



LUNDS UNIVERSITET

Inlämningsuppgift Tidsserieanalys
Undersökning av temperaturdata i Östersund

Croneborg, Claes 19981015-7199

STAH14

Lärare: Peter Gustafsson

30 januari 2023

Innehåll

1	Kortsiktig prognos	3
2	Långsiktigt prognos	8
2.1	Inledande analys av datamaterialet	8
2.2	Modell och forecasting	10
3	Förändringar över tidsperioden	13
4	Slutsats	18
A	Kort-data	19
B	Lång-data	23

Inledning

I denna undersökning kommer statistiska metoder användas för att analysera temperaturen som uppmäts på Östersunds-flygplats. Vi kommer bryta ned serien i dess modellkomponenter med hjälp av *stl()* och *decompose()*-funktionerna för att skilja på den deterministiska trenden $\mu(t)$ och stokastiska komponenten $X(t)$, alltså:

$$Y(t) = \mu(t) + X(t)$$

För att identifiera en stationär process med konstant medelvärde och varians undersöks detta genom ett Augmented Dickey-Fuller (ADF)-test. Om processen inte är stationär måste den istället justeras, vilket kan göras genom att differentiera serien:

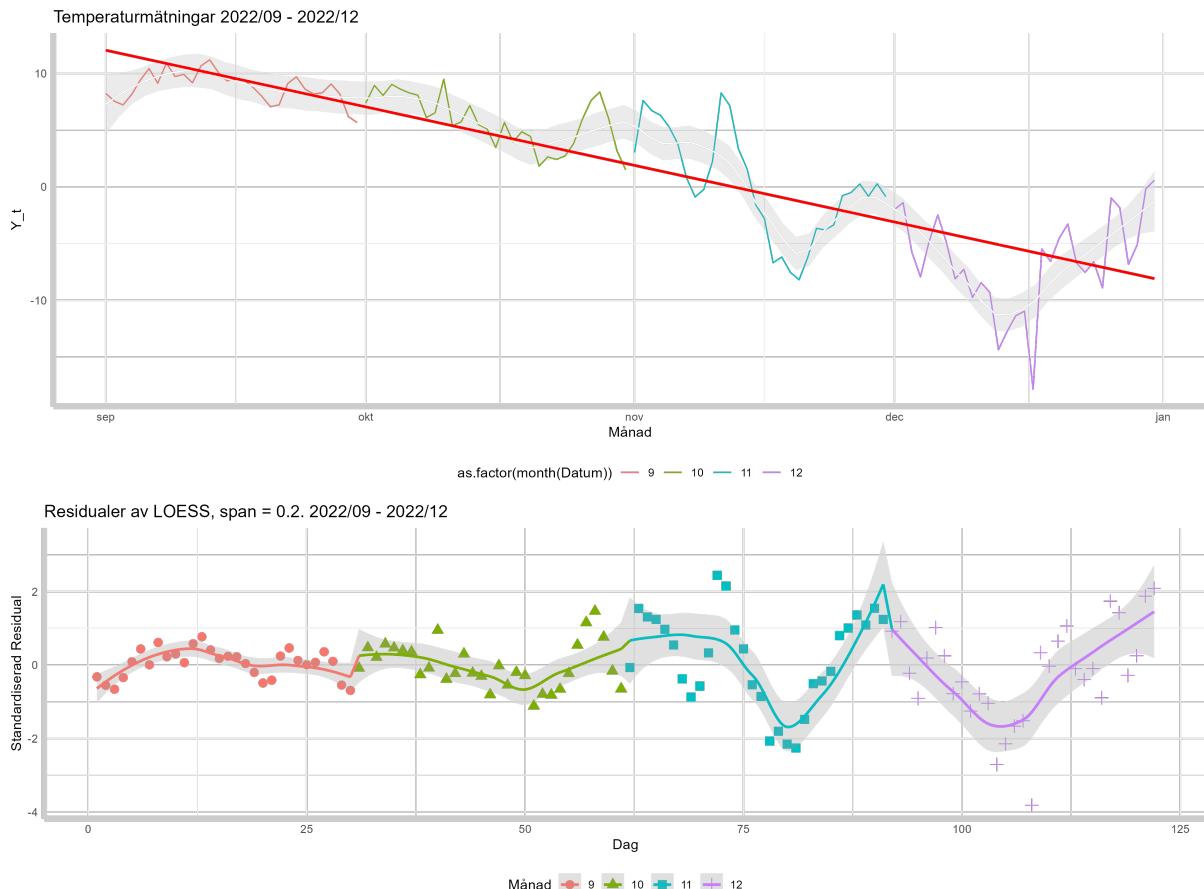
$$y'_t = y_t - y_{t-1}$$

För att sedan kunna validera modellen behöver modellen uppfylla antagandena: homoskedastiska residualer, obeorende residualer och även att de inte är autokorrelerade. Undersökningen tittar på två olika tidshorisonter, en kortiktig med 14-dagars prognos och en långsiktig sedan första temperaturmätning år 1950. Avslutningsvis kommer flera modeller jämföras där den bästa modellen med bäst prediktionsförmåga har lägst PRESS-värde:

$$PRESS = \sum_{i=1}^k (Y_i - \hat{Y}_i)^2$$

1 Kortsiktig prognos

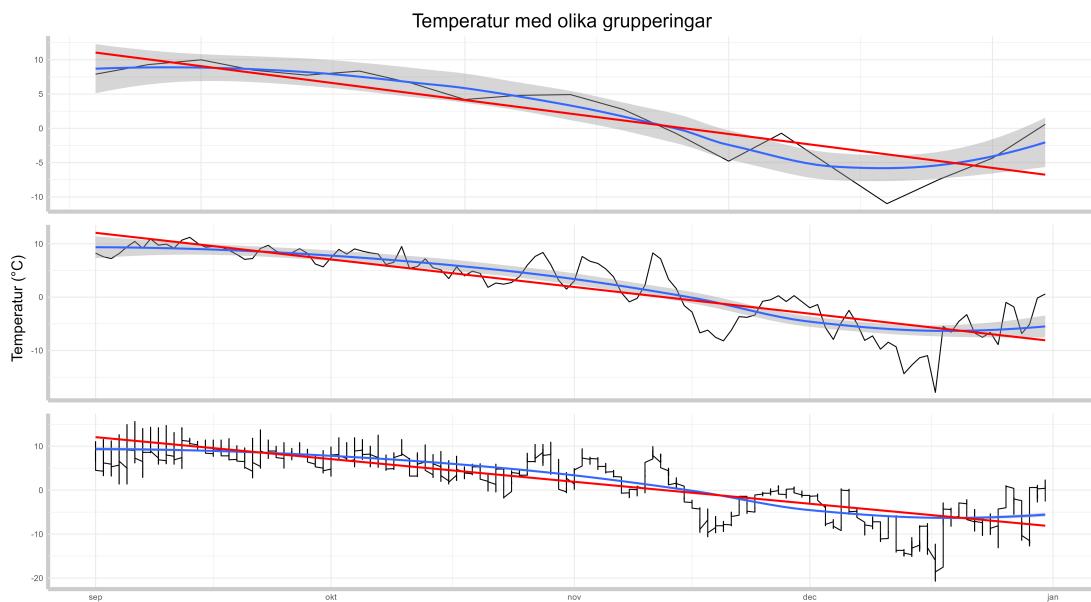
Datamaterialet som används för den kortsiktiga prognosen sträcker sig från September till årsskiftet. Materialet består av timvisa mätningar vilket under undersökningens gång kan och kommer grupperas på olika sätt. För att få en inledande förståelse av datans struktur och mönster visualiseras den genomsnittliga temperaturen per dag i figur 1.1. I huvudsak



Figur 1.1: Översikt av temperatur från 09/22 - 12/22

kommer datamaterialet undersökas med tre olika frekvenser. Timvis data, data grupperad per dag och data grupperad per vecka. De två sistnämnda processerna kommer således vara ett medelvärde av temperaturerna per dag respektive vecka. Den timvisa serien ger mest information, men veckodata kan ge en bättre skattning av en eventuell trend och dagliga data ger ett mellanting. Slutligen kommer de olika modellerna jämföras för att se huruvida skattningarna och modellens prediktionsförmåga varierar beroende på mängden data. Inledningsvis kommer vi undersöka den deterministiska modellkomponenten $\mu(t)$ i syfte att upptäcka en trend i data om en sådan finns. När vi har grupperat dataen på de olika sättens ser vi att såväl intercept som Tid är signifikanta för alla modeller. Vi ser även att förklaringsgraden, alltså R^2 , är bäst för regressionen med veckovisdata. Efter ha studerat figur 1.2 kan vi ganska snabbt inse att de genomsnittliga avvikelserna från

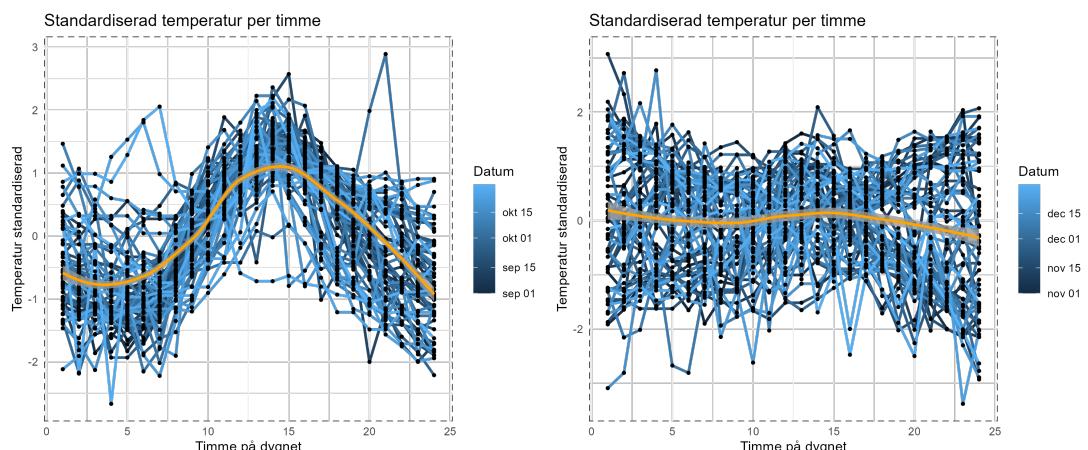
den linjära regression kommer vara allt högre desto fler datapunkter modellen 'behöver passa'. Enligt residualplottarna A.1 & A.2 för den linjära regression ser vi tydliga mönster



Figur 1.2: Datamaterial per datagruppering

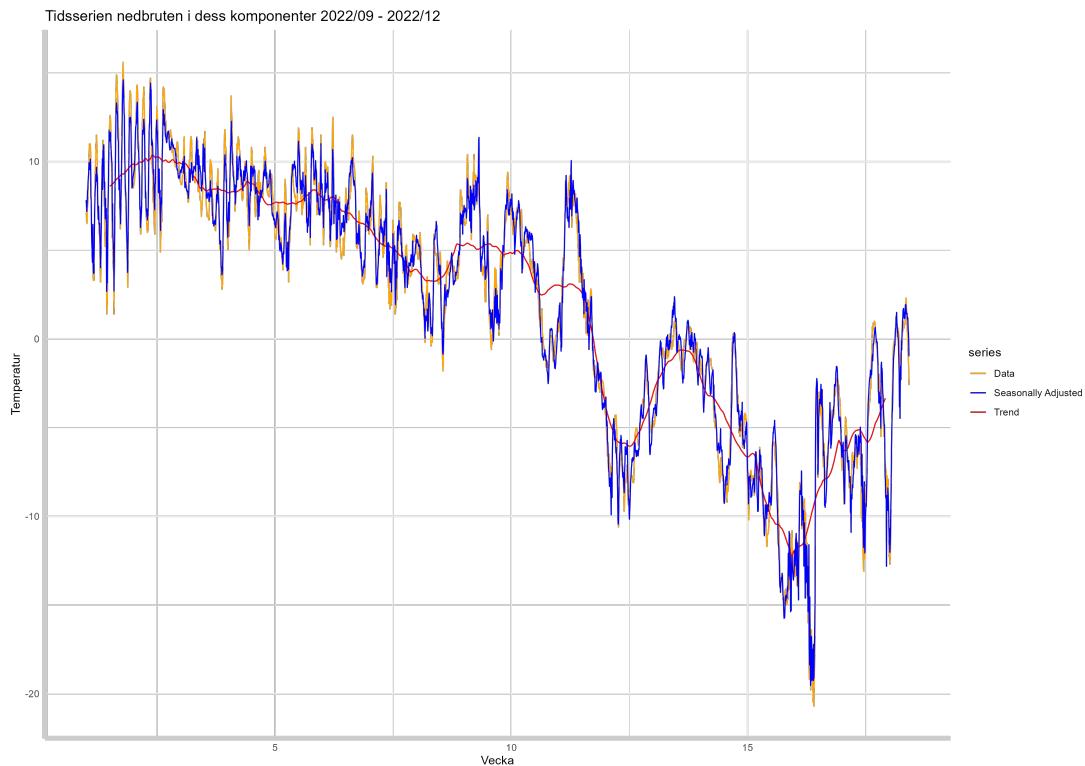
i residualerna. Att notera är däremot hur den veckovisa modellen verkar vara den mest stabila då residualerna i QQ-plotten följer den teoretiska linjen bättre än de andra. Med det sagt bryts antagandet om att residualerna är oberoende och modellen kan därför inte valideras.

Resultatet av ADF-testet i tabell A.2 visar att timvisa serien är den enda stationära. Det är därför nödvändigt att försöka göra processerna med dagsvis och veckovis där differentiering skapade en stationär dagvis tidsserie enligt tabell A.3. Till skillnad från den dagvisa tidsserien har den timvisa serien istället 24 observationer per dag, således finns också möjligheten att försöka lösgöra trend- och säsongskomponenten. För att illustrera den periodiska cykeln ser vi i figur 1.3 hur temperaturen varierade beronde på tiden på dygnet.



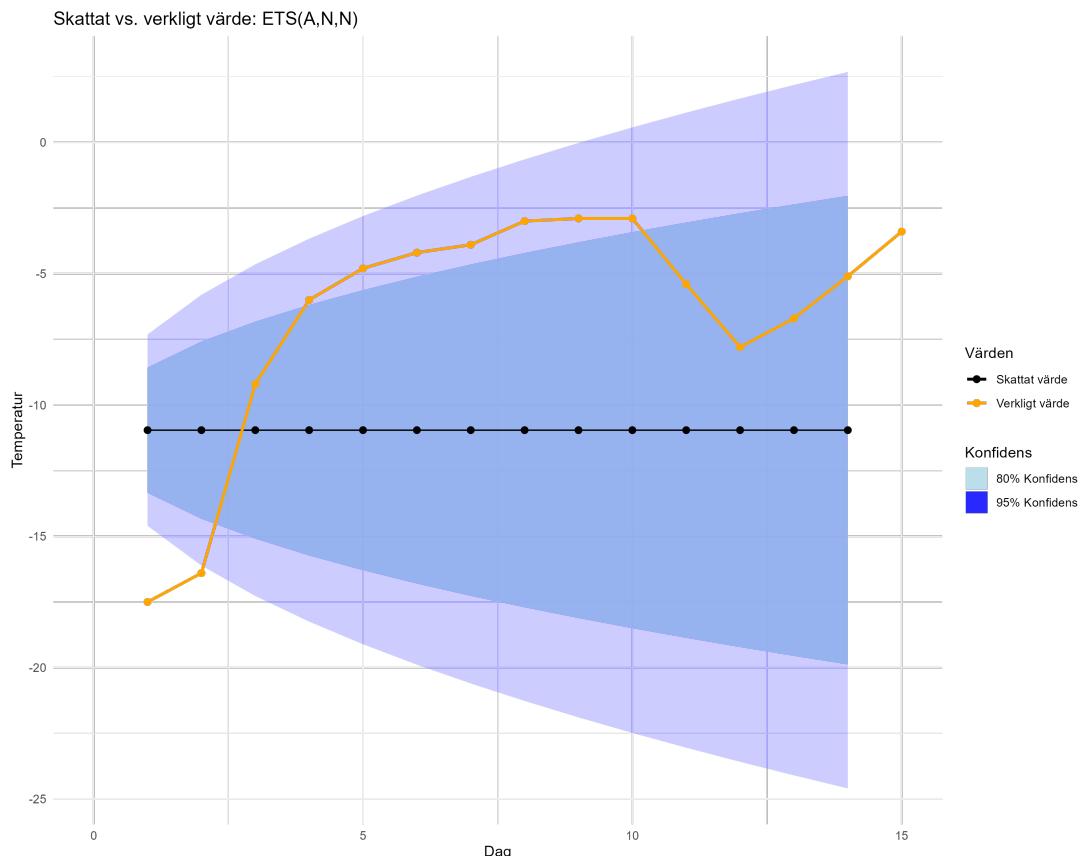
Figur 1.3: Säsongtrend timvis data, notera tidsintervall

Figuren är uppdelad i olika tidsintervall där vi kan se att den säsongsbetonade komponenten förändras desto senare på året mätningarna gjordes. Genom decompose()-funktionen extraheras de skattade trend- samt säsongsjusterad serie där de olika komponenterna visualiseras i figur 1.4.

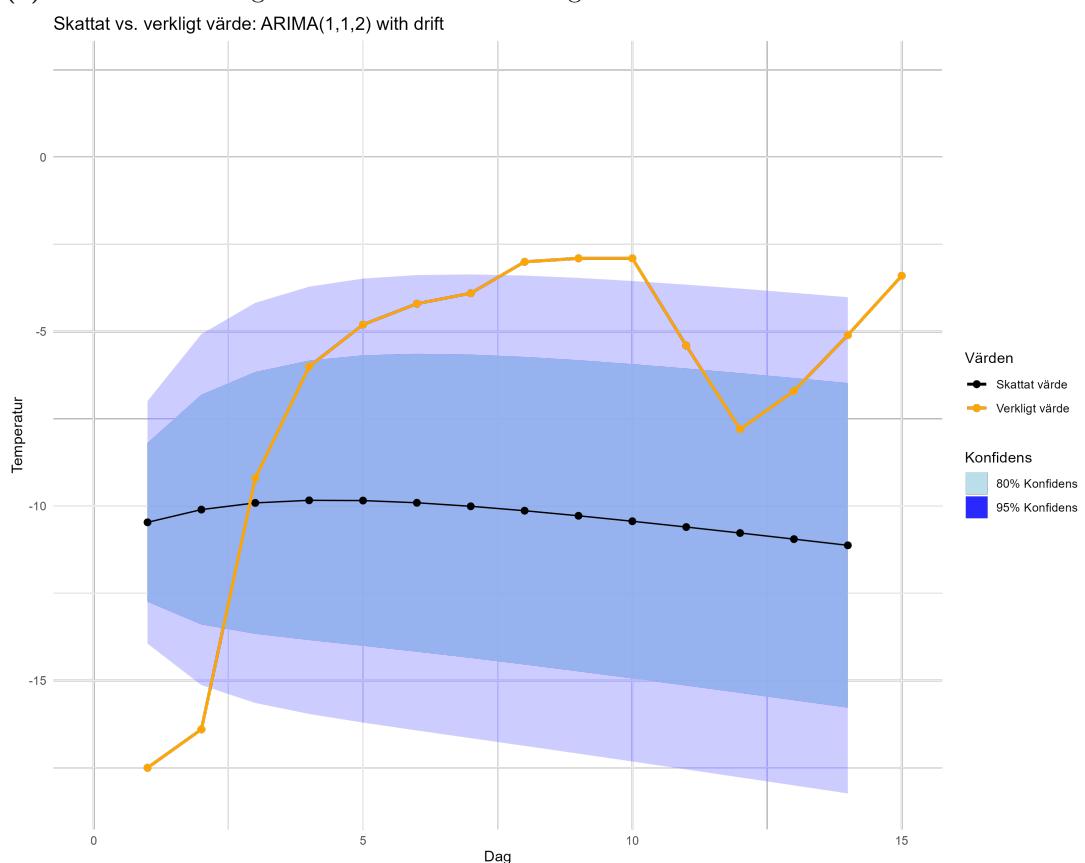


Figur 1.4: Timvis data modellkomponenter enligt de olika linjerna

När vi undersökt datamaterialet är nästa steg att bygga en modellen som passar datamaterialet bäst. Här kommer användas två olika metoder; ETS (Error, Trend, Seasonal) och ARIMA-modeller. ETS är en modell som använder exponential smoothing som kombinerar säsongsvariationer och trender för att förutsäga framtida värden. ARIMA använder autoregressiva integrerade moving average-modeller för att förutsäga framtida värden. Med hjälp av de inbyggda funktionerna `ets()` och `auto.arima()` i 'forecast'-paketet kan vi skatta flertalet modeller automatiskt. För att sedan kunna validera modellerna behöver antagandena om; Residualerna $\sim \mathcal{N}(0, \sigma^2)$ med konstant varians (*i.e.* homoskedasticitet) men även att de inte lider av autokorrelation. Den enda modell vilken uppfyller alla antagandena är ARIMA-modell på dagsvis data. För ETS-modellen är det en av punkterna som inte faller inom det 95% konfidensintervallet på ACF-plotten. Residualerna för ETS-modellen är både normalfördelade men även tillräckligt konstanta över tid, vilket ändå ger oss en anledning att jämföra modellernas prediktionsförmåga. Figur 1.5 illustrerar de två modellernas predikterade värde de 14 senaste dagarnas temperatur, gentemot de verkliga värdena. Det är tydligt att modellerna har svårt att generalisera på ny data då avvikelserna från det verkliga värdet är stora. Vi ser även att det predikterade intervallet ökar med tiden, alltså blir vår skattning mer osäker desto längre tidsperiod man predikterar.

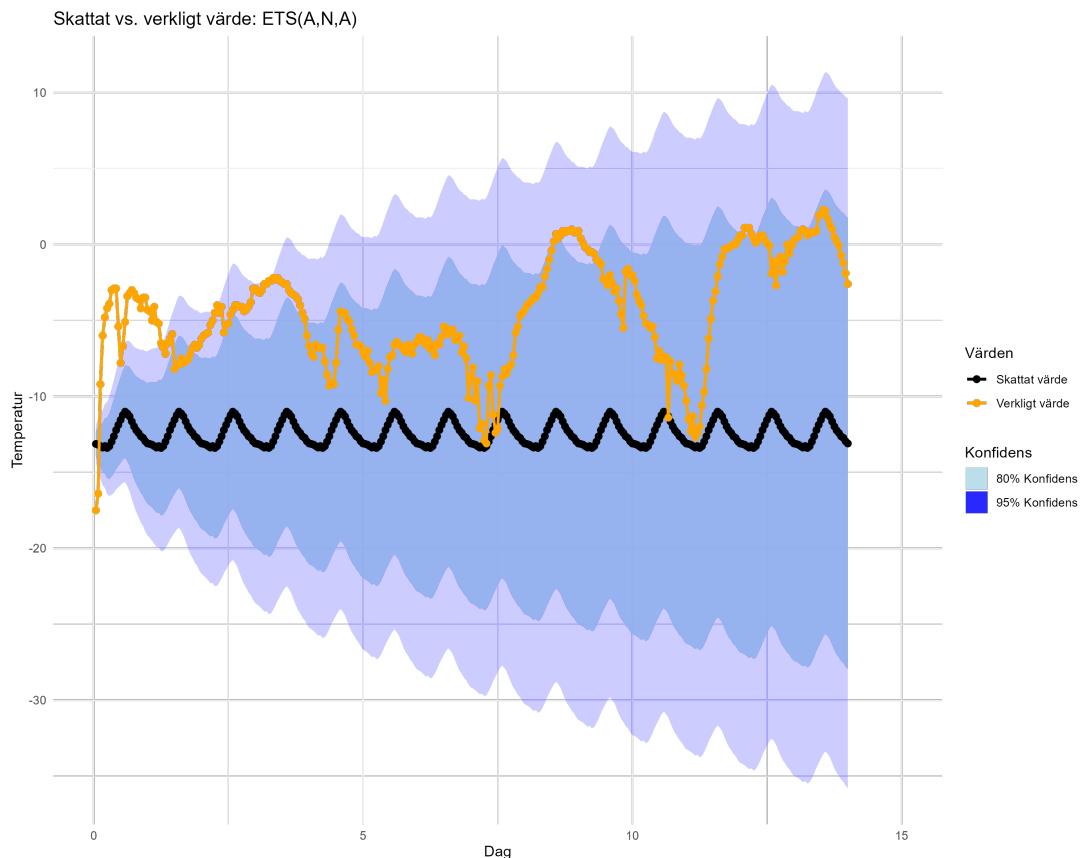


(a) Skattat vs. verkligt ETS-modell värde 14-dagar

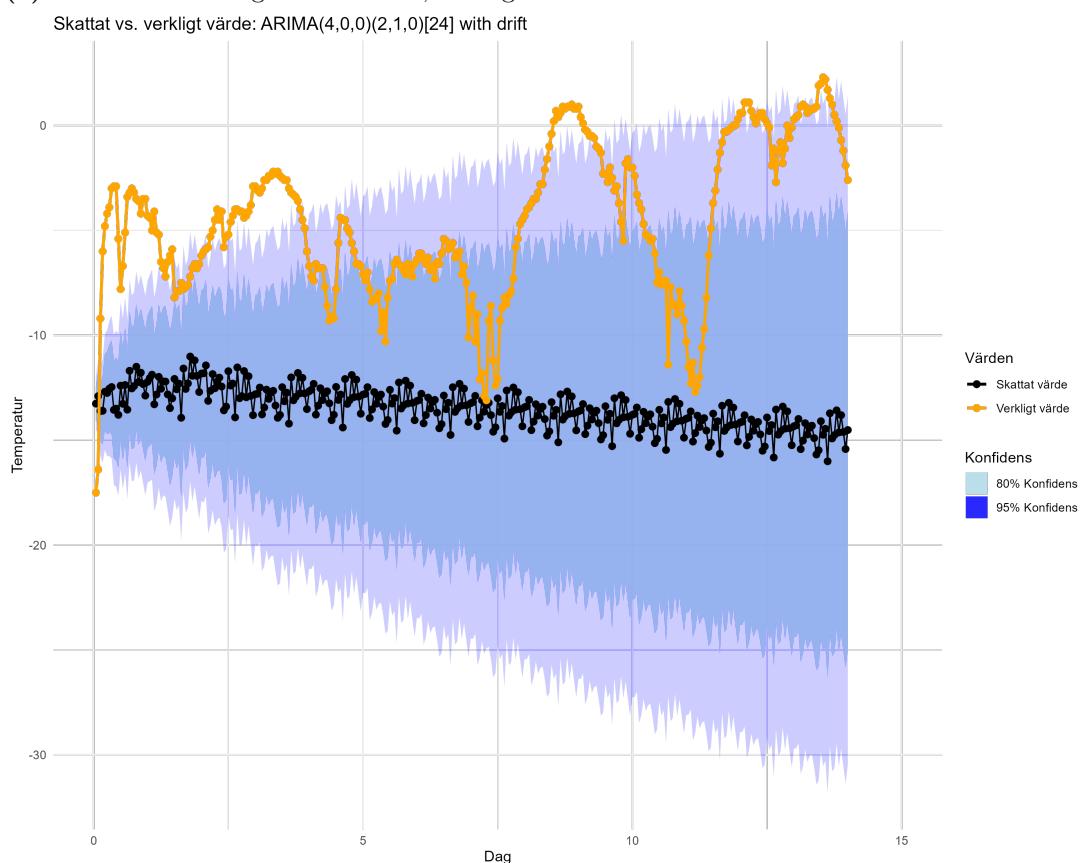


(b) Skattat vs. verkligt ARIMA-modell värde 14-dagar

Figur 1.5: Skattat vs. verkligt av bästa två modeller, ETS & ARIMA



(a) Skattat vs. verkligt ETS-modell, 14-dagar timvis data



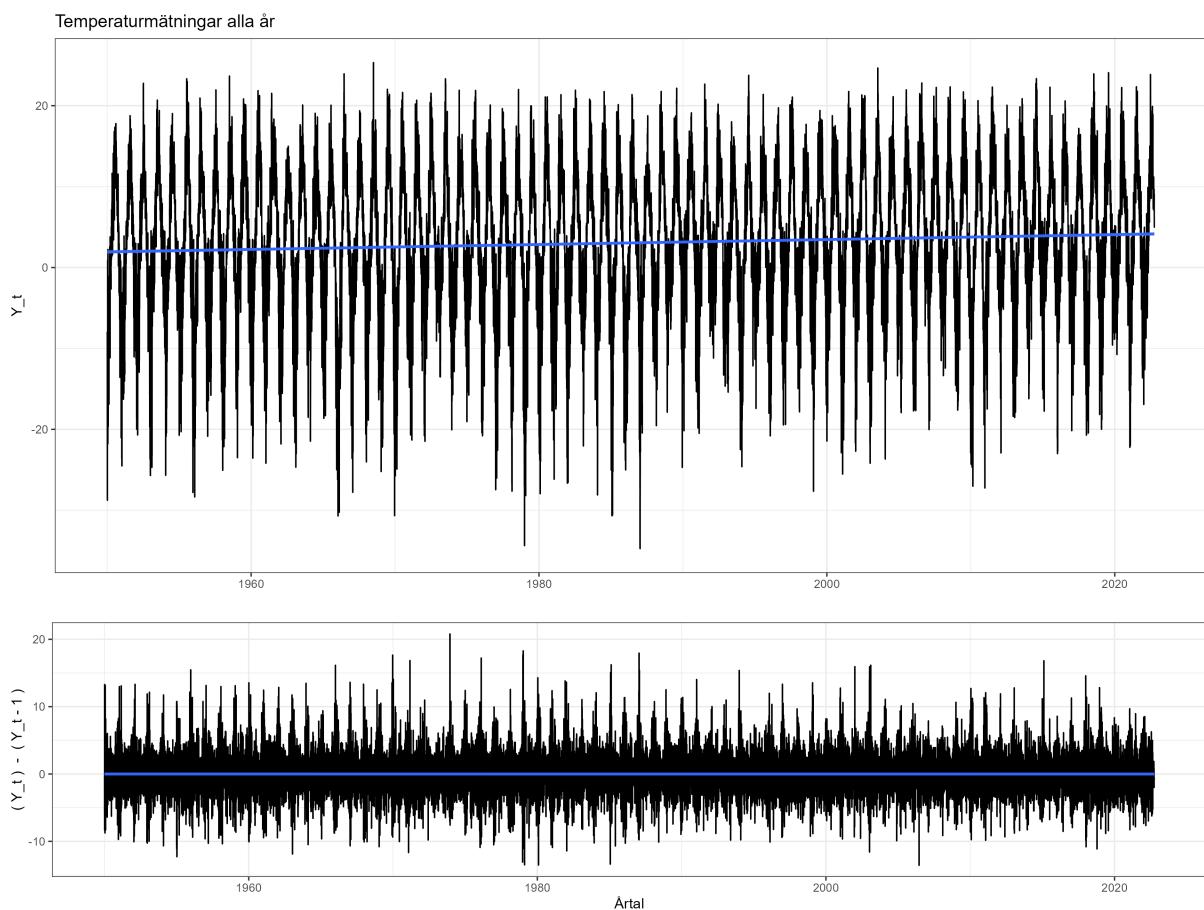
(b) Skattat vs. verkligt ARIMA-modell, 14-dagar timvis data

Figur 1.6: Skattat vs. verkligt timvis data, ETS & ARIMA

2 Långsiktigt prognos

2.1 Inledande analys av datamaterialet

Processen, eller tillvägagångssättet, kommer vara densamma när vi undersöker tidsserien med lång tidshorisont. Inledningsvis plottas temperaturen över alla år för att få en förståelse för datamaterialet och hur den kan komma att bete sig. Den styrlinje, alltså



Figur 2.1: Temperatur över alla år

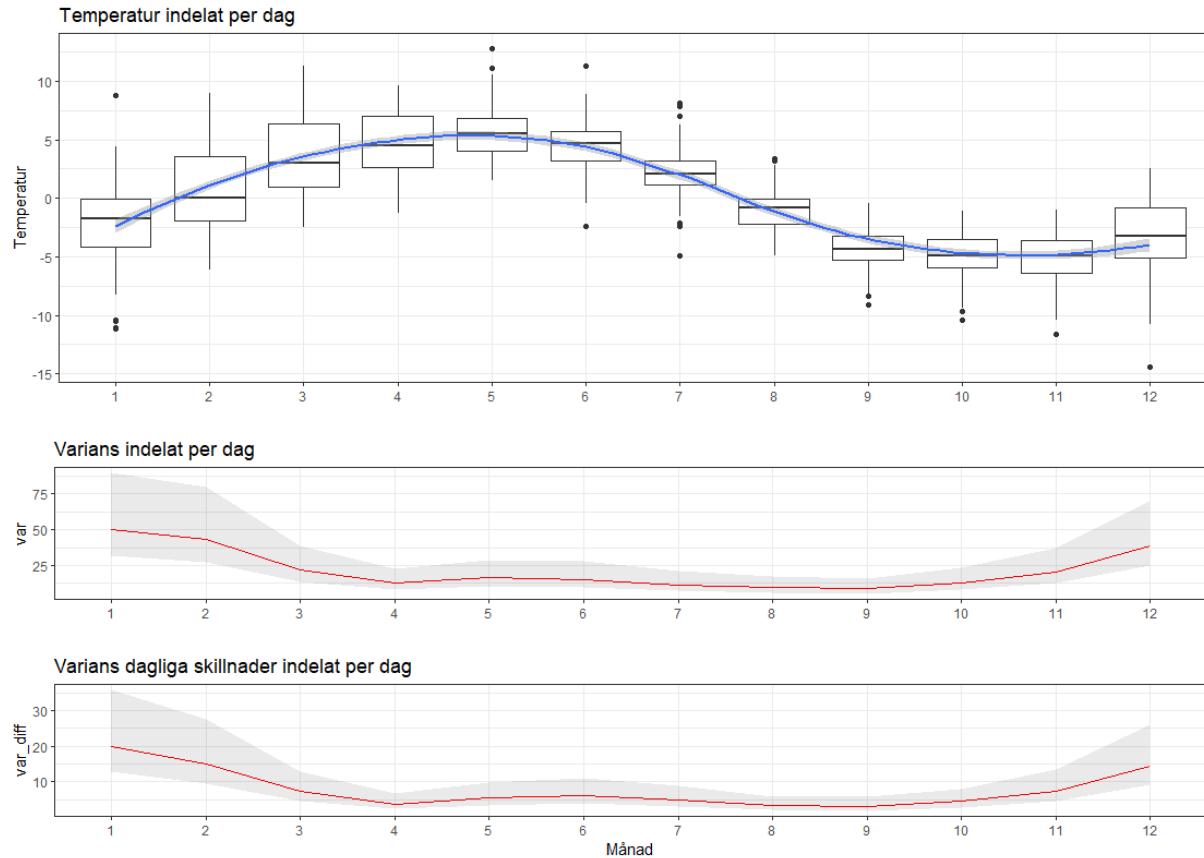
en enkel linjär regression där $\text{Temperatur} \sim \text{Tid}$, ger signifikans enligt regressionsstatistiken i tabell B.1. Detta är självklart en längre trend vilken kan differentieras bort och illustreras då av den nedre figuren.

Även om det föreligger en långsiktig uppåtgående trend kan vi enligt den differentierade grafen se hur differenser- na verkar konstanta över tid. Vi kan enligt ADF-testet konfirma att processen är stationär, där resultatet i tabellen visar på urpsrungsdataan Y_t . Det är därför in-

Augmented Dickey-Fuller test	
Statistika	-10.105
Lag order	29
P-värde	≤ 0.01
Mothypotes	Stationär

Tabell 2.1: ADF-test, daglig 8 data

te nödvändigt att differentiera även om risken för att göra ett skattningsfel är än mindre då Statistikan för differentierade serien är -36.356 på samma lag. Även om vi kan bli av med trenden i serien återstår det faktum att vi har ett mycket starkt säsongsberonde. Det blir kallare på vintern och varmare under sommaren. För att ge en illustration hur säsongskomponenten ungefärligt ser ut är temperaturen grupperad per månad i figur 2.2.



Figur 2.2: Caption

Att notera är de nedre graferna som istället visar hur stor variansen är av de dagliga temperaturmätningar vilka är grupperade per månad. Variationerna i såväl temperaturmätningar men också de dagliga skillnader från dag-till-dag ökar desto kallare det är. Med andra ord kan vi uppleva större skillnader från dag-till-dag i december jämförelsevis med juli. När vi bryter ned datamaterialet till dess komponenter kan tydligare se hur det finns en långt uppåtgående trend samtidigt som vi har en svängande säsongskomponent för varje år.

2.2 Modell och forecasting

Återigen kommer vi använda oss av de inbyggda funktionerna `ets()` och `auto.arima()` i syfte att bygga en modell.

En fingervisning på hur väl väl modell passar datamaterialet är att undersöka residualernas standardavvikelse. Då residualerna är avvikelser mellan det predikterade- och verkliga värdet eftersträvas minsta möjliga residual-standardavvikelse. I tabell 2.2 kan vi se att ETS-modellen på den ursprungliga tidsserien har lägst residual-standardavvikelse. För att jämföra modellerna och dess prediktionsförmåga på ny data undersöks istället modellernas PRESS-värde.

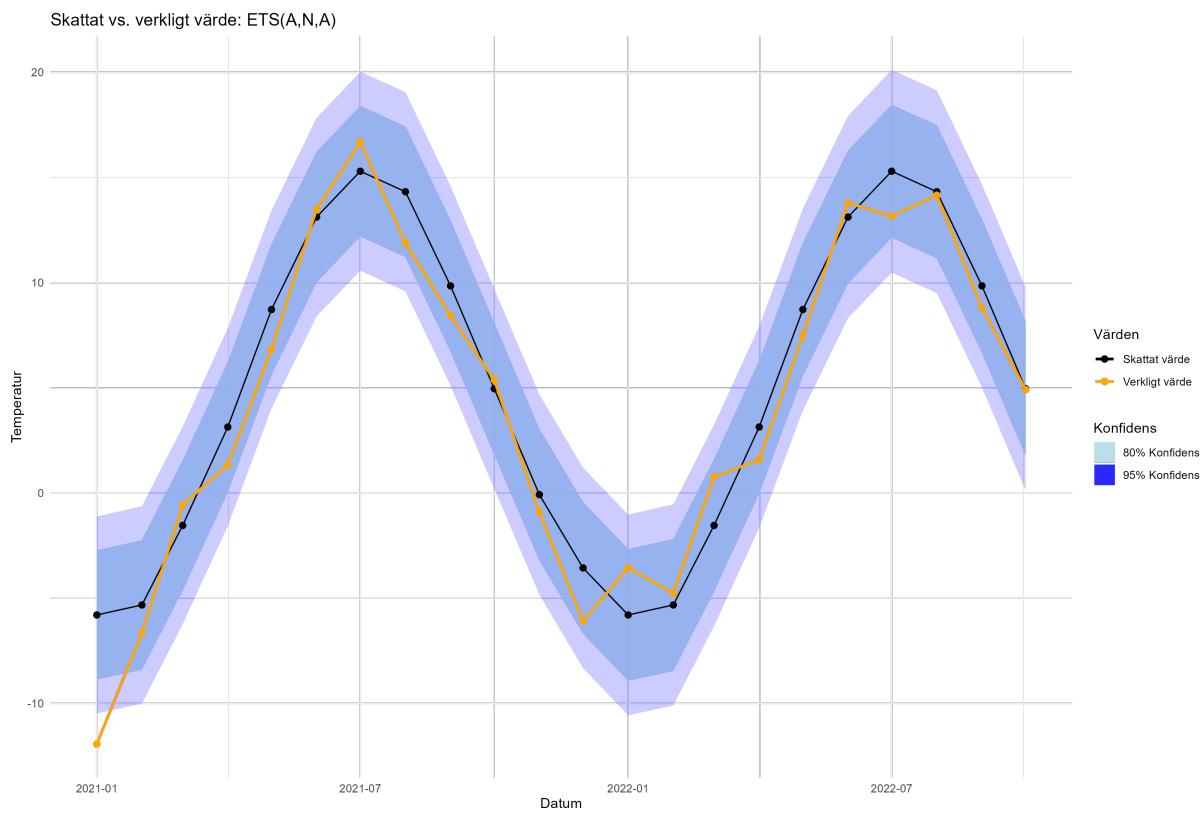
Vi finner i tabell 2.3 att `ETS_Y` har bäst prediktionsförmåga med lägst *PRESS*-värde. För att illustrera detta ges en visualisering i figur 2.3 av de skattade värdena gentemot det verkliga uppmätta värdena. Med undantag för den första datapunkten ligger alla andra uppmätta temperaturen inom ETS-modellens 95% konfidensintervall. ARIMA-modellen har en datapunkt fler än `ETS`-modellen utanför det 95% konfidensintervallet, även om alla punkter för båda modeller ligger inom det 80% konfidensintervallet. Vi kan konstatera av diagnostikplottarna i figur B.1 och B.3 att residualerna är normalfordelade. `ETS`-modellen verkar till synes hantera autoregressionen bättre än `ARIMA`-modellen även om den försöker korrigera för detta. QQ-plottarna visar att `ARIMA`-modellen har svårare att hantera extremvärder med större avvikelse vid svansarna i jämförelse med `ETS`-modellen. Sammanfattningsvis anser jag att, framförallt `ETS`-modellen, går att validera men att avvikelsen vid extremvärdena bör tas i åtanke. Vi kan nu göra en framtida prediktion där vi i figur 2.4 ser hur respektive modeller predikterar temperaturen tio år framöver.

Modell	SD
<code>ETS_Y</code>	2.39
<code>ETS_DY</code>	2.812
<code>ARIMA_Y</code>	2.737
<code>ARIMA_DY</code>	2.601

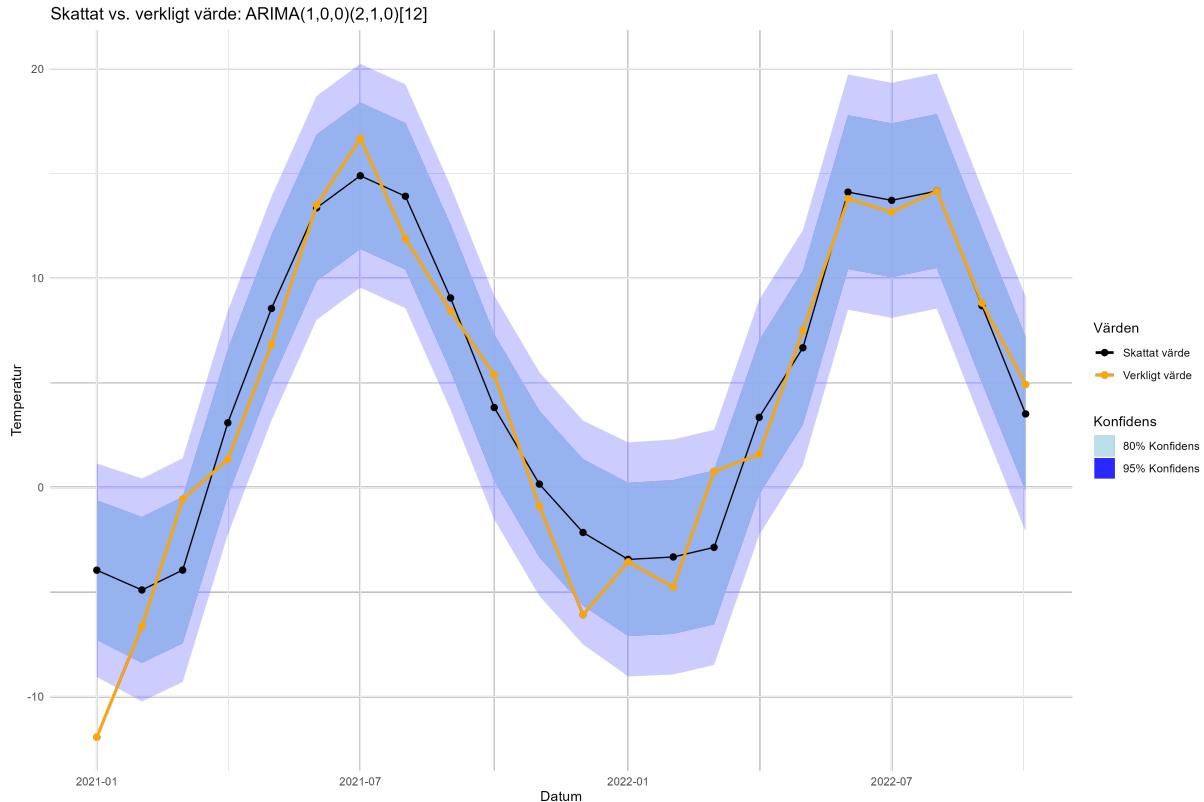
Tabell 2.2: Sd per modell

Modell	PRESS
<code>ETS_Y</code>	85.08
<code>ETS_DY</code>	1679.68
<code>ARIMA_Y</code>	132.34
<code>ARIMA_DY</code>	161.14

Tabell 2.3: PRESS-värde

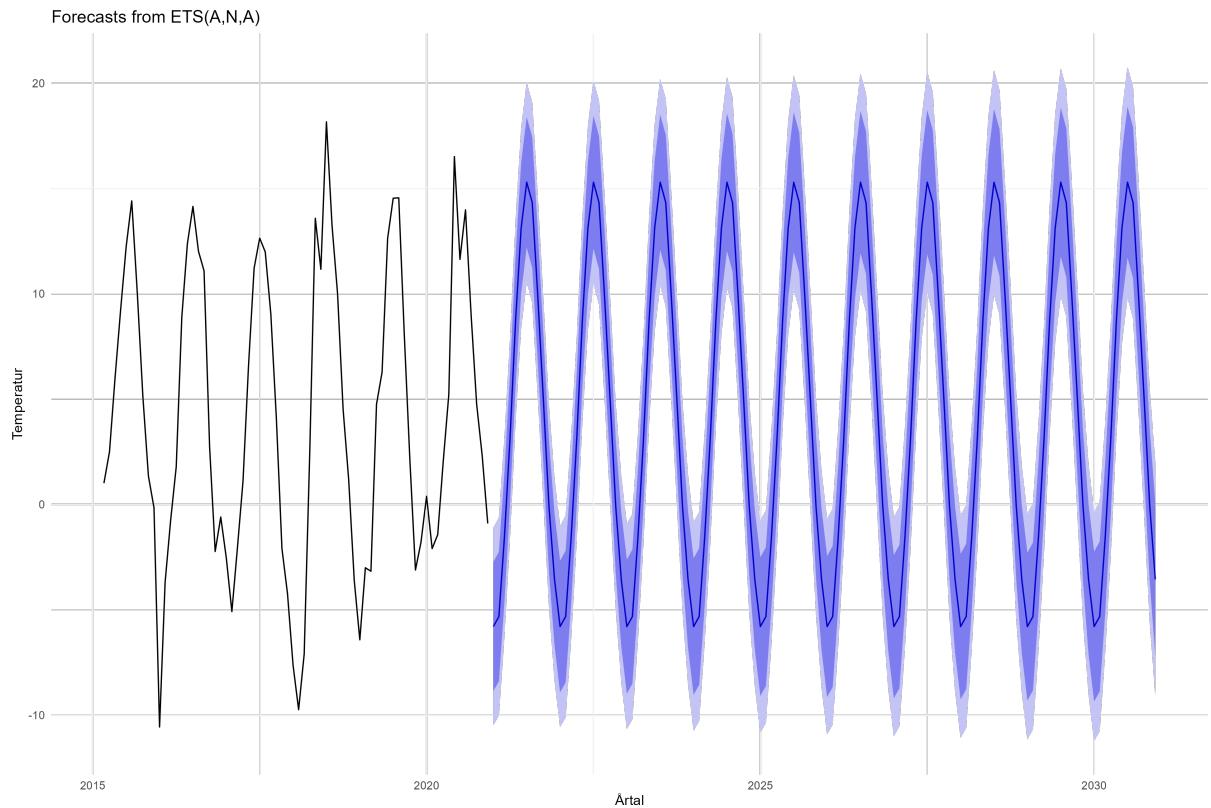


(a) Skattat vs. verkligt värde på testdata ETS

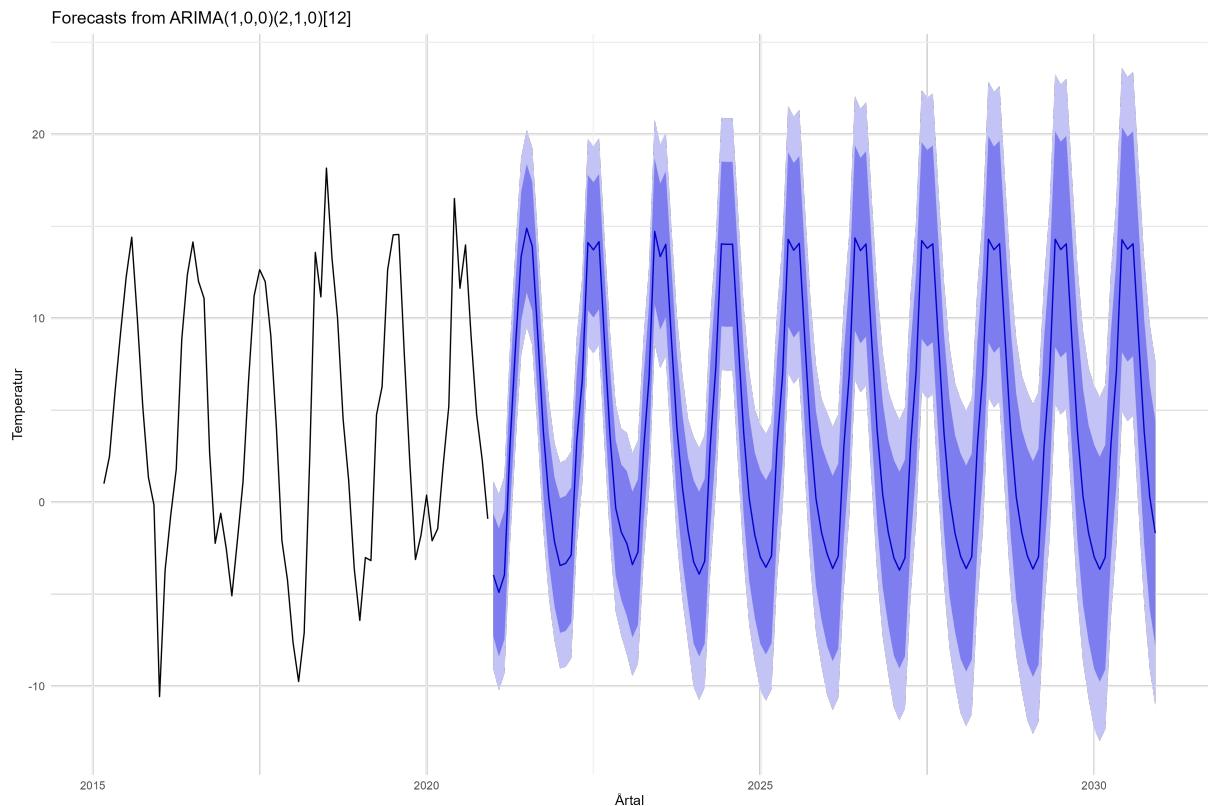


(b) Skattat vs. verkligt värde på testdata ARIMA

Figur 2.3: Jämförelse av ETS & ARIMA prediktionsförmåga på data 2021/01 - 2022/10



(a) Forecast ETS-modell 10 år framöver



(b) Forecast ARIMA-modell 10 år framöver

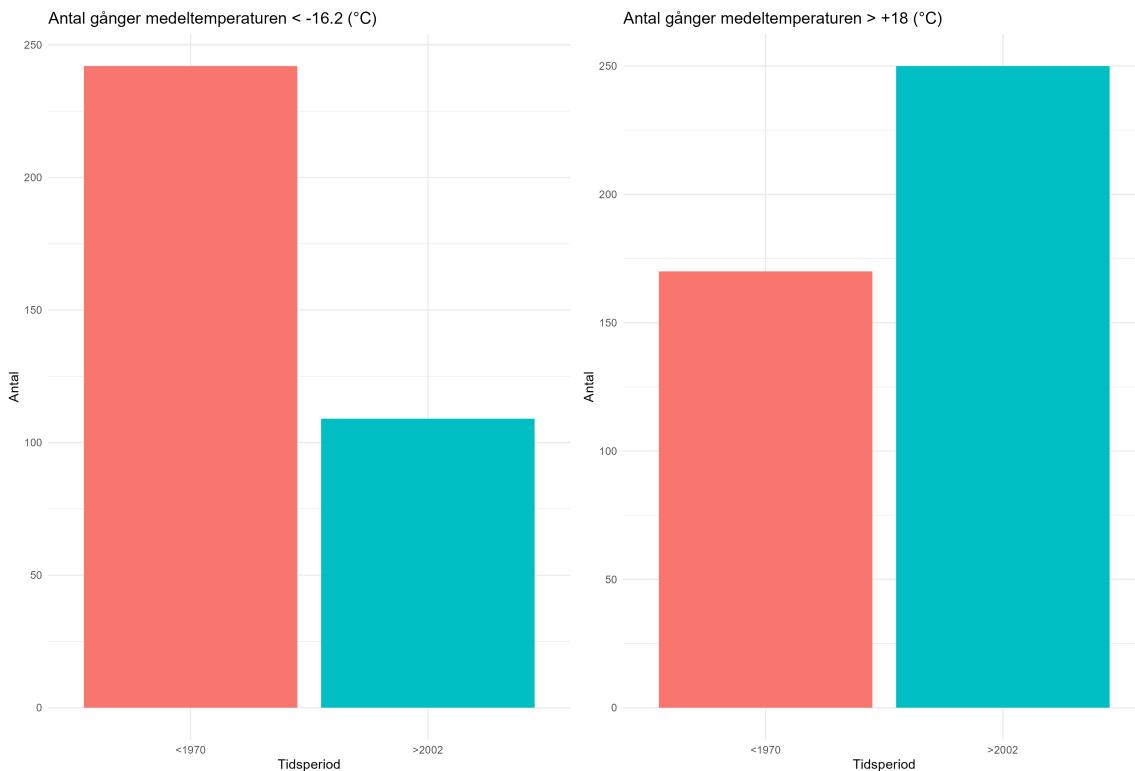
Figur 2.4: Forecasting av bästa två modeller, ETS & ARIMA

3 Förändringar över tidsperioden

För att undersöka förändringen av sannolikheten för extrem-värme/kyla delas materialet upp ytterligare i två perioder. SMHI definierar dessa perioder som vinter (december-februari) och sommar (juni-augusti). Här används kvantilfunktionen i R för att bestämma vilka medeltemperaturen som kan anses vara 'extremvärder'. För sommarperioden är detta $\geq 97.5\%$ kvantilen, och för vinterperioden är det 2.5% kvantilen. Vi undersöker därmed hur många gånger medeltemperaturen understigit -16.2018 och överstigit 17.9893 per respektive tidsperiod. För de två tidsperioderna skiljer sig detta åt gnaska markant. I figur 3.1 ser vi hur många gånger per repsktive periode det varit ett extremväder. Under tidsperioden 1950 – 1970

Kvantil	Temperatur ($^{\circ}C$)
2.5%	-16.2018
97.5%	17.9893

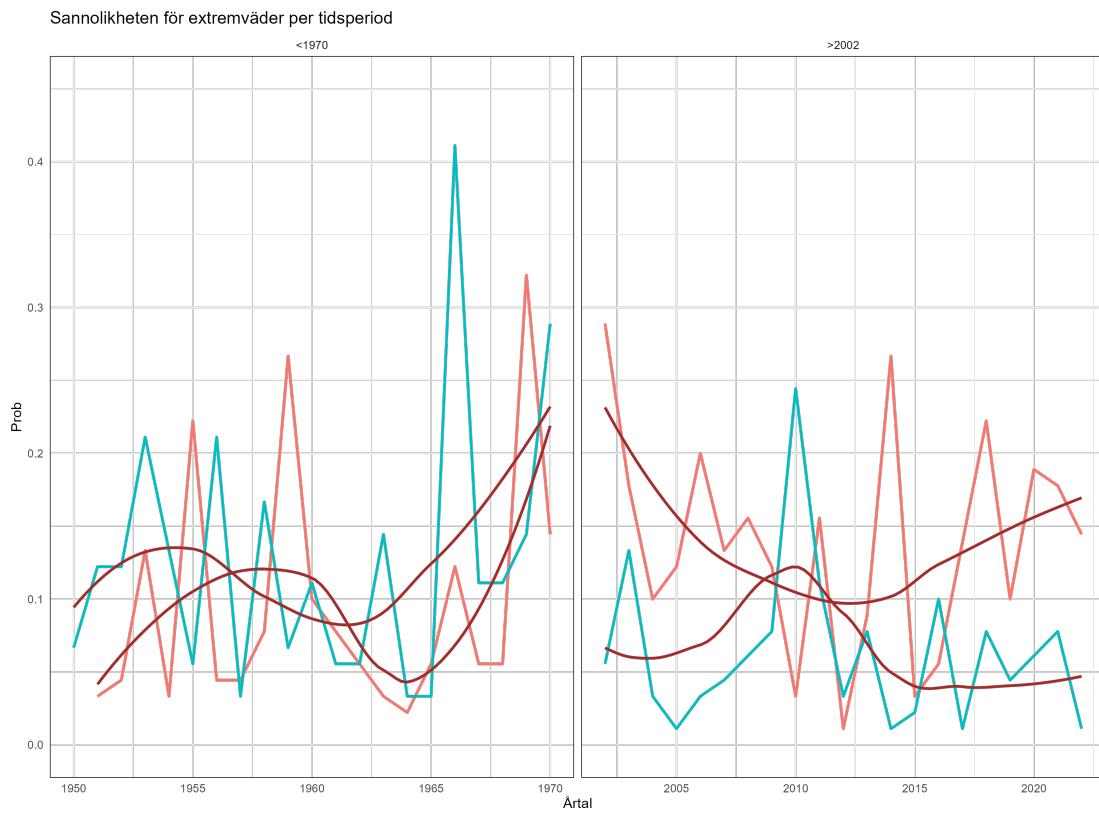
Tabell 3.1: Kvantilernas medeltemperatur($^{\circ}C$)



Figur 3.1: Antal gånger där medeltemperaturen indelat per tidsperiod under- och översteg kvantilerna 0.025% & 0.975%.

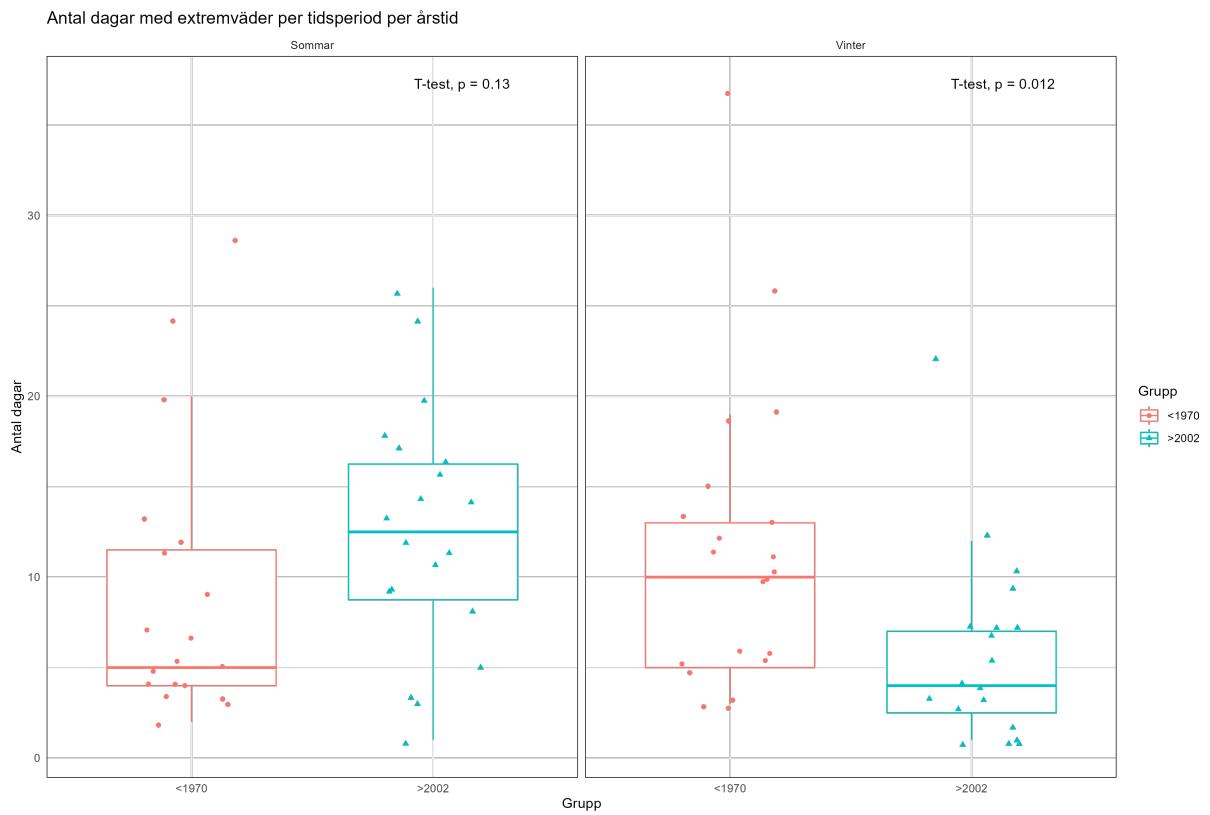
var det närmare 2.5 gånger så fler antal dagar med medeltemperaturen under $-16.2(^{\circ}C)$. Sommarperioden är istället det omvänta där tidsperioden 2002 – 2022 hade närmare 1.5 gånger så fler dagar med medeltemperatur över $18(^{\circ}C)$. Sannolikheten p (*extremväder*) beräknas då enligt ekvation 3.1:

$$p = \frac{\text{Antal gynnsamma utfall (i.e. antal dagar med extremväder)}}{\text{Antalet möjliga utfall (i.e. antal dagar under tidsperioden)}} \quad (3.1)$$



Figur 3.2: Sannolikheten för extremväder per tidsperiod med 'loess' styrlinje

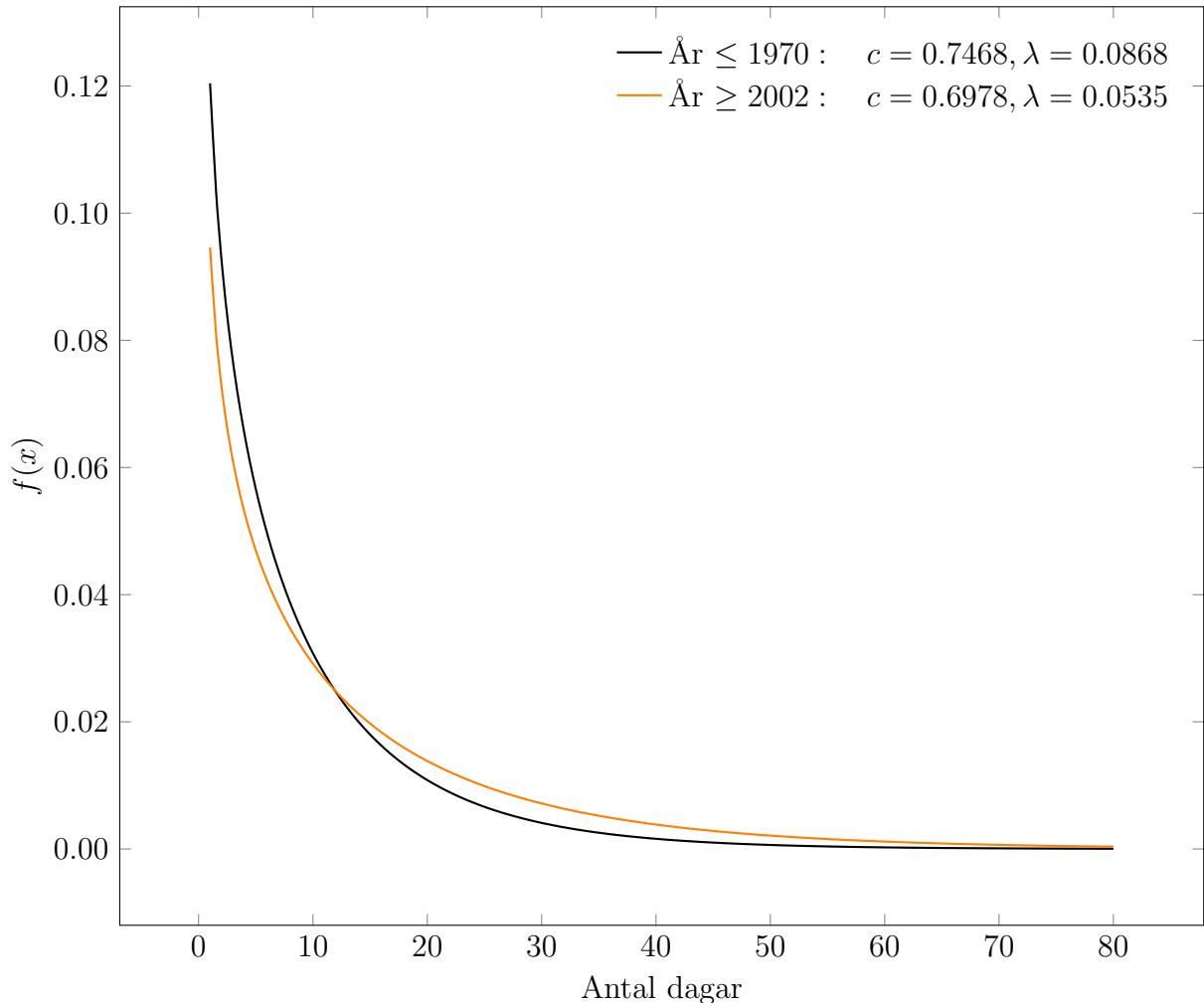
I figur 3.2 ser vi istället sannolikheten p för extremväder per respektive tidsperiod och år. Under tidsperioden 2002 – 2022 kan vi se att denna sjunkit till en låg stabil nivå för kallt väder. Samtidigt har sannolikheten för ovanligt varmt väder ökat. Intressant är dock att notera hur det verkade gå i en cykel för tidsperioden 1950-1970.



Figur 3.3: Boxplot med t-test för antal dagar med extremvärder

För att det ska vara statistiskt säkerställt att det skiljer sig mellan perioden måste kunna påvisas att det i genomsnitt är fler dagar per tidsperiod och årstid med extremvärder. I figur 3.3 kan vi se, med korresponderande t-test av medelvärdet, att vi faktiskt inte kan statistiskt säkerställa en skillnad mellan tidsperioderna. Däremot skall nämnas att få observationer kan ligga till grund för detta.

$$f(x) = \frac{\lambda^c}{\Gamma(c)} \cdot x^{c-1} \cdot \exp\{-\lambda x\}, \quad x \in (0, \infty)$$



Figur 3.4: Täthetsfunktion Gamma-fördelningen

Ett alternativ sätt att undersöka väderförändringarna över perioderna är att titta på hur antalet dagar mellan varje stora temperaturskillnad skiljer sig åt. Detta görs genom att standardisera temperaturskillnaderna från dag-till-dag genom ekvation 3.2:

$$std.Change = Z = \frac{x - \bar{x}}{\sigma} \quad (3.2)$$

Sedan beräknas antalet dagar mellan daglig temperaturförändring, där:

$$|std.Change| \geq 1.96$$

En fördelningsfunktionen har sedan skattats med hjälp av maximum-likelihood-metoden i programmet R och paketet 'fitdistrplus', där resultatet visade att gamma-fördelningen passar datan bäst. För period 1 (1950-1970) har shape-parametern skattats till 0.74684647, medan rate/scale-parametern skattats till 0.08683334. För perioden 2002-2022 har dessa parametrar istället skattats till 0.69778908 respektive 0.05353958. Täthetsfunktionerna

illustreras i 3.4. För att undersöka huruvida vi kan säkerställa att de två perioderna tillhör olika fördelningar har kan då göras ett likelihood-ratio-test. Detta visade på att vi inte heller här kan statistiskt säkerställa att de tillhör olika fördelningar även om skattningarna av modellparametrarna var olika.

4 Slutsats

I denna tidsserieanalys jämfördes olika modeller och deras prediktionsförmåga med två olika tidsintervall, en kort och en längre. Resultaten visade att modellen fungerade bättre när den fick mer data, vilket innebär att den längre perioden ger mer tillförlitliga prediktioner. Slutligen undersöktes huruvida det har skett förändringar av vädret sett över hela perioden. Här kan vi statistiskt säkerställa att medeltemperaturen stigit. Lite förvånande är dock att vi inte kan säkerställa att sannolikheten för extremväder, alltså mycket kallt eller mycket varmt väder, ökat med åren.

Kod: Kod för alla modeller, grafer och beräkningar och data finns att hämta på GitHub.

Bilaga A

Kort-data

	<i>Dependent variable:</i>		
	Temperatur		
	(Timvis)	(Daglig)	(Veckovis)
time(Datum)	-0.007*** (0.0001)	-0.166*** (0.009)	-0.989*** (0.134)
Constant	12.131*** (0.140)	12.217*** (0.622)	12.036*** (1.529)
Observations	2,927	122	19
R ²	0.705	0.750	0.762
Adjusted R ²	0.705	0.748	0.748
Residual Std. Error	3.793 (df = 2925)	3.413 (df = 120)	3.201 (df = 17)
F Statistic	6,982.911*** (df = 1; 2925)	359.767*** (df = 1; 120)	54.452*** (df = 1; 17)

Note:

*p<0.1; **p<0.05; ***p<0.01

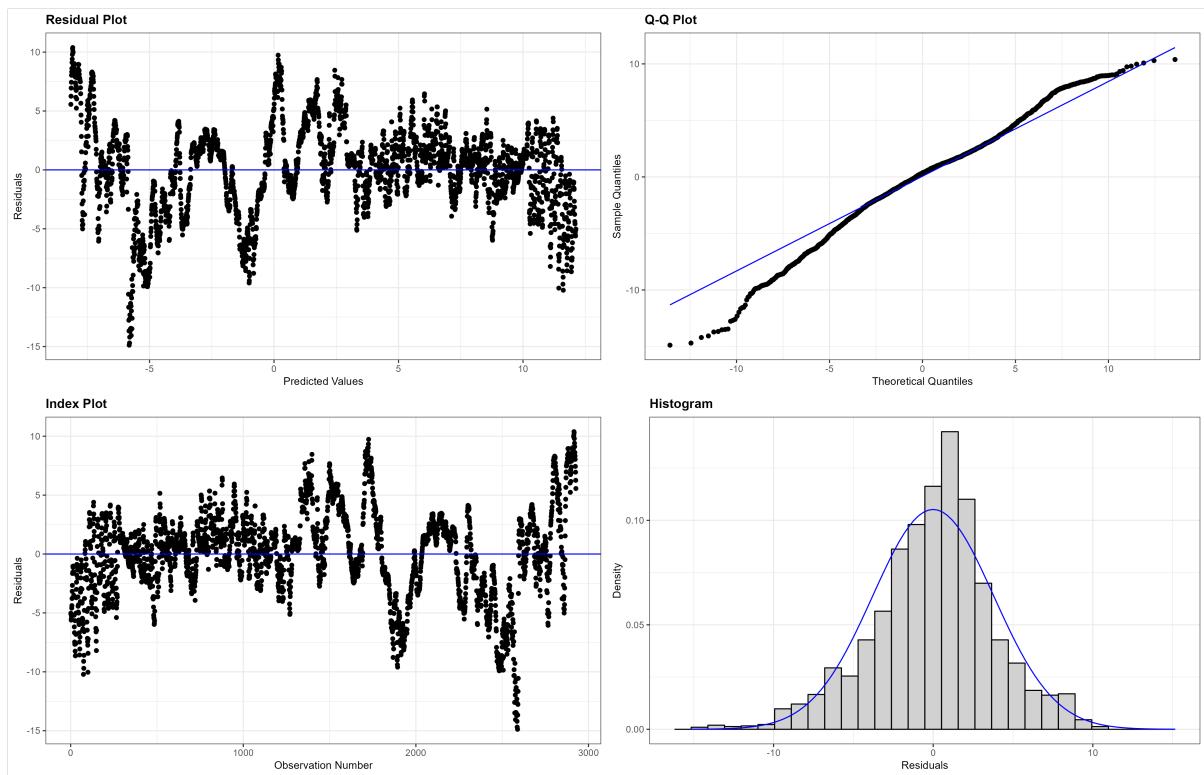
Tabell A.1: Regressionsstatistik enkel linjär regression, kort data med olika datagrupperingar

	Timvis	Daglig	Veckovis
Statistika	-4.5178	-2.9289	-1.54
Lag order	14	4	2
P-värde	0.01	0.1907	0.7476
Mothypotes	stationary	stationary	stationary

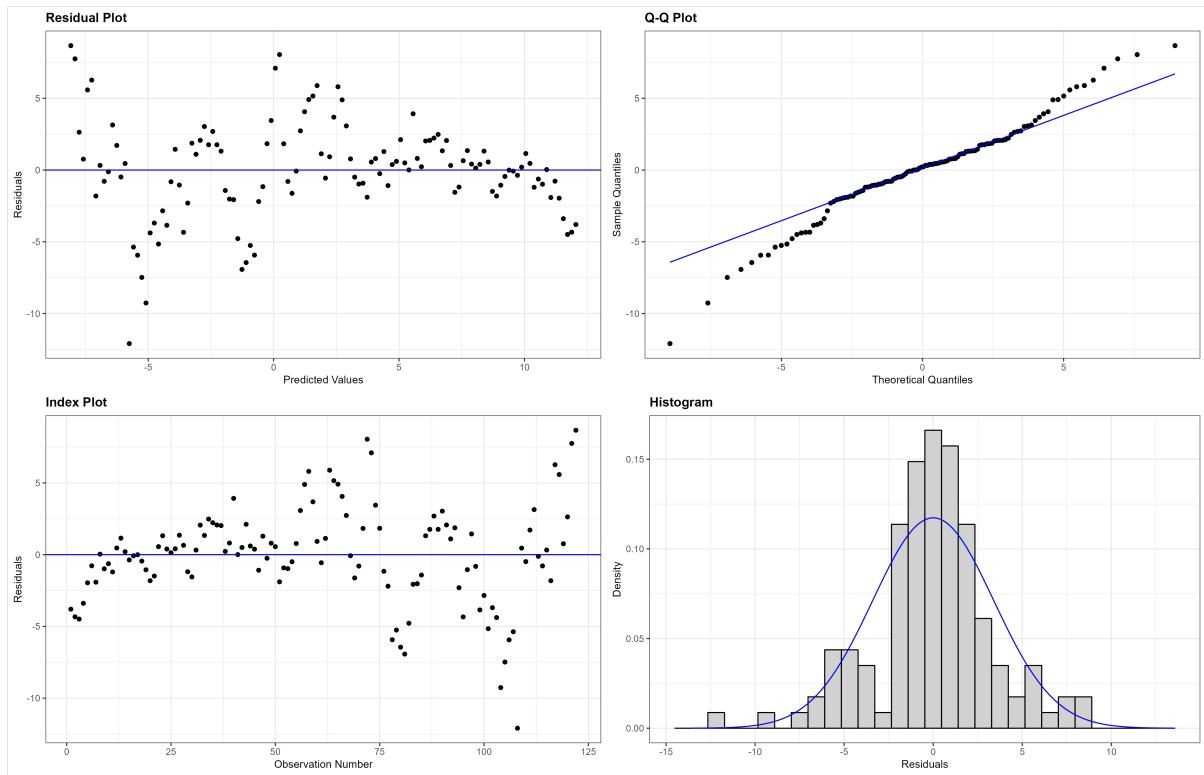
Tabell A.2: Augmented Dickey-Fuller test om stationär, kort data

	Timvis	Daglig	Veckovis
Statistika	-4.5178	-6.0527	-0.5188
Lag order	14	4	2
P-värde	0.01	0.01	0.9734
Mothypotes	stationary	stationary	stationary

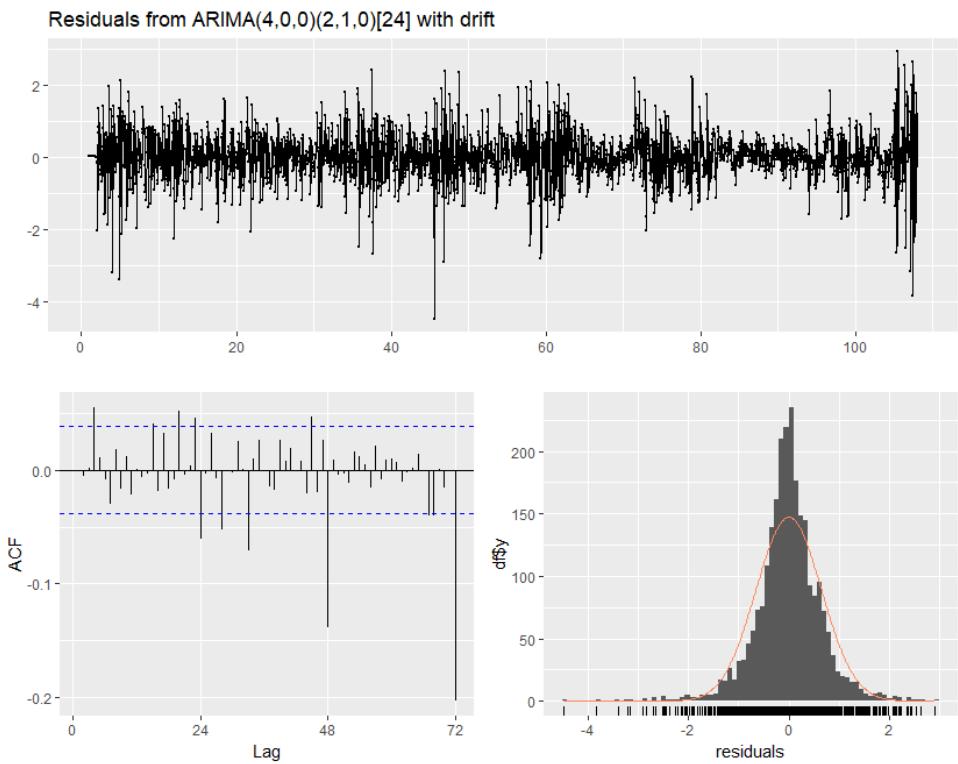
Tabell A.3: Augmented Dickey-Fuller test om stationär efter differentiering, kort data



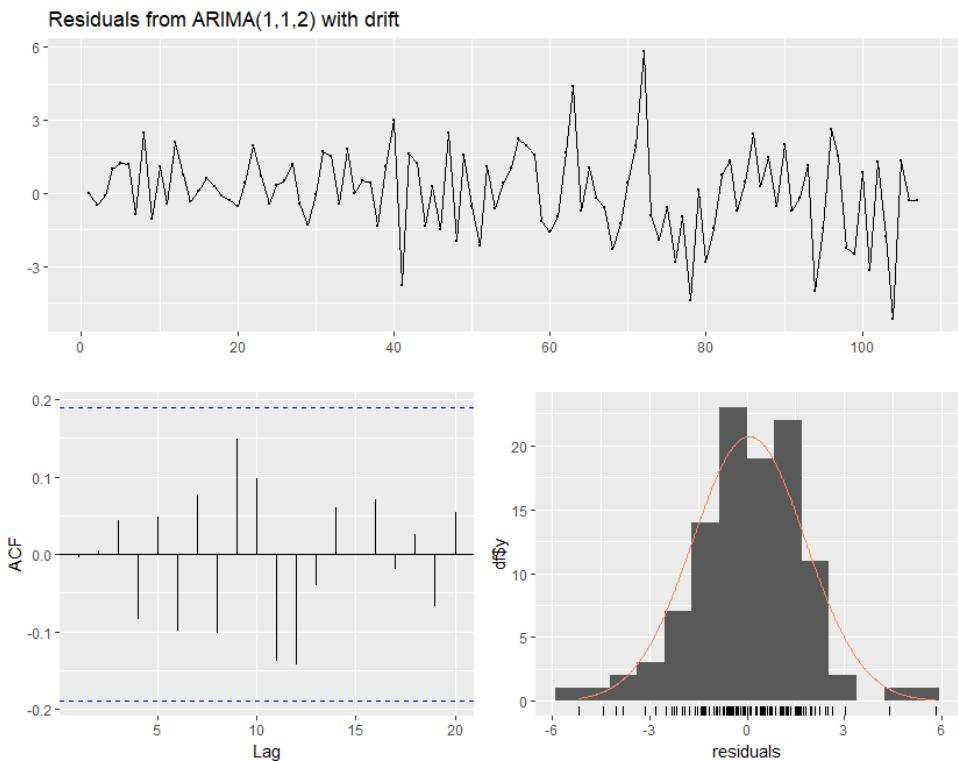
Figur A.1: Residualplottar enkel linjär regression kort timvis data



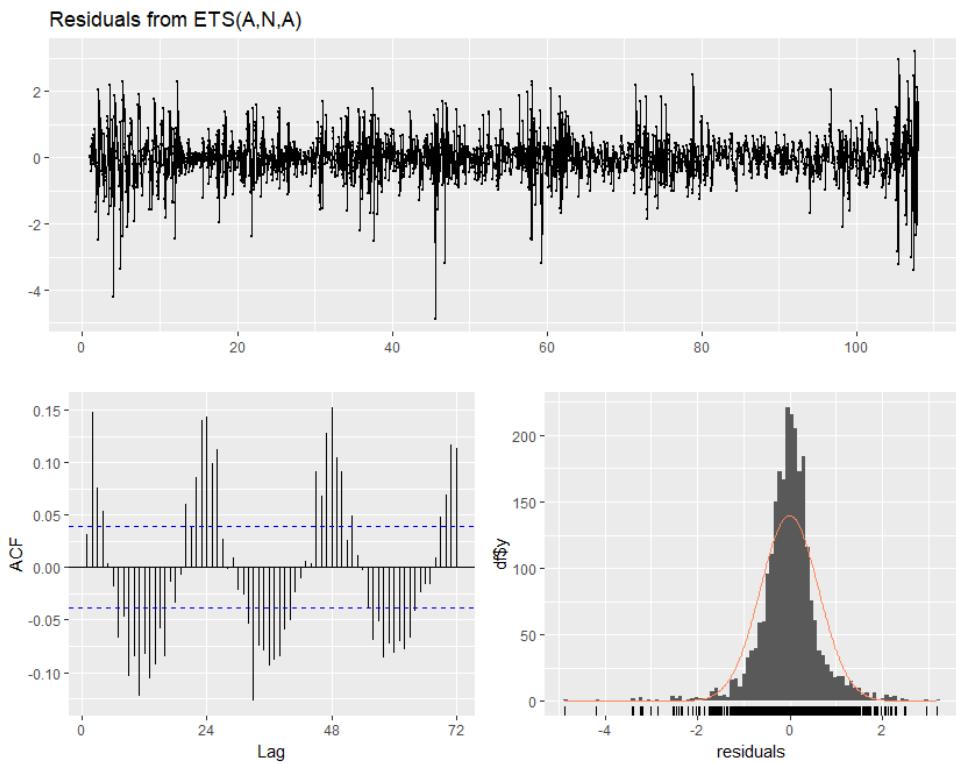
Figur A.2: Residualplottar enkel linjär regression kort daglig data



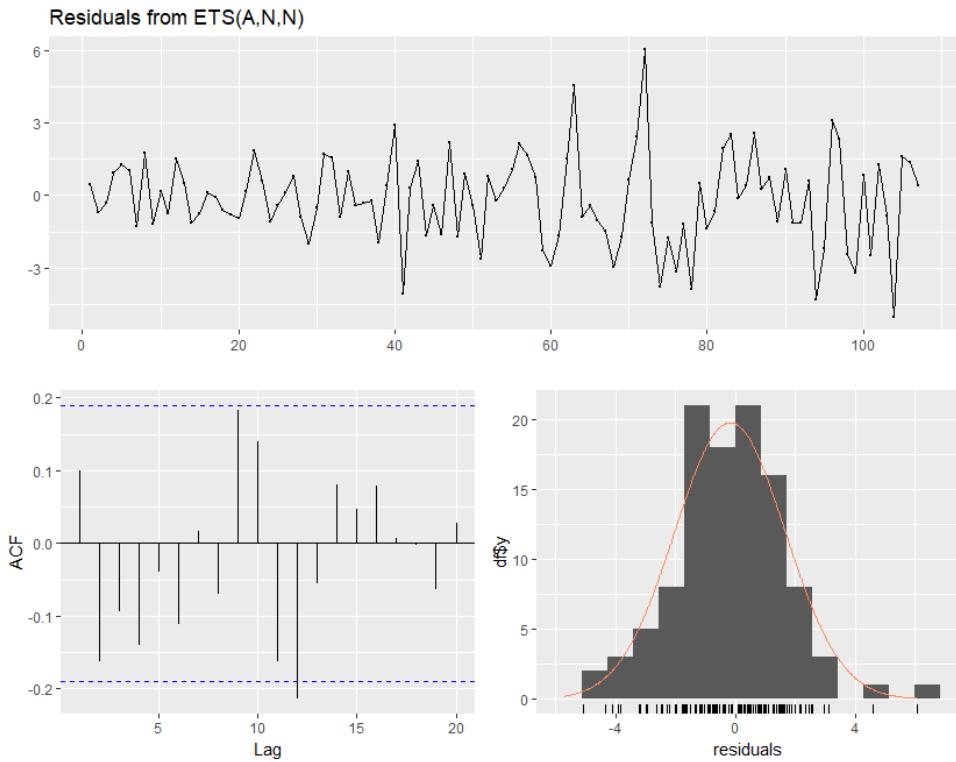
Figur A.3: Modelldiagnostik ARIMA_Y timme



Figur A.4: Modelldiagnostik ARIMA_Y dag



Figur A.5: Modelldiagnostik ETS_Y timme



Figur A.6: Modelldiagnostik ETS_Y dag

Bilaga B

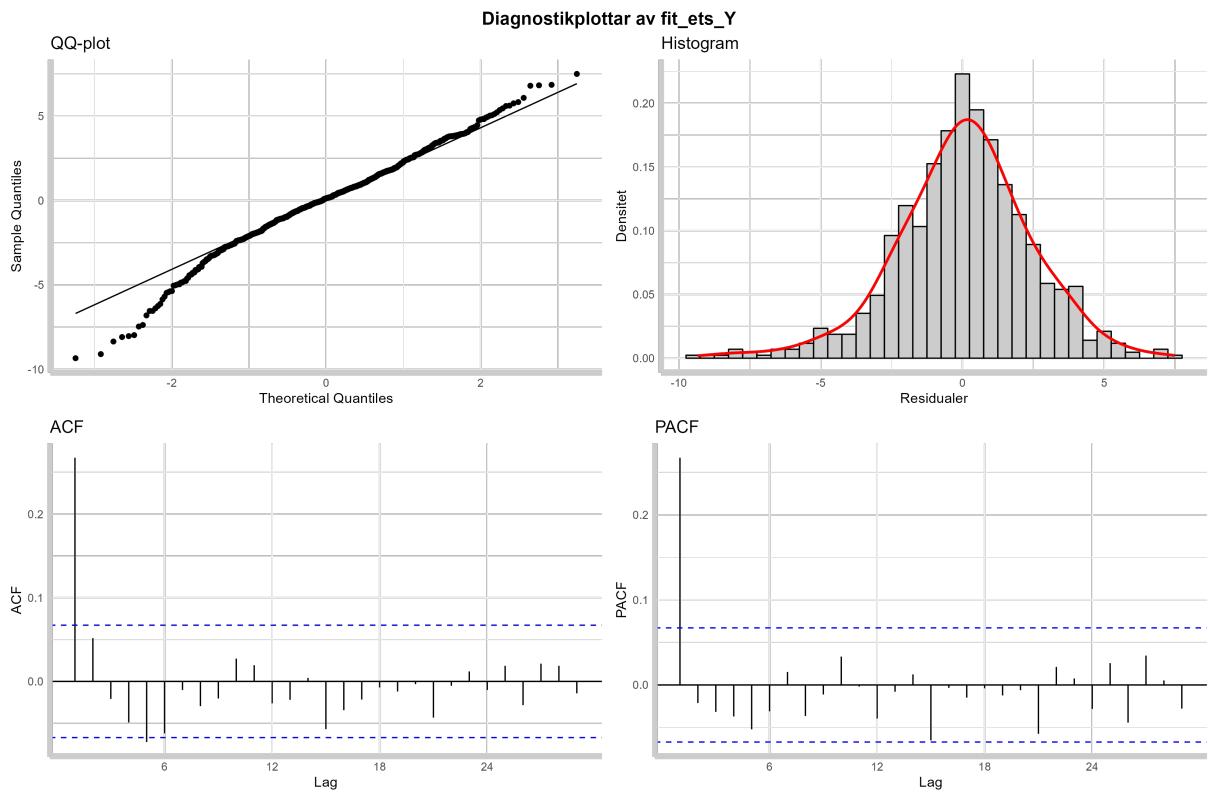
Lång-data

	<i>Dependent variable:</i>		
	Medeltemperatur		
	(Daglig)	(Veckovis)	(Månadsvis)
time(Datum)	0.0001*** (0.00001)	0.001*** (0.0001)	0.003** (0.001)
Constant	1.929*** (0.108)	1.754*** (0.271)	1.870*** (0.531)
Observations	26,544	3,853	874
R ²	0.005	0.006	0.007
Adjusted R ²	0.005	0.006	0.006
Residual Std. Error	8.780 (df = 26542)	8.407 (df = 3851)	7.841 (df = 872)
F Statistic	141.652*** (df = 1; 26542)	22.928*** (df = 1; 3851)	5.968** (df = 1; 872)

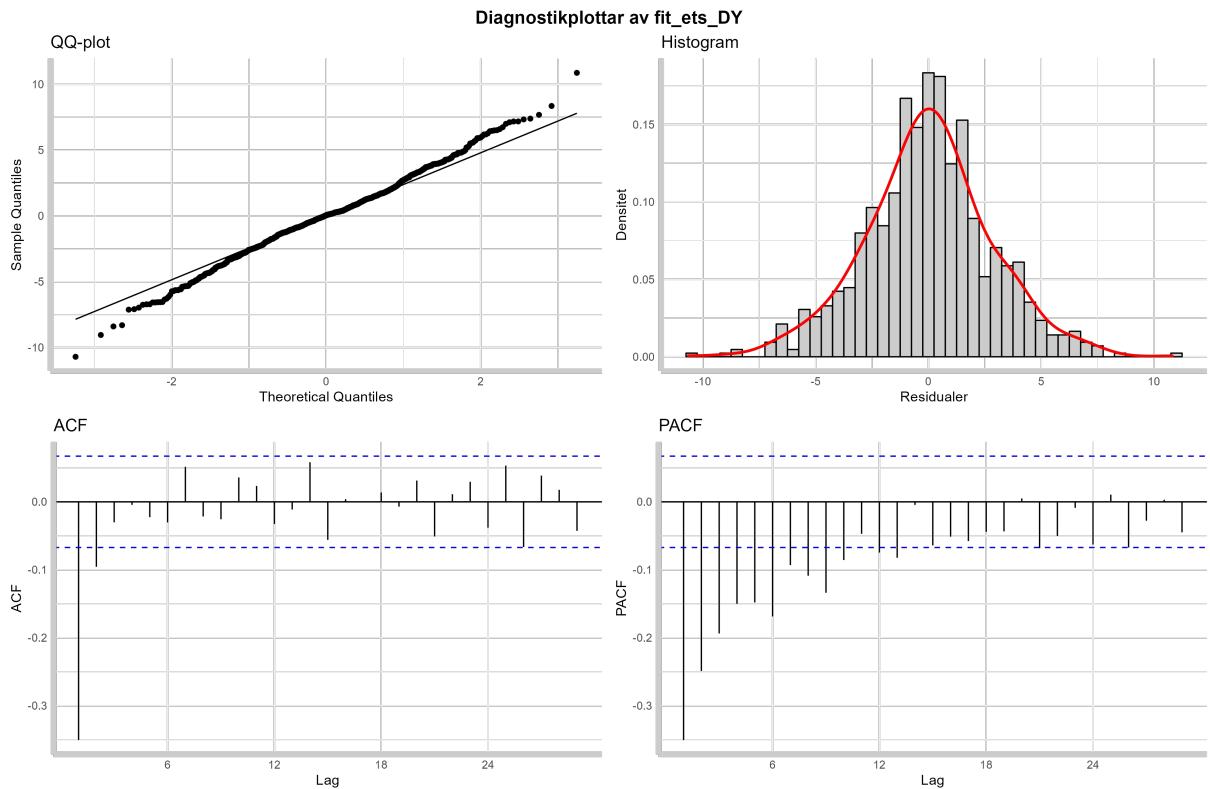
Note:

*p<0.1; **p<0.05; ***p<0.01

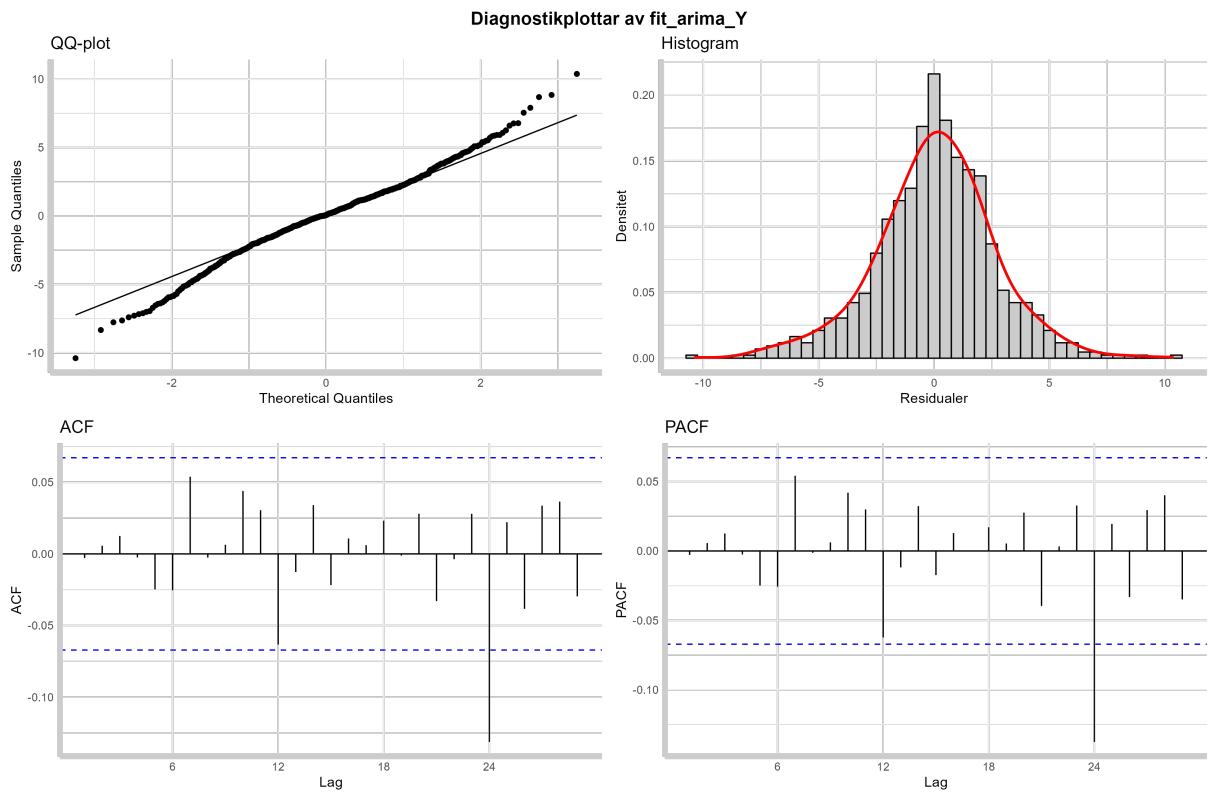
Tabell B.1: Regressionsstatistik enkel linjär regression, alla år med olika datagrupperingar



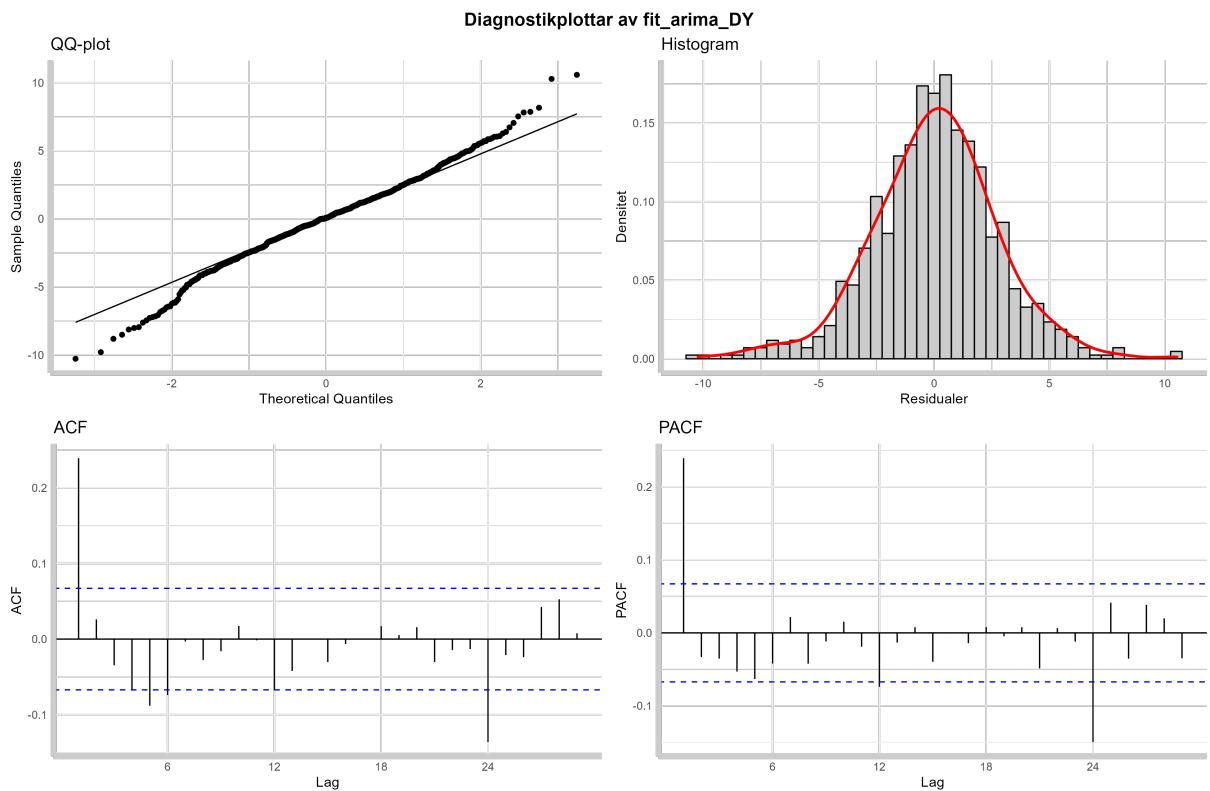
Figur B.1: Modelldiagnostik ETS_Y



Figur B.2: Modelldiagnostik ETS_DY



Figur B.3: Modelldiagnostik ARIMA_Y



Figur B.4: Modelldiagnostik ARIMA_DY