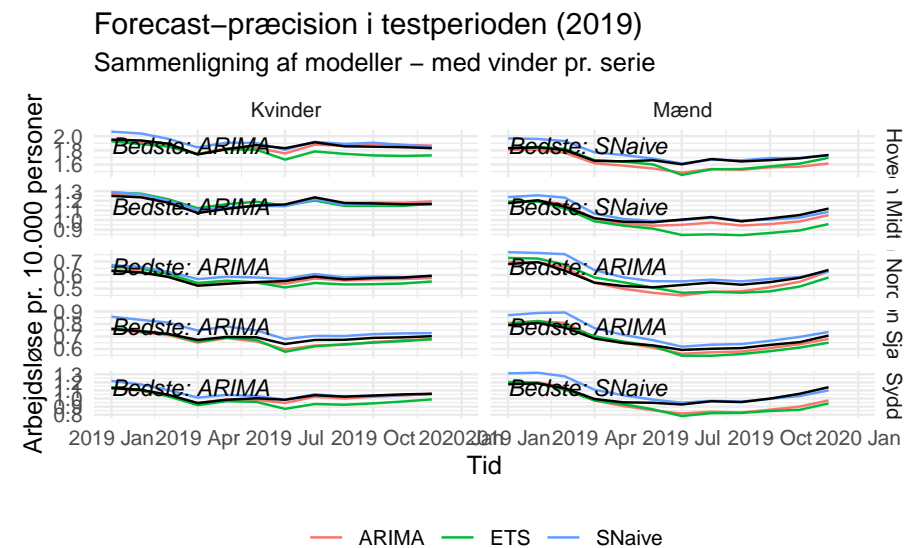
```
# A tibble: 10 x 5
  kon      region      .model  RMSE  MAPE
```

	<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):

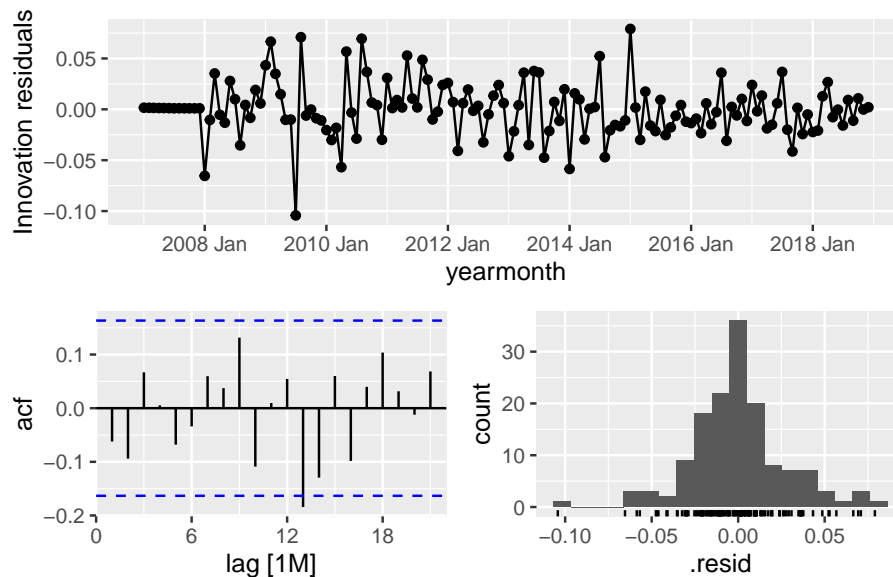
```
# A tibble: 10 x 5
  kon      region      .model  RMSE  MAPE
  <fct>   <fct>         <chr>   <dbl> <dbl>
1 Kvinder Region Hovedstaden ARIMA  0.0376 1.30
2 Kvinder Region Midtjylland ARIMA  0.0250 1.44
3 Kvinder Region Nordjylland ETS    0.0185 2.27
4 Kvinder Region Sjælland   ARIMA  0.0170 1.38
5 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 ek-

sempel 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 fejleddene. 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 ukorrigerede 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

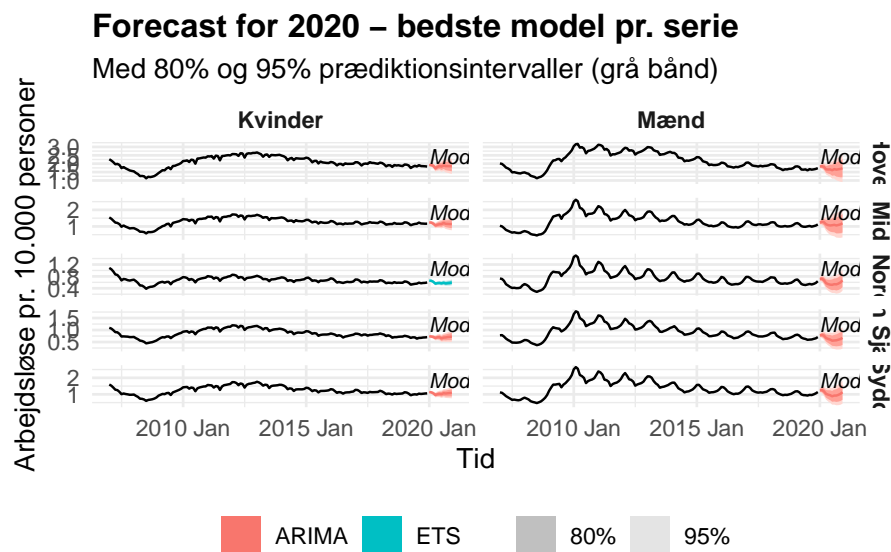
Forecasts for 2020 visualiseres for alle serier med prædiktionsintervaller på 80% og 95%. Den bedste model for hver serie fremgår af farven på forecast-linjen og angives desuden som label i hvert facet.

```

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