

Projektarbete i statistik

# Prediktion av kvävedioxidhalt i luften

732G39

William Wiik  
Oskar Storberg

Avdelningen för Statistik och maskininlärning  
Institutionen för datavetenskap  
Linköpings universitet  
Vårterminen, 20222

Handledare: Karl Wahlin, statistikkonsult  
Examinator: Isak Hietala, universitetsadjunkt

## **Sammanfattning**

Rapportens syfte är att anpassa en tidsseriemodell på data över kvävedioxid och därefter prediktera en vecka framåt i tiden, denna data är insamlad från en okänd italiensk stad som sträcker sig från mars 2004 till april 2005. Därefter är målet att jämföra modellerna för att se vilken som är bäst efter diverse lägesmått som går igenom i metoden. Genom rapporten redovisas tre modeller, den första modellen är multipel linjär regression som är dummykodad på dygnets timmar. Sedan användes ARIMA(2,1,2)(1,1,1) som har säsong differentiering på 24h och den sista modellen är en dämpad Holt-Winter modell. ARIMA och Holt-Winter hade en modell som blev överanpassad och uppfyllde inte antaganden, vilket innebar att de inte var lämpliga modeller för att göra prediktioner. Multipel linjär regression hade en modell som var underanpassad och som hade kvar autokorrelation. Slutsatsen blev alltså att det inte finns någon lämplig modell av de modeller som undersöktes att prediktera kvävedioxidhalten i en italiensk stad 1 vecka framåt.

## **Abstract**

The purpose of this report is to fit a timeseries model on data over carbon dioxide thereafter predict one week forward in time, this data is collected from an unknown Italian city where the data has been collected from March 2004 to April 2005. Further the goal will be to validate the created models to say which is the best, this will be done through various position measurements that can be read about in the method. Through the report various models are described such as multiple linear regression which is dummy coded every hour, furthermore we used ARIMA(2,1,2)(1,1,1) which is a seasonal ARIMA-model with differentiation on 24. Model 3 is a Holt-Winter that is a combination of additive and multiplicative, also known as damped Holt-Winter. ARIMA and Holt-Winter both had overfitted models and did not meet the requirements, this means none of the models are suitable for predictions. Multiple linear regression had an under adapted model that still had autocorrelation. The conclusion is therefor that none of the models that were investigated are suitable to predict the level of carbon dioxide in the unknown Italian city one week forward.

## **Förord**

Ett stort tack till Karl Wahlin som har väglett oss med denna rapport genom att ha föreslagit idéer och tankar.



# Innehåll

<b>1</b>	<b>Introduktion</b>	<b>1</b>
1.1	Bakgrund . . . . .	1
1.2	Syfte . . . . .	1
1.2.1	Frågeställningar . . . . .	1
1.3	Etiska och samhälleliga aspekter . . . . .	1
<b>2</b>	<b>Data</b>	<b>2</b>
<b>3</b>	<b>Metod</b>	<b>4</b>
3.1	Imputering . . . . .	4
3.2	Multipel linjär regression . . . . .	4
3.3	ARIMA . . . . .	5
3.4	Holt- Winter's metod . . . . .	6
3.5	Residualanalys . . . . .	7
3.5.1	Antagande om lika varians . . . . .	7
3.5.2	Antagande om normalfördelning . . . . .	7
3.5.3	Antagande om oberoende . . . . .	8
3.6	Durbin-Watson . . . . .	8
3.7	SAC och SPAC . . . . .	8
3.8	Jämförelsemått . . . . .	9
3.8.1	Förklaringsgrad . . . . .	9
3.8.2	MSE, MAD och MAPE . . . . .	9
3.9	Implementation i programvaror . . . . .	10
<b>4</b>	<b>Resultat</b>	<b>11</b>
4.1	Imputering . . . . .	11
4.2	Multipel linjär regression . . . . .	11
4.3	ARIMA . . . . .	15
4.3.1	Prediktion . . . . .	18
4.4	Holt-Winters metod . . . . .	19
4.4.1	Prediktion . . . . .	23
<b>5</b>	<b>Diskussion</b>	<b>25</b>

<b>6</b>	<b>Slutsats</b>	<b>26</b>
<b>7</b>	<b>Bilaga</b>	<b>28</b>
7.1	Bilaga A . . . . .	28
7.2	Bilaga B . . . . .	28
7.3	Bilaga C . . . . .	29
7.4	Bilaga D . . . . .	29

## Figurer

1	Koncentration av kvävedioxid i luften . . . . .	2
2	Visualisering av imputering . . . . .	11
3	Residualanalys, Multipel linjär regression . . . . .	13
4	SAC, Multipel linjär regression . . . . .	14
5	Differentierade tidserien, ARIMA . . . . .	15
6	Residualanalys, ARIMA . . . . .	16
7	SAC, ARIMA . . . . .	17
8	Faktiska värdena mot de anpassade värdena, ARIMA . . . . .	18
9	Prediktion, ARIMA . . . . .	19
10	Residualanalys, Holt-Winter . . . . .	21
11	SAC, Holt-Winter . . . . .	22
12	Faktiska värdena mot de anpassade värdena, Holt-Winter . . . . .	23
13	Prediktion, Holt-Winter . . . . .	24

## Tabeller

1	Information om bortfall, 1 . . . . .	3
2	Information om bortfall, 2 . . . . .	3
3	Information om bortfall, 3 . . . . .	3
4	Koefficienter, Multipel linjär regression . . . . .	12
5	Förklarande mått, Multipel linjär regression . . . . .	12
6	Durbin-Watson, Multipel linjär regression . . . . .	14
7	Koefficienterna för SARIMA(2,1,2)(1,1,1)[24] . . . . .	16
8	Förklarande mått, ARIMA . . . . .	16
9	Olika modeller $AIC_c$ -värden . . . . .	20
10	Koefficienterna för Holt-Winter (1-15) . . . . .	20
11	Koefficienterna för Holt-Winter (16-30) . . . . .	20
12	Typ av modell, Holt-Winter . . . . .	20
13	Förklarande mått, Holt-Winter . . . . .	21

# 1 Introduktion

## 1.1 Bakgrund

Karolinska institutet har bedrivit forskning inom området kväveoxider, vilket gasen kvävedioxid uppmärksamas som en påfrestning mot miljö samt folkhälsa. Detta leder till att det är av intresse att undersöka mängden av kvävedioxid i luften för att sedan kunna göra prediktioner. Undersökningen har genomförts med hjälp av sensorer som har befunnits sig i en stark förorenad stad i Italien. Mer detaljerat om N<sub>0</sub><sub>2</sub>, som är den kemiska beteckningen av kvävedioxid, är att den uppstår vid förbränning eller oxidation av kväveoxid. Det är en giftig gas, då gasen kan irritera luftvägarna för att sedan skada lungorna. För höga halter av N<sub>0</sub><sub>2</sub> kan gasen vara dödlig, vilket gör det ännu mer intressant att undersöka detta vidare. Kvävedioxid är även direkt farlig mot miljön, eftersom gasen tillsammans med solljus medverkar bildandet av marknära ozon. Kvävedioxid kan även bidra till övergödning och försurning av mark och vatten. (Naturvårdsverket; Andersson, 2022)

## 1.2 Syfte

Rapporten riktar sig till att prediktera mängden av kvävedioxid som släpps ut i en italiensk stad.

### 1.2.1 Frågeställningar

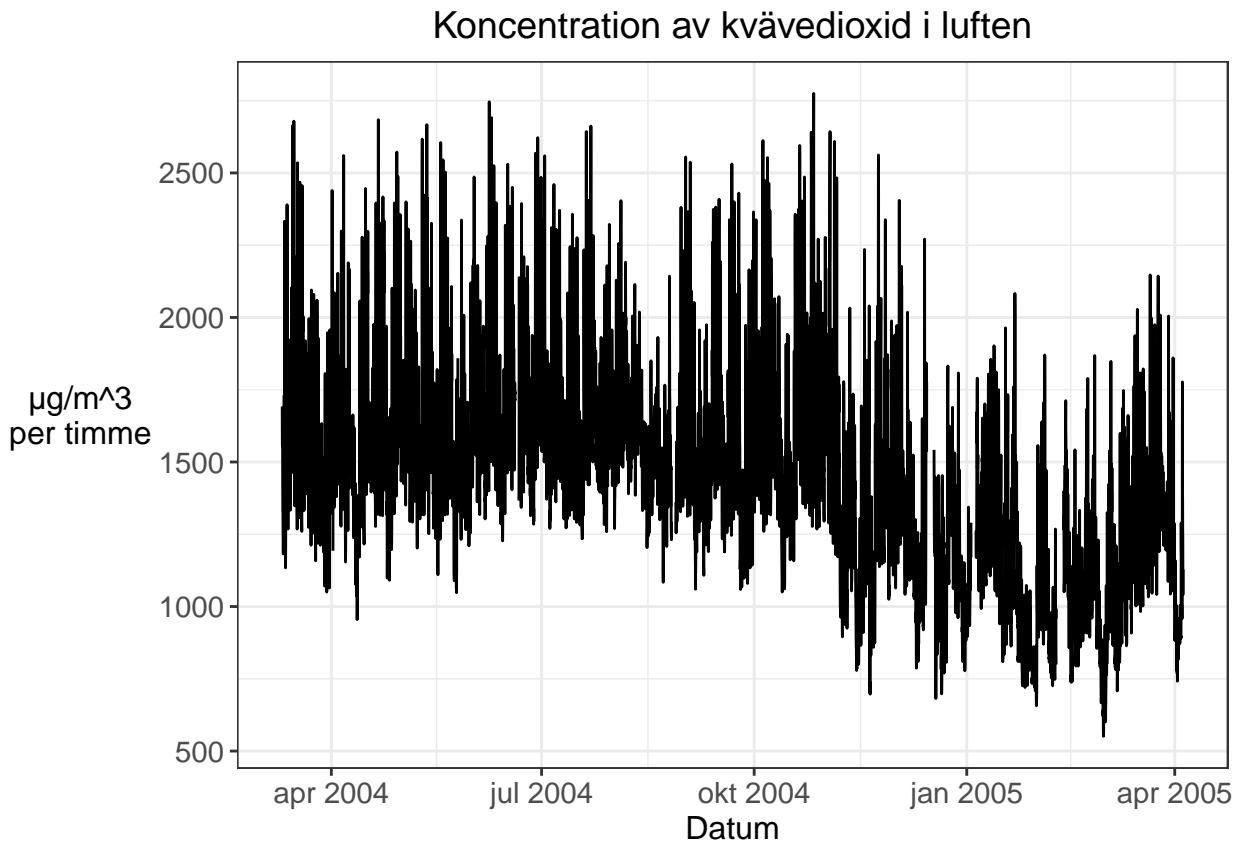
- Hur ser en lämplig modell ut för att prediktera kvävedioxidhalten i en italiensk stad 1 vecka framåt?

## 1.3 Etiska och samhälleliga aspekter

Kontroller av föroreningsnivåer kan jämföras via förorenade städer i Italien och bidra till att kontrollera vad som klassas som högt förorenade områden. Med denna information kan uppgångar ses i kontamineringsnivå genom prediktioner en vecka framåt i tiden, detta skulle kunna förberedda städer om vilka tidpunkter som utsläppen är höga och därmed förbättra folkhälsan.

## 2 Data

Datamaterialet i rapporten innehåller 9357 observationer på en variabel, PT08.S4.N02., vilket är kvävedioxid. Varje observation beskriver den genomsnittliga koncentrationen per timme och insamlingen har skett från mars 2004 till april 2005 i en italiensk stad vid ett allvarligt förorenat område. Det finns saknade värden i datamaterialet och dessa har markerats med -200 som sedan kodats om till NA. Enheten för koncentrationen är mikrogram/m<sup>3</sup> per timme som förkortas till  $\mu\text{g}/\text{m}^3$  per timme. (UCI, 2016).



Figur 1: Koncentration av kvävedioxid i luften

I figur 1 visas hela tidserien och det går att avläsa att halten befinner sig runt 1000-2500  $\mu\text{g}/\text{m}^3$  per timme fram till december 2004 där halten sänks till intervallet 500-2000.

Det finns 366 tidpunkter där värden saknas och i dessa fall kommer imputation användas. I tabell 1 syns mer information om bortfallet bland annat hur många "NA-luckor" som finns, vilket menas med hur många intervall med saknade värden som finns. I tabell 2 kan det avläsas hur många bortfall det förekommer för respektive timme och det verkar vara förhållandevis slumpmässigt. Dock verkar inte bortfallet vara slumpmässigt i tabell 3, då det förekommer betydligt mer saknade värden vid onsdag, torsdag och fredag. Bortfallet kan därför antas vara MAR, *missing at random*, då sannolikheten att det förekommer saknade värden är större för vissa dagar. Det är dock svårt att veta varför bortfallet är högre vissa dagar på grund av den begränsade information om hur mätningarna gick till. Det kan alltså inte helt fastställas att bortfallet är MAR, men det är det som rapporten

kommer anta.

Tabell 1: Information om bortfall, 1

Längd av tidserie	Antal NA	% av NA i tidserien	Antal NA-luckor	Medellängd av NA-luckor
9357	366	3.91%	16	22.875

Tabell 2: Information om bortfall, 2

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Antal NA för respektive timme	18	16	15	14	13	13	14	14	14	12	13	14	13	13	16	15	16	18	17	19	17	17	17	18

Tabell 3: Information om bortfall, 3

	Söndag	Måndag	Tisdag	Onsdag	Torsdag	Fredag	Lördag
Antal NA för respektive veckodag	27	32	44	75	70	102	16

## 3 Metod

### 3.1 Imputering

Hantering av saknade värden i en tidserie görs genom imputering. Linjär imputering är en metod för att framställa nya datapunkter inom ett intervall av kända datapunkter. För att ta reda på värdet av dessa nya datapunkter används formel 1. (Bayen & Siauw, 2015).

$$y = y_1 + (x_1 - x) \frac{(y_2 - y_1)}{(x_2 - x_1)} \quad (1)$$

där  $x_1$  och  $y_1$  är de första koordinaterna

$x_2$  och  $y_2$  är de andra koordinaterna

$x$  är x-koordinaten där imputeringen ska göras

### 3.2 Multipel linjär regression

Multipel linjär regression är en lämplig modell när det finns en beroende variabel  $y_t$  och minst två förklarande variabler. Modellen visas enligt ekvation 2. (Bowerman, O'Connell & Koehler 2005).

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon_t \quad (2)$$

$\beta_0$  är interceptet där linjen skär y-axeln.  $\beta_1$  till  $\beta_k$  är koefficienterna för prediktorn  $x_k$  som tillsammans bestämmer lutningen för regressionslinjen.  $\epsilon_t$  är modellens felprecision även kallad slumpterm. Det finns vissa krav som måste uppfyllas på slumptermen  $\epsilon_t$  och dessa krav kommer beskrivas i kapitel 3.5. Sedan finns det ytterligare ett krav, vilket är att det ska finnas ett linjärt samband mellan responsvariabeln och de förklarande variablerna. (Bowerman, O'Connell & Koehler 2005).

I många fall används kvalitativa variabler och dessa behövs hanteras med dummyvariabler. Detta innebär att de kvalitativa variablerna endast kan anta värdet 1 eller 0. Variabeln antar värdet 1 om analysenheten innehåller egenskapen som är av intresse och 0 annars. Hur många dummyvariabler som ska ingå i modellen bestäms genom att skapa  $c - 1$  dummyvariabler, där  $c$  är antalet kategorier. Det är alltså en kategori som utelämnas, denna kategori kallas referenskategori och koefficienterna för de  $c - 1$  dummyvariablerna kommer visa de förväntade skillnaderna, gentemot referenskategorin. (Bowerman, O'Connell & Koehler 2005).

Skattningarna för modellens samtliga  $\beta$  koefficienter skattas med minsta kvadratmetoden och ekvation 2 kan skrivas om till matrisform enligt ekvation 3 för att få en enklare bild hur minsta kvadratmetoden fungerar. (Bowerman, O'Connell & Koehler 2005).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \quad (3)$$

Där

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{pmatrix} \quad (4)$$

$\mathbf{X}$  kallas för designmatrisen och består av samtliga datapunkter enligt ekvation 5.

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{t1} & X_{t2} & \cdots & X_{tk} \end{pmatrix} \quad (5)$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad (6)$$

$$\mathbf{E} = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_t \end{pmatrix} \quad (7)$$

För att skatta modellens samtliga  $\beta$  koefficienter används formel 8.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (8)$$

### 3.3 ARIMA

ARIMA är en modell som predikterar framtida värden baserat på föregående värden i tidsserier. Ett krav för att göra en ARIMA-modell är att tidserien är stationär, vilket menas med att tidserien har ett konstant väntevärde och en konstant varians över tiden. Om tidserien inte är stationär kan den differentieras tills den blir stationär. Ifall trend ses i tidserien skall originalserien differentieras med 1 tidsförskjutning, se ekvation 9. (De Livera, Hyndman & Snyder 2010).

$$y'_t = y_t - y_{t-1} \quad (9)$$

$y'_t$  är den stationära tidsserien

$y_t$  är originalserien

Om säsongsvariation finns i tidserien skall originalserien differentieras för säsong, se ekvation 10.

$$y'_t = y_t - y_{t-L} \quad (10)$$

där  $L$  är säsongslängden

Den stationära tidsserien  $y'_t$  ska sedan modelleras med ARIMA(p,d,q) som har tre delar. Den första delen är en autoregressiv modell med ordning p som betecknas AR(p), se ekvation 11.

$$y'_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t \quad (11)$$

$$c = \mu(1 - \phi_1 - \dots - \phi_p) \quad (12)$$

$\epsilon_t$  är vitt brus, alltså slumptermen. Alla  $\epsilon_t$  är oberoende och normalfördelade med väntevärde 0 och konstant varians.

$\phi_p$  är en lutningskoefficienterna för AR-delen som skattas med minsta kvadratmetoden.  
 $\mu$  är medelvärdet av den stationära tidsserien  $y'_t$ .

Den andra delen är en moving average modell med ordning q som betecknas MA(q).  $y'_t$  modelleras här endast med vitt brus, se ekvation 13.

$$y'_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \quad (13)$$

$\theta_q$  är en lutningskoefficienterna för MA-delen, som skattas med minsta kvadratmetoden.

Kombineras dessa två delar får en ARMA-modell, ARMA(p,q), se ekvation 14. För att göra modellen komplett läggs en tredje del till, vilket är d som beskriver hur många differentieringar för trend som har gjorts. Detta bildar ARIMA(p,d,q). Om till exempel originalserien har differentierats en gång på trend är d=1. (De Livera, Hyndman & Snyder 2010).

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \quad (14)$$

En ARIMA-modell kan även inkludera en säsongsdel där endast det L:te tidsavståndet modelleras, L står för säsongslängden. Detta bildar SARIMA(p,d,q)(P,D,Q)<sub>L</sub>. Ekvation 17 visar en SARIMA(1,1,1)(1,1,1)<sub>12</sub> med bakåstegsoperatorer som betecknas med B. (De Livera, Hyndman & Snyder 2010).

$$By'_t = y'_{t-1} \quad (15)$$

$$B^2 y'_t = y'_{t-2} \quad (16)$$

$$(1 - \phi_1 B) (1 - \Phi_1 B^{12}) (1 - B) (1 - B^{12}) y'_t = (1 + \theta_1 B) (1 + \Theta_{1,12} B^{12}) \varepsilon_t \quad (17)$$

$\Phi$  och  $\Theta$  är respektive parametrar för SAR(P) och SMA(Q), vilket skattas på samma som AR(p) och MA(q) fast där endast det L:te tidsavståndet modelleras. (De Livera, Hyndman & Snyder 2010).

### 3.4 Holt- Winter's metod

Holt- Winter's metod är ett sätt att modellera tre aspekter av en tidserie, nivå  $\ell_t$ , trend  $b_t$  och säsong  $s_t$ . De två vanligaste varianterna av Holt-Winter är en ren additiv och ren multiplikativ modell. Det finns dock andra varianter av Holt- Winter's metod, vilket är en kombination av en additiv och multiplikativ modell där säsong, trend och felet är antingen additiv eller multiplikativ. Sedan finns det även Holt- Winter's dämpade metod som dämpar trenden och lämpar sig för tidserier som har en linjär trend med en snabb onaturlig avvikelse av trenden. För att ta reda på om säsong, trend och felet ska vara additiva eller multiplikativa testas många kombinationer av modeller och den modell som får lägst  $AIC_c$ , se ekvation 18, anses vara den mest lämpade. (De Livera, Hyndman & Snyder, 2010).

$$AIC_c = -2\ln(\text{likelihood}) + 2k \left( \frac{n}{n-k-1} \right) \quad (18)$$

*likelihood* är likelihood funktionen av den anpassade modellen.

*k* är antal parametrar i modellen.

*n* är antal observationer.

Denna rapport kommer rikta in sig på när säsongen och felet är multiplikativ och trenden är additiv. Det kan förkortas till ETS(M,A<sub>d</sub>,M), ETS = ErrorTrendSeason och det nedsänkta d:et står för att trenden är dämpad. Modellen visas enligt ekvation 19. (De Livera, Hyndman & Snyder, 2010).

$$y_t = (\ell_{t-1} + \phi b_{t-1}) s_{t-1}^{(L-1)} (1 + \epsilon_t) \quad (19)$$

$$\ell_t = (\ell_{t-1} + \phi b_{t-1}) (1 + \alpha \epsilon_t) \quad (20)$$

$$b_t = \phi b_{t-1} \beta (\ell_{t-1} + \phi b_{t-1}) \epsilon_t \quad (21)$$

$$s_t^{(0)} = s_{t-1}^{(L-1)} (1 + \gamma \epsilon_t) \quad (22)$$

$$s_t^{(i)} = s_{t-1}^{(i-1)}, \quad i = 1, \dots, L-1 \quad (23)$$

Där  $\alpha$ ,  $\beta$ ,  $\gamma$  är utjämningsparametrar för nivå, trend och säsong. Dessa parametrar kan anta värden mellan 0 och 1 och reglerar vilken andel av nivå, trend och säsong som står för högst eller lägst förklaring. Ett värde nära 0 på  $\alpha$  tolkas som att nivån aldrig uppdateras, medan ett värde nära 1 tolkas som att nivån uppdateras efter varje observation.

Ett värde nära 0 på  $\beta$  tolkas som att trenden är linjär, medan ett värde nära 1 tolkas som att trenden uppdateras efter varje observation. Ett värde nära 0 på  $\gamma$  tolkas som att säsongen är fix, medan ett värde nära 1 tolkas som att säsongen uppdateras efter varje observation.

$\phi$  står för dämpningsparamtern och ligger mellan 0 och 1, men värdet är sällan mindre än 0.8. Värdet på  $\phi$  nära 1 tolkas som att modellen knappt går att särskilja från en icke-dämpad modell.

*L* står för säsongs längden.

$\epsilon_t$  står för felet och antas vara oberoende och normalfördelade med väntevärde 0 och konstant varians.

För att skatta  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\phi$  och de intiala värdena  $\ell_0$ ,  $b_0$ ,  $s_0$ ,  $s_{-1}, \dots, s_{-L+1}$  används maximum-likelihood. (De Livera, Hyndman & Snyder, 2010).

## 3.5 Residualanalys

Inom de tre modellerna som tagits upp, multipel linjär regression, ARIMA och Holt-Winter är utvärdering av modellens antagande en viktig aspekt. Det finns tre krav på residualerna och dessa kontrolleras genom att visualisera residualerna i olika diagram.

### 3.5.1 Antagande om lika varians

Det första kravet är att residualerna ska ha lika varians, vilket kan kontrolleras i ett spridningsdiagram över residualerna. Om två parallella linjer kan läggas över och under punkterna är kravet uppfyllt.

### 3.5.2 Antagande om normalfördelning

Det andra kravet är att residualerna ska vara normalfördelade. Detta krav kan kontrolleras genom att skapa två diagram, ett histogram och ett kvantildiagram, även kallat för QQ-plot.

### 3.5.3 Antagande om oberoende

Det sista kravet är att residualerna inte ska vara autokorrelerade, alltså att de inte ska vara beroende av varandra. Kontroll av detta antagande görs först genom ett spridningsdiagram på residualerna i observationsordning. För att antagandet ska vara uppfyllt ska mönstret se slumpmässigt ut, men detta kan vara svårt att se, vilket gör att det finns andra metoder att kontrollera antagandet. SAC, SPAC och Durbin-Watson test är tre olika metoder för att undersöka om autokorrelation finns.

## 3.6 Durbin-Watson

Ett antagande för modeller är att residualerna ska vara oberoende. Detta kan kontrolleras med ett Durbin-Watson test där teststatistika kan ligga mellan 0 till 4 och 2 är ingen förekomst av autokorrelation, vilket innebär bäst resultat på testet. Förekomst av autokorrelation innebär att residualerna är beroende av varandra. Testet visas enligt ekvation 24. (Bowerman, O'Connell & Koehler 2005).

$$\frac{\sum(e_t - e_{(t-1)})^2}{\sum e_t^2} \quad (24)$$

Ovan i formeln syns  $e_t$  som är residualerna för den anpassade modellen och räknas ut enligt ekvation 25.

$$e_t = y_t - \hat{y}_t \quad (25)$$

Där  $y_t$  är det faktiska värdet i tidpunkt  $t$  och  $\hat{y}_t$  är det skattade värdet i tidpunkt  $t$ . (Bowerman, O'Connell & Koehler 2005).

## 3.7 SAC och SPAC

Sample autorcorrelation(SAC) mäter korrelationen mellan observationer i tidsserien som separeras med  $k$ -laggar, eftersom detta är en korrelation kan  $r_k$  anta värden mellan -1 och 1. Där -1 är negativ korrelation och 1 positiv korrelation, ekvationen 26 visar beräkningen för  $r_k$  och ekvation 27 visar beräkningen för medelvärdet av  $y'_t$ . (Bowerman, O'Connell & Koehler 2005).

$$r_k = \frac{\sum_{t=b}^{n-k} (y'_t - \bar{y}') (y'_{t+k} - \bar{y}')}{\sum_{t=b}^n (y'_t - \bar{y}')^2} \quad (26)$$

$$\bar{y}' = \frac{\sum_{t=b}^n y'_t}{(n - b + 1)} \quad (27)$$

Sample partial autocorrelation(SPAC) mäter korrelationen för ett urval av observationerna mot de laggade observationerna, där ekvation 28 är delberäkning av ekvation 29. (Bowerman, O'Connell & Koehler 2005).

$$r_{kk} = \begin{cases} r_1 & \text{om } k = 1 \\ (r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}) / (1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j) & \text{om } k = 2, 3, \dots \end{cases} \quad (28)$$

$$r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j} \text{ för } j = 1, 2, \dots, k-1 \quad (29)$$

## 3.8 Jämförelsemått

### 3.8.1 Förklaringsgrad

Förklaringsgrad är ett mått på hur mycket av variationen som en modell förklrar. Det beräknas genom att dividera den förklarade variationen med den totala variationen. Om förklaringsgraden ska jämföras mellan olika modeller är den justerande förklaringsgraden ett mer rättvist mått, då den straffar modeller med många variabler. (Bowerman, O'Connell & Koehler 2005).

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (30)$$

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \quad (31)$$

där  $n$  = totala antal observationer,  $p$  = antal oberoende variabler

### 3.8.2 MSE, MAD och MAPE

För att jämföra modeller kan MSE, MAD och MAPE användas som jämförelsemått. Alla dessa mått räknas ut genom att använda residualerna enligt ekvation 25. (Bowerman, O'Connell & Koehler 2005).

Det första måttet är MSE, *mean squared error* och beräknas enligt ekvation 31. (Bowerman, O'Connell & Koehler 2005).

$$MSE = \frac{1}{n} \sum_{t=1}^n (e_t)^2 \quad (32)$$

MAD, *mean absolute deviation*, mäter hur stort felet är i genomsnitt och skillnaden mellan MSE är att MAD inte har det kvadrerade felet utan endast absolutbeloppet av felet. Detta gör att större värden på  $e_t$  inte bidrar lika mycket i MAD än vad det gör i MSE. MAD beräknas enligt ekvation 32. (Bowerman, O'Connell & Koehler 2005).

$$MAD = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (33)$$

Det sista måttet, MAPE som står för *mean absolute percentage error* mäter hur många procent som prediktionen ligger från de faktiska värdena, i genomsnitt. Det beräknas enligt ekvation 33. (Bowerman, O'Connell & Koehler 2005).

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \quad (34)$$

### 3.9 Implementation i programvaror

Paket och funktioner som har använts i R redovisas här.

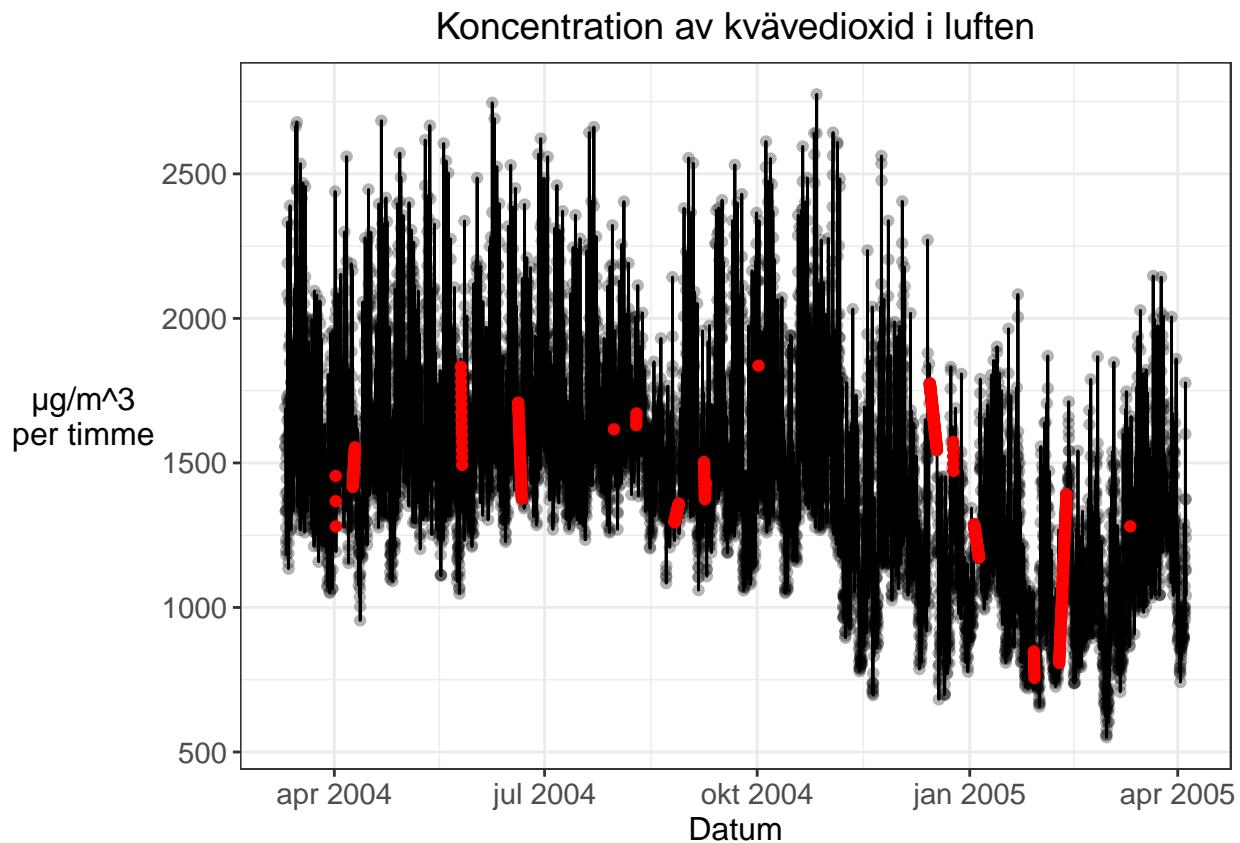
Paketet *imputeTS* (Moritz & Bartz-Beielstein, 2017) har använts där funktionen *na\_interpolation* användes för att göra linjär imputering.

Paketet *forecast* (Hyndman et al., 2022) har också använts där funktionerna *ETS*, *ARIMA* och *forecast* har använts. *ETS* räknar ut parametrarna för en Holt-Winters modell och *ARIMA* räknar ut parametrarna för en ARIMA modell. Den sista funktionen, *forecast*, genomför beräkningar för att ta fram prediktioner i framtiden med tillhörande prediktionsintervall.

## 4 Resultat

### 4.1 Imputering

Imputeringsmetoden som användes var linjär imputering. Anledningen till att denna imputering användes beror på att tidserien endast sträcker sig över drygt ett år, vilket gör att en säsongsimputering hade varit olämpligt. Istället används därför en enklare imputering, linjär imputering. I figur 2 har hela tidserien visualiseras med röda prickar som indikerar på att imputation har använts vid denna tidpunkt.



Figur 2: Visualisering av imputering

### 4.2 Multipel linjär regression

En multipel linjär regressionsmodell har skapats och består av 25 variabler, där 23 av variablerna är dummys för varje timme på ett dygn. Timme "04:00" har använts som jämförelsetimme, eftersom det var den timmen där utsläppen var som lägst, vilket gör tolkningarna tydligare. I tabell 4 redovisas alla skattade parametrar och tillhörande t-test. I tabell 5 visas olika förklarande mått för modellen.

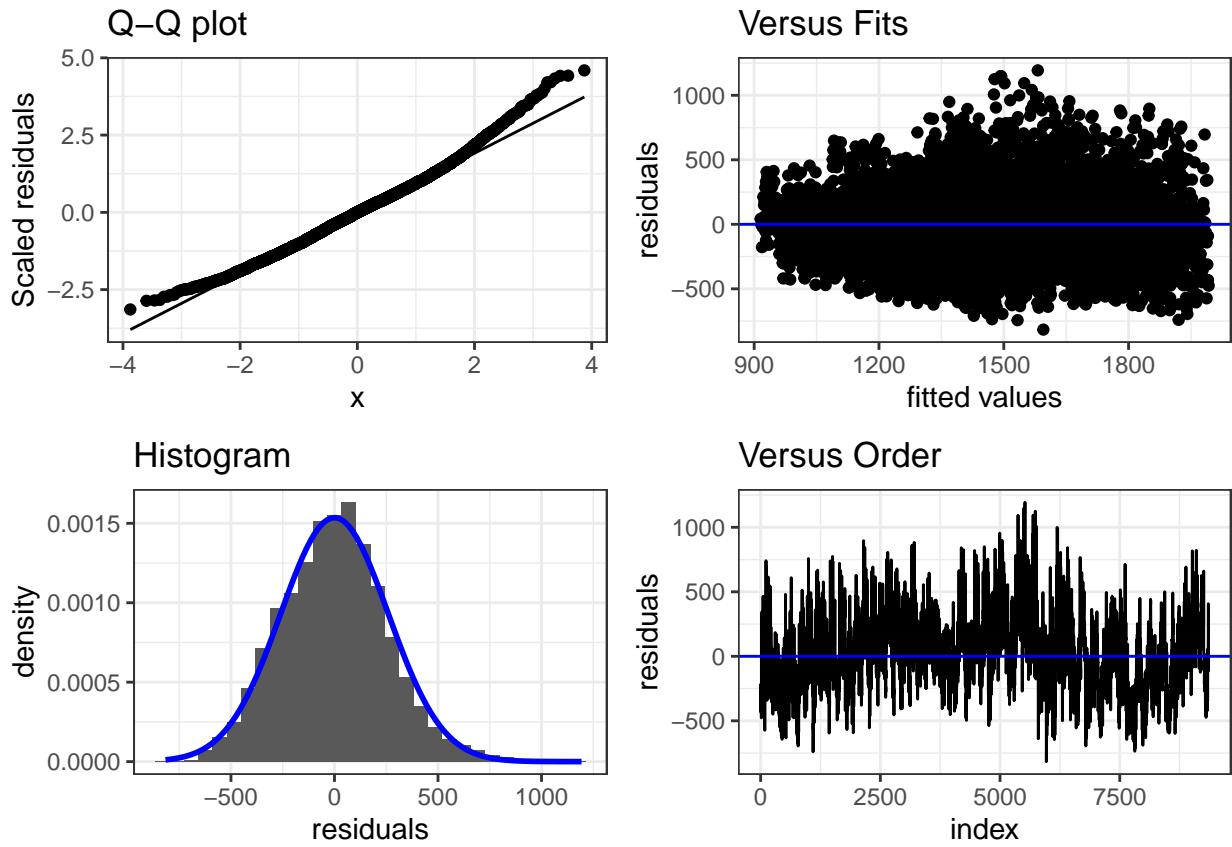
Tabell 4: Koefficienter, Multipel linjär regression

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1539.554	13.968	110.217	0.000
time	-0.067	0.001	-67.060	0.000
X00.00	158.856	18.624	8.530	0.000
X01.00	103.082	18.624	5.535	0.000
X02.00	51.813	18.624	2.782	0.005
X03.00	16.456	18.624	0.884	0.377
X05.00	9.747	18.624	0.523	0.601
X06.00	77.653	18.624	4.170	0.000
X07.00	277.239	18.624	14.886	0.000
X08.00	453.864	18.624	24.370	0.000
X09.00	391.963	18.624	21.046	0.000
X10.00	299.668	18.624	16.090	0.000
X11.00	248.476	18.624	13.342	0.000
X12.00	237.059	18.624	12.729	0.000
X13.00	237.264	18.624	12.740	0.000
X14.00	208.856	18.624	11.214	0.000
X15.00	205.604	18.636	11.033	0.000
X16.00	231.931	18.636	12.445	0.000
X17.00	321.217	18.636	17.236	0.000
X18.00	410.721	18.624	22.053	0.000
X19.00	451.462	18.624	24.241	0.000
X20.00	389.414	18.624	20.909	0.000
X21.00	261.001	18.624	14.014	0.000
X22.00	198.846	18.624	10.677	0.000
X23.00	183.481	18.624	9.852	0.000

Tabell 5: Förklarande mått, Multipel linjär regression

r.squared	r.squared.adj	MSE	MAPE	MAD
0.43	0.43	67455.47	15.11	253.26

En linjär regressionsmodell har krav som måste vara uppfyllda, vilket togs upp i metodkapitlet. För att kontrollera dessa antaganden kommer fyra diagram att undersökas som visas i figur 3.

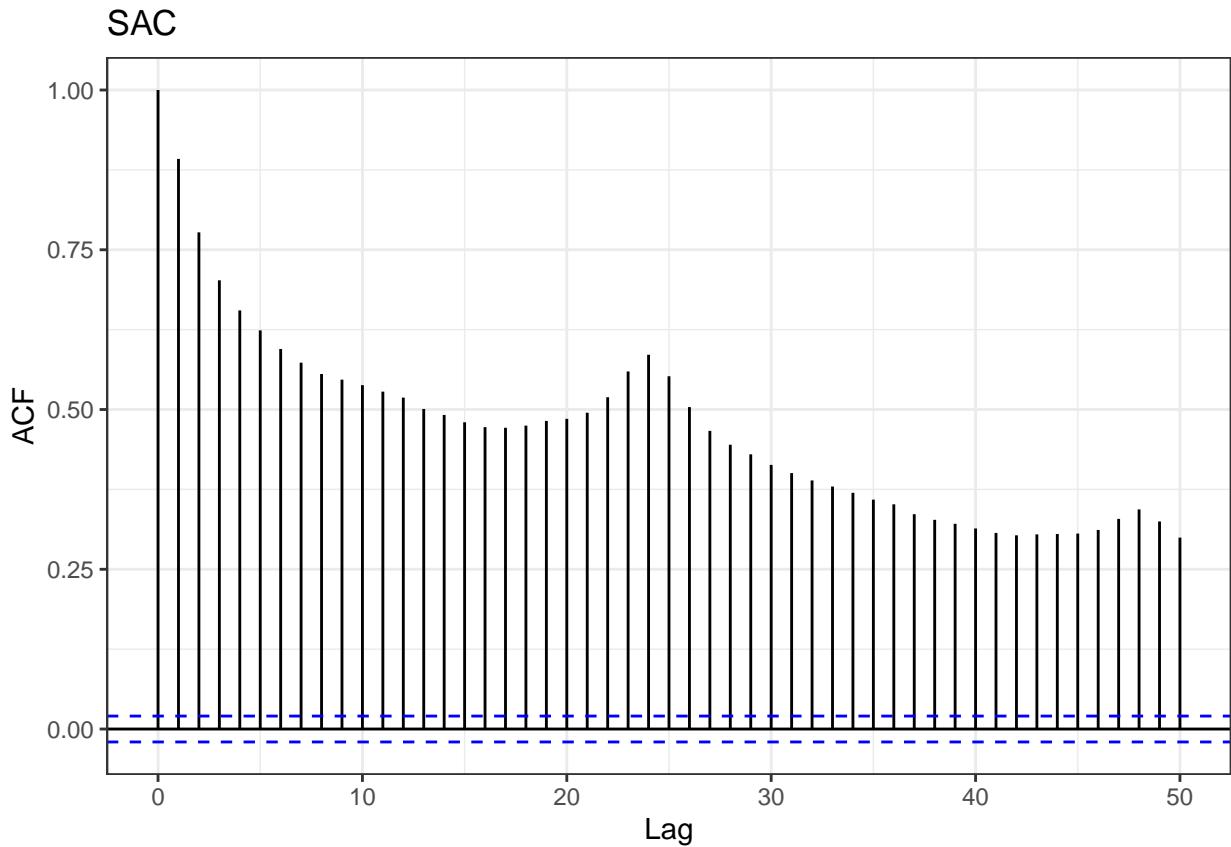


Figur 3: Residualanalys, Multipel linjär regression

I diagrammen till vänster, "Q-Q plot" och "Histogram", kan antagandet om normalfördelade residualer undersökas. I histogrammet går det att se att kurvan ser normalfördelad ut. Dock är svansen till höger är längre än den vänstra, vilket gör att residualerna inte är perfekt normalfördelade. Detta bekräftas även i "Scaled residuals" där linjen avviker lite i ändarna.

I diagrammet längst upp till höger, "Versus Fits", kan antagandet om lika varians mellan residualerna undersökas. Variansen är lite mindre i början av diagrammet, men annars är den lika stor. Antagandet kan anses vara uppfyllt.

I det sista diagrammet, "Versus order", undersöks om residualerna är beroende. Mönstret ser beroende ut, vilket indikerar på att det skulle finnas autokorrelation i residualerna. Detta kommer att undersökas noggrannare i figur 4 där SAC utreds.



Figur 4: SAC, Multipel linjär regression

Tabell 6: Durbin-Watson, Multipel linjär regression

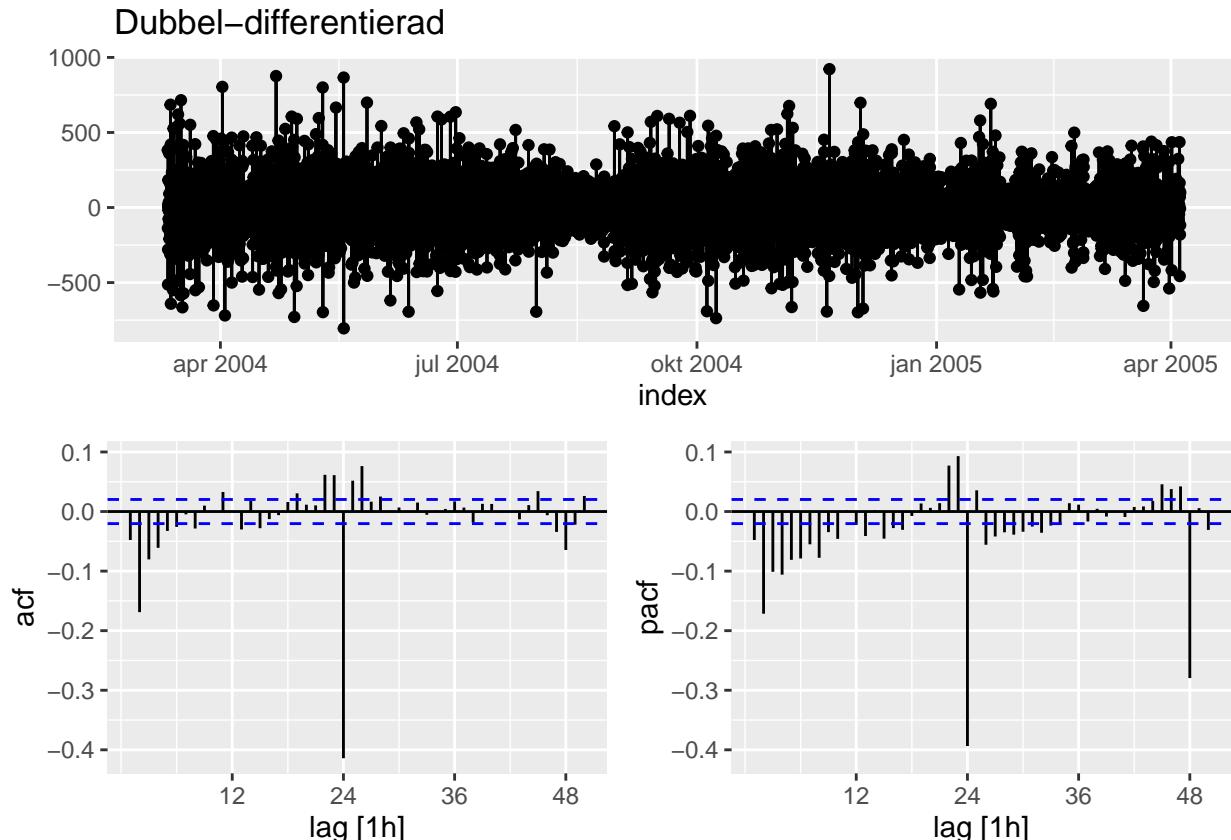
DW-test
0.21547

I figur 4 framgår det tydligt att det finns en positiv autokorrelation som minskar successivt. Detta bekräftas av test-statistikan i tabell 6, då Durbin-Watson värdet är mindre än 1. Det går även att se ett vågigt mönster, vilket indikerar på att det finns ett säsongmönster. Kravet om oberoende residualer kan således inte antas.

Veckodagar lades till som ytterligare variabler i modellen för att försöka minska autokorrelationen, men SAC fick samma mönster som i figur 4. Detta gör att prediktioner för denna modell blir totalt ointressant, då antaganden inte uppfylls. I de två nästföljande delkapitel kommer en ARIMA och en Holt-Winter modell att skattas för att försöka uppfylla antaganden.

### 4.3 ARIMA

För att modellera en ARIMA-modell behöver tidsserien vara stationär. Med tanke på att tidserien har ett säsongsmönster på timmar kommer en differentiering på 24 att göras. När detta genomfördes blev tidsserien fortfarande icke-stationär, vilket innebar att en differentiering på 1 var nödvändig. Detta känns ologiskt, då tidsserien har ett förhållandevis konstant väntevärde, men differentieringen var nödvändig för att uppfylla kravet på stationaritet. Den differentierade tidsserien visas i figur 5 med tillhörande SAC och SPAC.



Figur 5: Differentierade tidsserien, ARIMA

Nu är tidsserien stationär och analyser av modell kan påbörjas. Det är svårt att avgöra vilken ordning  $AR(p)$ ,  $MA(q)$ ,  $SAR(P)$  och  $SMA(Q)$  ska ha genom att kolla på SAC och SPAC. Detta beror förmodligen på att alla ordningar ( $p,q,P$  och  $Q$ ) är skilda från noll, vilket leder till att det är väldigt svårt att ta fram ordningar genom att endast kolla på SAC och SPAC. Detta gjorde att ett urval av SARIMA-modeller skapades och sedan jämfördes  $AIC_c$  och SAC för residualerna för de olika modellerna. Den slutgiltiga modellen blev en SARIMA(2,1,2)(1,1,1)<sub>24</sub>, då den hade lågt  $AIC_c$  och knappt någon autokorrelation i SAC.

I tabell 7 visas skattningarna av dessa komponenter och i tabell 8 har de förklarande mätten även räknats ut. Förklaringsgraden är väldigt hög, 90%, vilket menas med att 90% av tidsseriens variation förklaras av modellen. MSE, MAPE och MAD har fått betydligt lägre värden jämfört med modellen i multipel linjär regression.

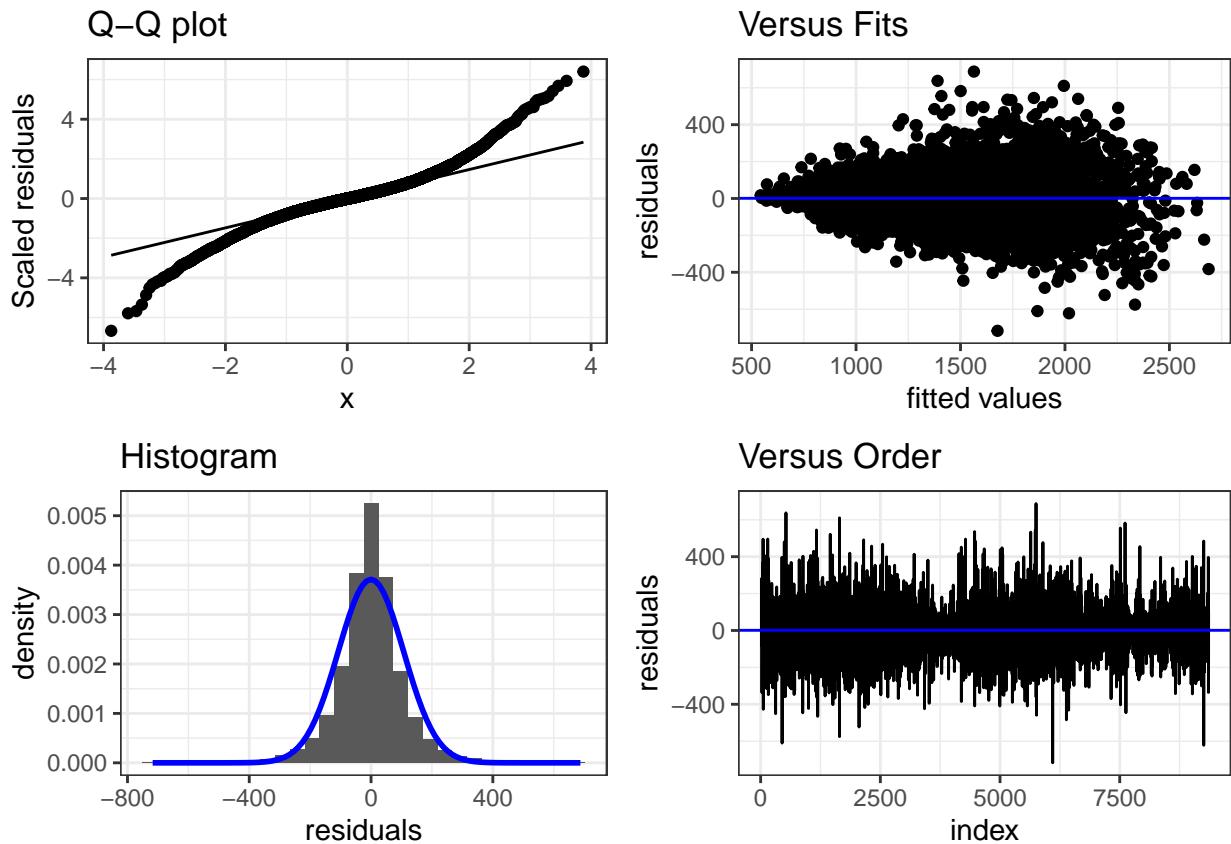
Tabell 7: Koefficienterna för SARIMA(2,1,2)(1,1,1)[24]

term	estimate	std.error	statistic	p.value
ar1	0.5404	0.0763	7.0790	0.0000
ar2	0.0880	0.0627	1.4024	0.1608
ma1	-0.6217	0.0747	-8.3271	0.0000
ma2	-0.2798	0.0722	-3.8766	0.0001
sar1	0.1310	0.0114	11.5386	0.0000
sma1	-0.9399	0.0042	-223.6465	0.0000

Tabell 8: Förklarande mått, ARIMA

r.squared	r.squared.adj	MSE	MAPE	MAD
0.90196	0.9019	11546.87	5.14332	79.20367

En SARIMA-modell har likt en linjär regressionsmodell krav som måste vara uppfyllda. För att kontrollera dessa antaganden kommer fyra diagram att undersökas som ses i figur 6.

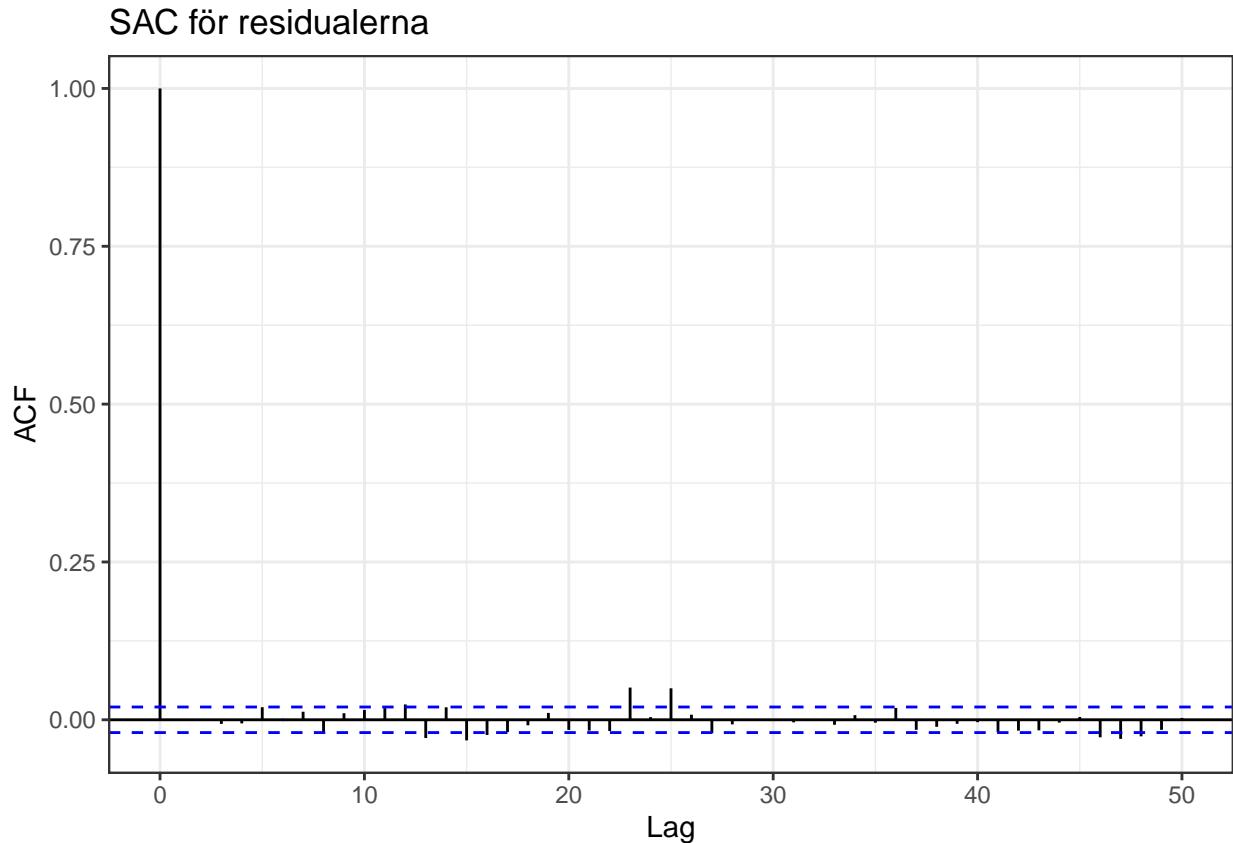


Figur 6: Residualanalys, ARIMA

Kravet om normalfördelade residualer uppfylls inte, vilket går att se i "Q-Q plot", eftersom observationerna avviker i början såväl som slutet. I histogrammet går det även att se att normalfördelnings antagandet inte uppfylls, då kurvan är väldigt spetsig och underdimensionerad.

"Versus Fits" visar på en växande varians som följs av en minskande varians. Detta gör att kravet om konstant varians inte uppfylls.

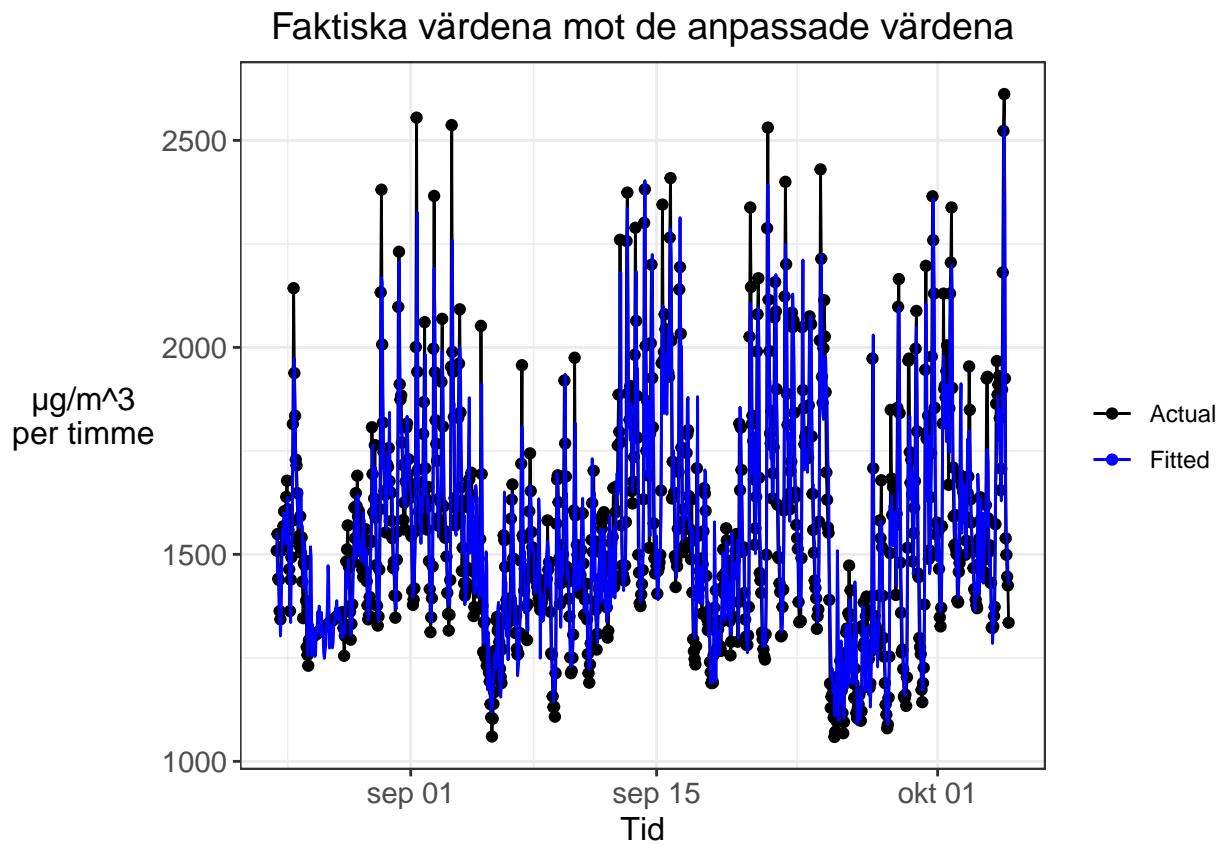
Det sista diagrammet, "Versus Order", är det svårt att avgöra om mönstret är slumpmässigt eller ej. Därför kommer SAC att undersökas i figur 7.



Figur 7: SAC, ARIMA

I diagrammet går det att avläsa att det knappt finns någon autokorrelation för residualerna. Det finns några spikar som går utanför konfidensintervallet, men dessa är endast två stycken, lag 23 och 25, och därför anses det knappt finnas någon autokorrelation. Kravet om oberoende residualer kan alltså sägas vara uppfyllt.

De faktiska värdena kommer att visualiseras mot de anpassade värdena för att få en uppfattning hur bra anpassningen är i ARIMA-modellen.

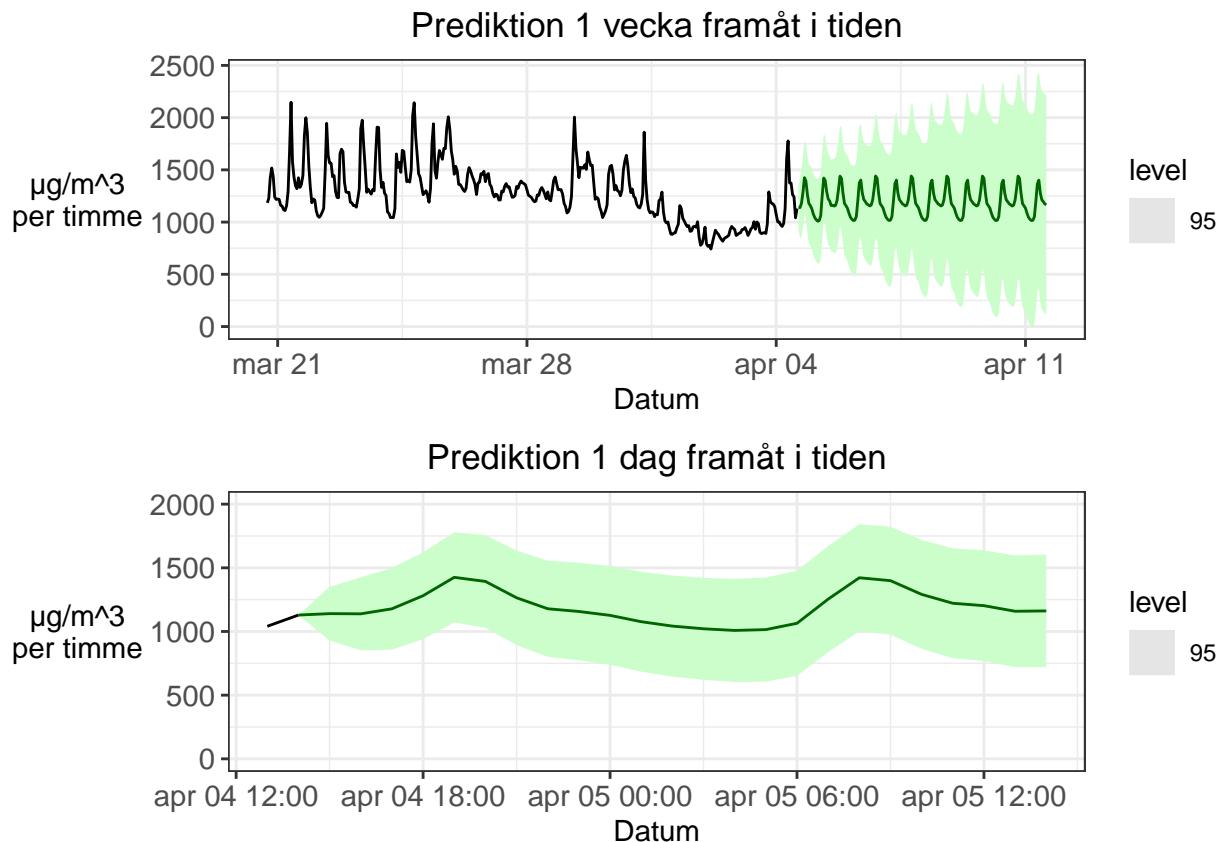


Figur 8: Faktiska värdena mot de anpassade värdena, ARIMA

I figur 8 visas de faktiska värdena och de anpassade värdena från ARIMA modellen med endast observation 4000-5000. De anpassade värden har anpassat sig väldigt bra till de observerade värdena, vilket indikerar på en överanpassning, då modellen fängar upp slump.

#### 4.3.1 Prediktion

Framtida prediktioner har gjorts för en vecka, alltså från 2005-04-04 15:00 till 2005-04-11 14:00. I figur 9 visas två diagram, det första visar en del av tidserien med de faktiska värdena tillsammans med en veckas prediktion. Det undre diagrammet visar prediktionen av ett dygn (2005-04-04 15:00 till 2005-04-05 14:00) för att få en tydligare bild hur halten av kvävedioxid förändras för varje timme. Den gröna skuggan i diagrammen är ett 95 procentigt prediktionsintervall av prediktionerna.



Figur 9: Prediktion, ARIMA

I diagrammen kan det avläsas att prediktionen har ett återkommande mönster för varje dag, vilket är rimligt med tanke på att det är en SARIMA med säsong på 24. Det går att se att halten är störst runt 07:00 till 09:00 och 19:00 till 20:00 och att den är minst vid timmarna 00:00 till 05:00. Detta känns logiskt, eftersom tidpunkterna där halten är störst är tidpunkter som människor tenderar att åka till och från jobbet. Detta innebär att trafiken förmögligen är större vid dessa tidpunkter, vilket i sin tur innebär att mer kvävedioxid släpps ut från bilar.

#### 4.4 Holt-Winters metod

En modell har skattats med Holt-Winters metod genom att jämföra  $AIC_c$  värdet för alla olika kombinationer av additiva och multiplikativa termer för felet, trenden och säsongen. M står för multiplikativ, A står för additiv och N står för att termen inte existerar i modellen. I tabell 9 syns de 10 modellerna med lägst  $AIC_c$  och den modellen som har fått lägst värde är ETS(M,A<sub>d</sub>,M). Det är alltså en modell med multiplikativt fel och säsong och en additiv dämpad trend som lämpar sig bäst för tidserien om man utgår från  $AIC_c$ .

Tabell 9: Olika modeller  $AIC_c$ -värden

ETS modell	$AIC_c$
(M,A <sub>d</sub> ,M)	173087.9
(M,N,M)	173110.4
(M,M <sub>d</sub> ,M)	173137.3
(M,M,M)	173544.5
(M,A,M)	173716.0
(M,A,A)	174520.9
(M,N,A)	174551.4
(M,M,A)	174554.0
(M,M <sub>d</sub> ,A)	174559.9
(M,A <sub>d</sub> ,A)	174560.0

I tabell 10 och 11 syns skattningarna av koefficienterna för modellen.  $\beta$  har fått ett värde väldigt nära 0, vilket innebär att trenden är linjär. Trenden är dämpad, men  $\phi$  har fått ett värde på 0.98, vilket tolkas som att dämpningen är minimal och knappt har någon påverkan.  $\alpha$  har fått värdet 0.79 och kan tolkas som att nivån uppdateras efter varje observation. Den sista utjämningsparametern  $\gamma$  har fått värdet 0.17, vilket innebär att säsongen är fix. De resterande koefficienterna är de initiala värdena  $\ell_0$ ,  $b_0$ ,  $s_0$ ,  $s_{-1}, \dots, s_{-23}$  och de tolkas som startvärde för respektive komponent. Förklarande mått har även räknats ut och visas i tabell 13.

Tabell 10: Koefficienterna för Holt-Winter (1-15)

$\alpha$	$\beta$	$\gamma$	$\phi$	$\ell_0$	$b_0$	$s_0$	$s_{-1}$	$s_{-2}$	$s_{-3}$	$s_{-4}$	$s_{-5}$	$s_{-6}$	$s_{-7}$	$s_{-8}$
0.79	0	0.17	0.98	1413.55	5.27	0.92	1	0.95	1.01	1	1.05	1.07	1.07	1.2

Tabell 11: Koefficienterna för Holt-Winter (16-30)

$s_{-9}$	$s_{-10}$	$s_{-11}$	$s_{-12}$	$s_{-13}$	$s_{-14}$	$s_{-15}$	$s_{-16}$	$s_{-17}$	$s_{-18}$	$s_{-19}$	$s_{-20}$	$s_{-21}$	$s_{-22}$	$s_{-23}$
1.18	0.92	0.84	0.86	0.85	0.86	0.88	0.94	0.96	0.99	0.93	1.1	1.18	1.14	1.09

Tabell 12: Typ av modell, Holt-Winter

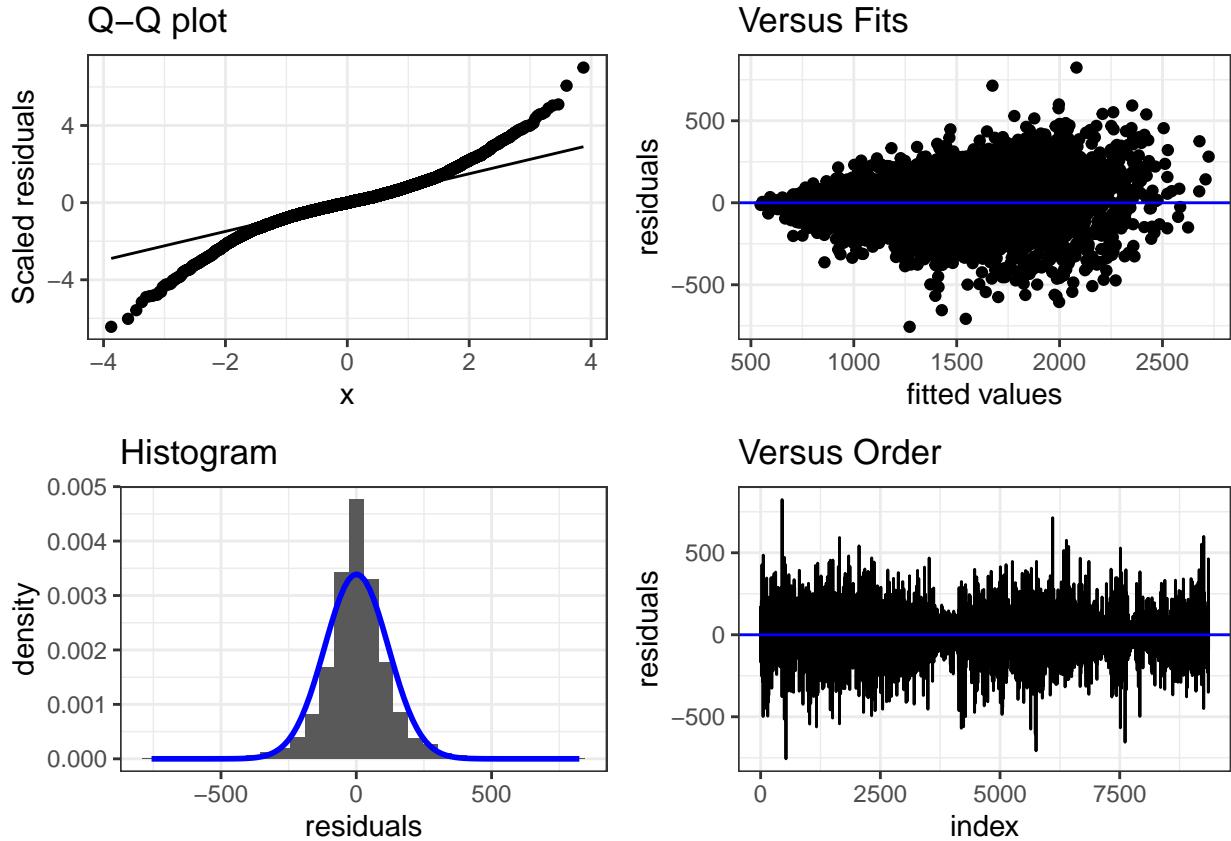
errortype	trendtype	seasontype	damped	period
M	A	M	TRUE	24

Förklaringsgraden har fått ett värde på 88%, vilket är 2 procentenheter lägre än ARIMA-modellen. MSE, MAPE och MAD har fått värden som är lite högre än ARIMA-modellen. Detta innebär att Holt-Winter modellen inte anpassar sina värden lika väl till de faktiska värdena jämfört med ARIMA-modellen.

Tabell 13: Förklarande mått, Holt-Winter

r.squared	r.squared.adj	MSE	MAPE	MAD
0.88258	0.88254	13829.84	5.64418	88.22991

Residualanalys ska även göras på Holt-Winter modellen för att undersöka om modellen uppfyller kraven.

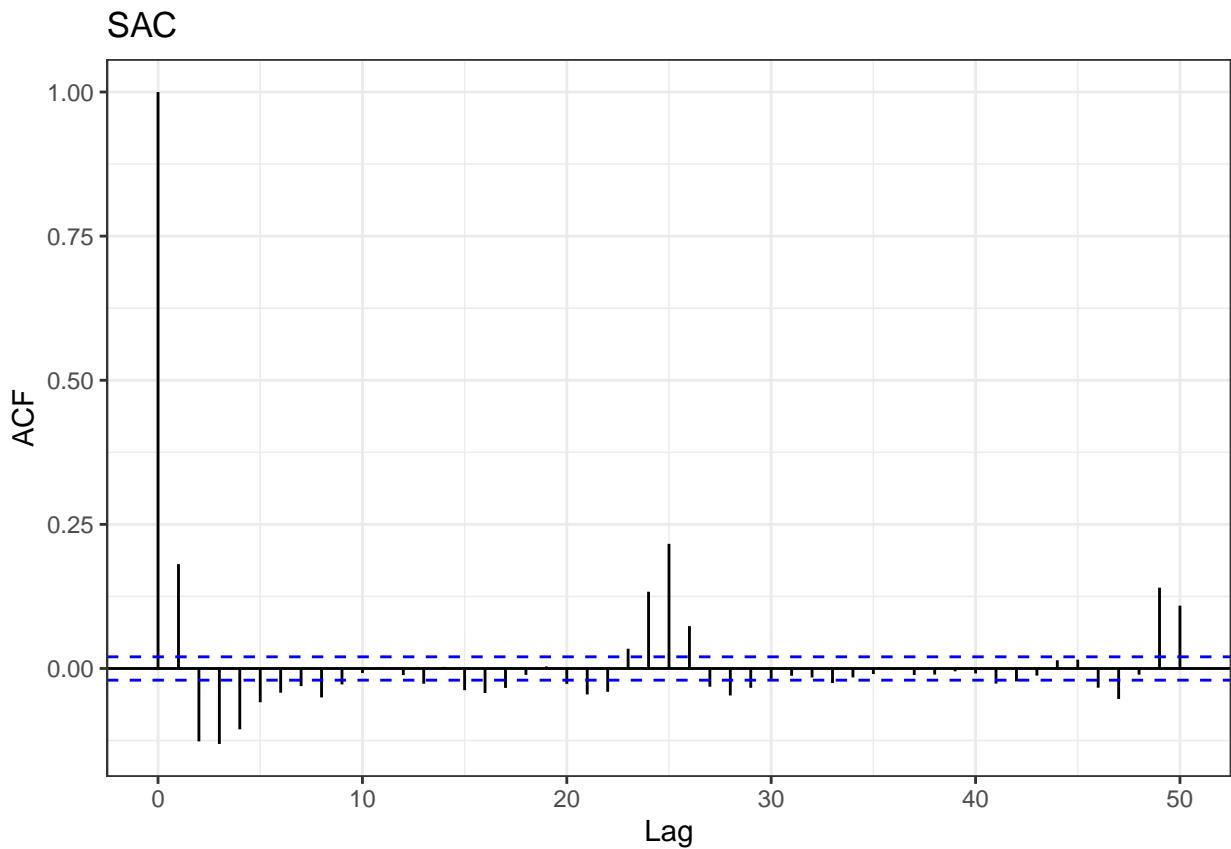


Figur 10: Residualanalys, Holt-Winter

I diagrammet "Versus Fits" syns en ojämn varians för residualerna, vilket gör att kravet om konstant varians inte uppfylls.

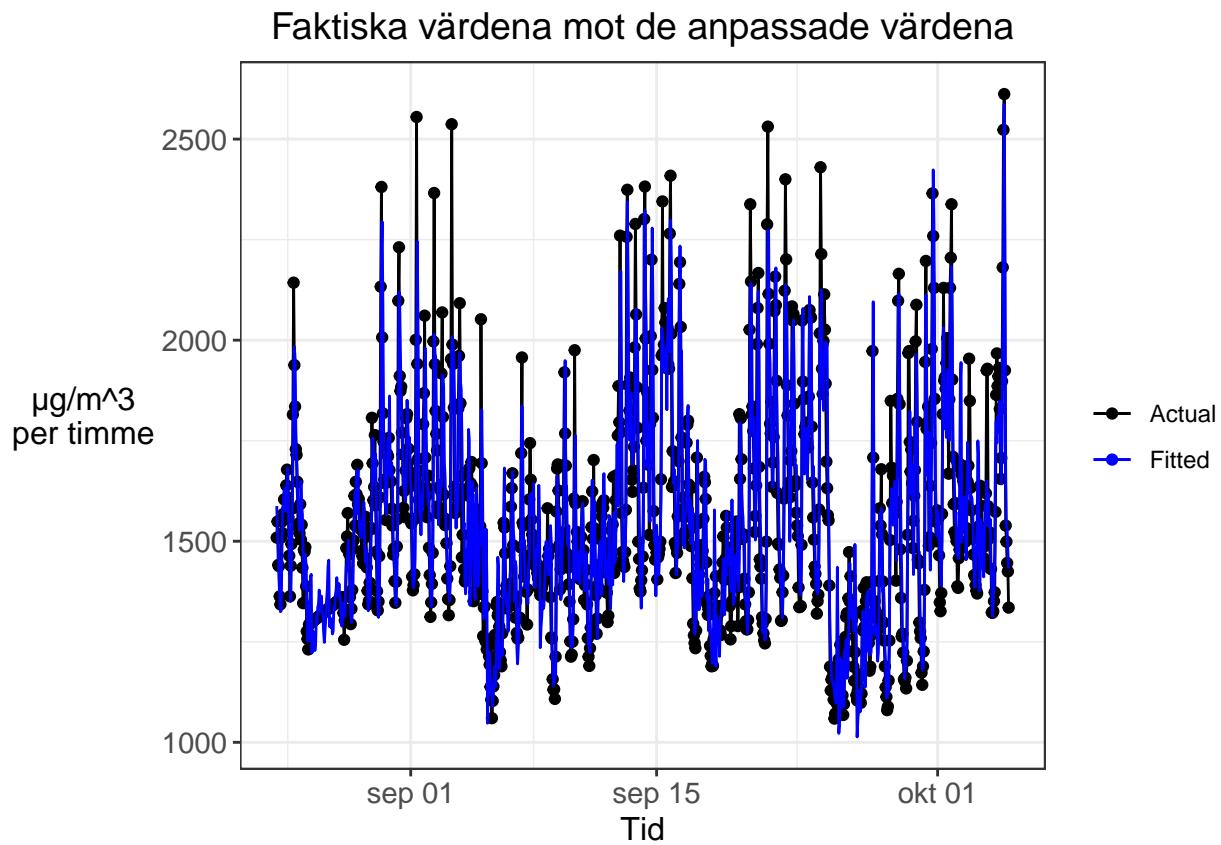
Vidare kan kravet om normalfördelade residualer undersökas i "Q-Q plot" och "Histogram". Det går tydligt att se att observationerna avviker i båda ändarna i "Q-Q plot" och att histogrammet är spetsigt med långa svansar. Detta innebär att kravet om normalfördelade residualer inte uppfylls.

Det sista diagrammet, "Versus Order" är återigen svårtolkat, då det är svårt att avgöra om mönstret är slumpmässigt eller ej. Detta undersöks vidare i figur 11 där SAC visas.



Figur 11: SAC, Holt-Winter

Det finns signifikanta spikar vid varje 24:e lag, vilket medför att det finns autokorrelation i residualerna. Holt-Winter har alltså inte lyckats få bort autokorrelation för säsongen och antagandet kan ej uppfyllas helt. Dock är autokorrelation betydligt mindre jämfört med modellen i multipel linjär regression. Prediktioner kommer att göras, men dessa bör analyseras försiktigt.

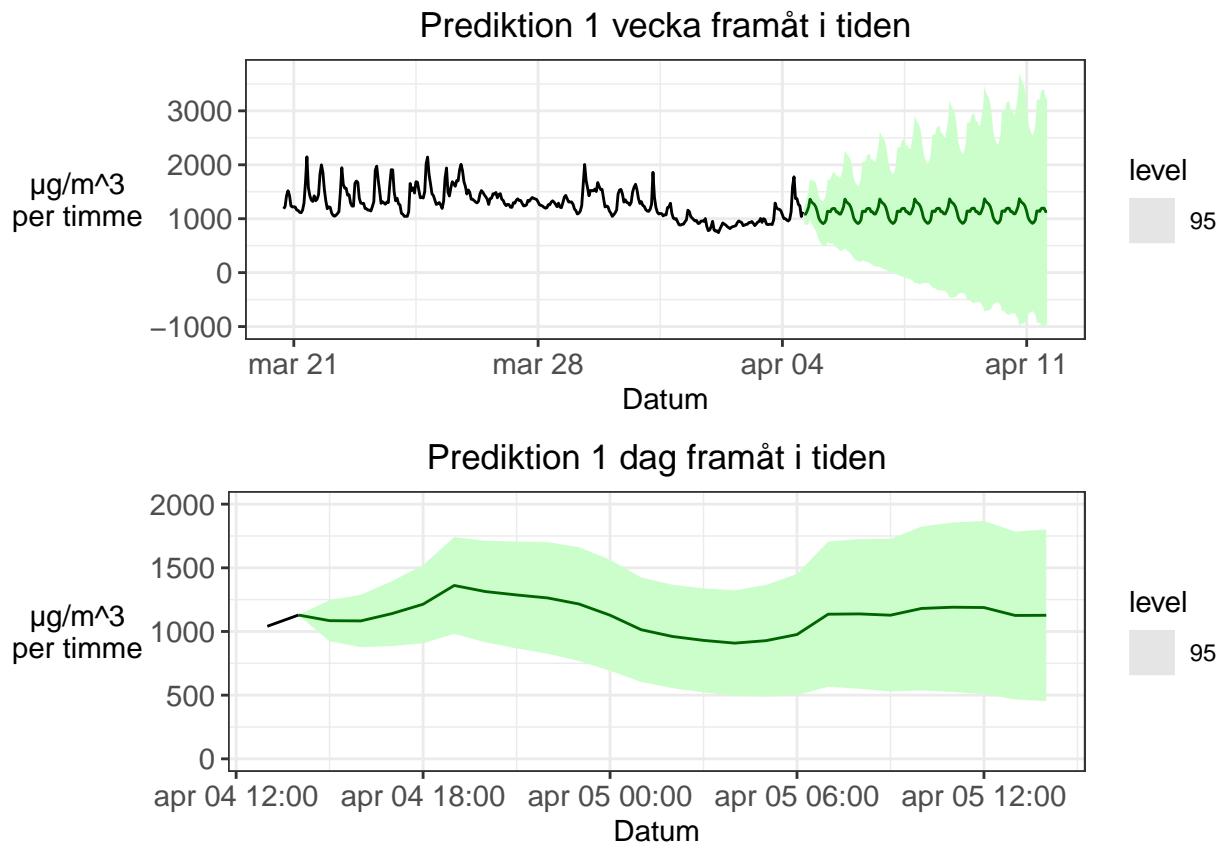


Figur 12: Faktiska värdena mot de anpassade värdena, Holt-Winter

I figur 12 visas de faktiska värdena och de anpassade värdena från Holt-Winter modellen. Även i denna modell visas endast observation 4000-5000. De anpassade värden har anpassat sig väldigt bra till de observerade värdena. Likt ARIMA-modellen indikerar detta på en överanpassning.

#### 4.4.1 Prediktion

Framtida prediktioner har gjorts för en vecka, alltså från 2005-04-04 15:00 till 2005-04-11 14:00. I figur 13 visas två diagram, det första visar en del av tidserien med de faktiska värdena tillsammans med en veckas prediktion. Det undre diagrammet visar prediktionen av ett dygn (2005-04-04 15:00 till 2005-04-05 14:00) för att få en tydligare bild hur halten av kvävedioxid förändras för varje timme. Den gröna skuggan i diagrammen är ett 95 procentigt prediktionsintervall av prediktionerna.



Figur 13: Prediktion, Holt-Winter

Det går att se att halten av kvävedioxid är störst runt 19:00-20:00 och att den är minst vid timmarna 03:00 till 05:00. Intervallet blir snabbt ganska brett, vilket innebär att prediktioner längre fram i tiden är osäkra skattningar.

## 5 Diskussion

Av de tre modellerna som har anpassats är det ARIMA-modellen som har valts som den bästa modellen för att prediktera halten av kvävedioxid. Det finns dock brister i ARIMA-modellen, ej nomalfördelade och icke-konstant varians i residualerna, vilket gör att det inte går att lita på prediktionerna i framtiden. Prediktionerna ger en indikation på framtiden, men de bör inte användas för att befästa slutsatser för framtiden. Den höga förklaringsgraden (90%) är ytterligare ett problem med ARIMA-modellen. Oftast är en hög förklaringsgrad att föredra, men när den går upp mot 90 procent kan det vara dåligt. Detta beror på att modellen har anpassat modellen för bra på träningsdata, vilket gör att slump har plockats upp av modellen. Detta gick tydligt att se i figur 8 där de faktiska värdena visualiseras mot de anpassade värdena.

En Holt-Winter modell skattades också, men modellen blev sämre än ARIMA-modellen, eftersom det fanns autokorrelation i Holt-Winter modellen. En förbättring på denna rapport skulle kunna vara att ändra tillvägagångssättet hur modell väljs till Holt-Winter, då denna rapport endast utgått från  $AIC_c$ . Ett förslag hade varit att ta fram fler mått för att jämföra modellerna åt och att jämföra residualerna för varje modell. Detta hade kanske resulterat i en modell med ingen autokorrelation, vilket hade varit en förbättring jämfört med den modell som har valts i denna rapport.

Den första modellen som skattades, multipel linjär regression, fick tydlig positiv autokorrelation, vilket gjorde att den modellen inte var av intresse när de kommer till prediktioner. Detta beror på att prediktionerna inte är sammhetsenliga och inga säkra slutsatser kan dras. Ett försök med att inkludera båda timmar och veckodagar som dummyvariabler genomfördes, men den positiva autokorrelationen fanns fortfarande kvar.

Ett möjligt problem till att det var problematisk att hitta en bra modell skulle kunna vara serien endast sträcker sig över ett år. En tidserie över flera år hade gjort det lättare att det identifiera trender och säsongsmönster för månader. Just nu är det omöjligt att säga att det finns säsongsmönster för månader, då det endast finns data för ett år. Om tidserien hade sträckt sig över 3 år hade ett potentiellt säsongsmönster för månader kunnat analyserats. Detta hade öppnat möjligheter för andra modeller, speciellt TBATS-modellen som kan hantera två säsongsmönster enligt Hyndman (Ottexts, 2018), i detta fall timmar och månader.

I rapporten *Predicting air quality using ARIMA, ARFIMA and HW smoothing* (Nimesh et. al., 2014) har prediktioner om luftkvalité genomförts. De använde tre olika modeller för att ta fram prediktioner, ARIMA, ARFIMA och Holt-Winter, vilket liknar denna rapportens modeller. Rapporten kom fram till att ARFIMA var den modell som var bäst lämpad för prediktioner, vilket är intressant, då denna modell ej testades i vår rapport. ARFIMA är en ARIMA-modell, men där "differentiering-parametern"  $d$  inte behöver vara ett heltal.

## 6 Slutsats

Syftet med rapporten var att prediktera mängden av kvävedioxid som släpps ut i en italiensk stad. För att besvara syftet ställdes följande frågeställning upp.

- Hur ser en lämplig modell ut för att prediktera kvävedioxidhalten i en italiensk stad 1 vecka framåt?

Det finns ingen lämplig modell för att prediktera kvävedioxidhalten, vilket innebär att det inte finns några säkra slutsatser om kvävedioxidhalten i en italiensk stad 1 vecka framåt.

## Referenser

- Andersson, N. 2022. *Kväveoxider*. Karolinska Institutet.  
<https://ki.se/imm/kvaveoxider>
- Bayen, A. & Siauw, T., 2015. *Linear Interpolation*. Sciencedirect.  
<https://www.sciencedirect.com/topics/engineering/linear-interpolation>
- Bowerman, B., O'Connell, R. & Koehler, A., 2005. *Forecasting, Time Series, and Regression*. USA: Brooks/Cole Cengage Learning.
- Hyndman, R. & Athanasopoulos, G., 2018. *Forecasting: Principles and Practice*. Otexts. <https://otexts.com/fpp2/>
- Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos F., Razbash, S., Wang, E. & Yasmeen, F. 2022. *Forecasting functions for time series and linear models*
- Naturvårdsverket. *Fakta om kväveoxider i luft*. <https://www.naturvardsverket.se/amnesomraden/luft/luftfororeningar-och-dess-effekter/fakta-om-kvaveoxider-i-luft>
- Moritz, S & Bartz-Beielstein, T. 2017. *imputeTS: Time Series Missing Value Imputation in R*.
- Ruby, N., Arora, S., Kusum Mahajan, K. & Nath Gill, A. 2014. *Predicting air quality using ARIMA, ARFIMA and HW smoothing*. [https://www.researchgate.net/publication/287930752\\_Predicting\\_air\\_quality\\_using\\_ARIMA\\_ARFIMA\\_and\\_HW\\_smoothing](https://www.researchgate.net/publication/287930752_Predicting_air_quality_using_ARIMA_ARFIMA_and_HW_smoothing)
- UCI Machine Learning Repository. 2016. *Air Quality Data Set*.  
<https://archive.ics.uci.edu/ml/datasets/Air+Quality>

## 7 Bilaga

### 7.1 Bilaga A

I bilaga A visas den R-kod som har använts för att hantera datamaterialet.

```
#-----LADDAR HEM DATAN-----
luft <- read.csv2("C:\\\\Users\\\\willi\\\\OneDrive\\\\Dokument\\\\AirQualityUCI\\\\AirQualityUCI.csv")
luft[luft == "-200"] <- NA
luft <- luft[1:9357, c(1,2,11,13:15)]

# Ersätter NA-värden med linjär imputation
luft_linear <- na_interpolation(luft$PT08.S4.N02., option = "linear")

# Skapar en tidssekvens som matchar det givna data
ts_hour <- as.POSIXct(timeSequence(from = "2004-03-10 18",
                                to = "2005-04-04 14", by = "hour"))

# Skapar ett ts-objekt
ts_luft <- ts(as.vector(luft_linear), start=c(2004, 1674),
               end = c(2005,2270), frequency=24*365)
```

### 7.2 Bilaga B

I bilaga B innehåller R-kod som har använts för att anpassa modellen med multipel linjär regression.

```
#-----
#----- MULTIPEL LINJÄR REGRESSION -----
#-----

# Skapar en designmatris
hour_val <- c(18:23, rep(0:23, 389) ,0:14)
X_hour <- matrix(0, nrow = 9357, ncol = 24)
colnames(X_hour) <- c("00:00","01:00","02:00","03:00","04:00","05:00","06:00",
                      "07:00","08:00","09:00","10:00","11:00","12:00","13:00",
                      "14:00","15:00","16:00","17:00","18:00","19:00","20:00",
                      "21:00","22:00","23:00")

for(i in 0:23){
  X_hour[,i+1] <- ifelse(hour_val == i, 1, 0)
}

X <- cbind(time = 1:9357, X_hour[,-5])
reg_data <- data.frame(temp = as.vector(luft_linear), X)

# Skapar modellen med hjälp av designmatrisen
```

```

lm_temp2 <- lm(temp~, data = reg_data)

# Räknar ut förklarande mått till modellen
rsq <- summary(lm_temp2)$r.squared
rsq_ad <- 1 - (((1-rsq)*(9357-1))/(9357-24-1))
mse <- mean(lm_temp2$residuals^2)
mape <- MAPE(lm_temp2$residuals,luft$PT08.S4.N02.)
mad <- mad(lm_temp2$residuals)

```

### 7.3 Bilaga C

I bilaga C innehåller R-kod som har använts för att anpassa ARIMA-modellen.

```

#-----
#-----ARIMA-----
#-----

df <- data.frame(index = seq(ymd_h("2004-03-10 18"), ymd_h("2005-04-04 14"), by = "hour"),
                  value = ts_luft)

fit <- as_tsibble(df) %>%
  model(
    best = ARIMA(value ~ pdq(2,1,2) + PDQ(1,1,1,24))
  )

# Plockar ut residualerna och de anpassade värdena
resi <- fit$best[[1]]$fit$est$.resid
fitted <- fit$best[[1]]$fit$est$.fitted

# Beräknar de förklarande mätten
rss <- sum((fitted-as.vector(luft_linear))^2)
tss <- sum((as.vector(luft_linear)-mean(as.vector(luft_linear)))^2)
rsq_arima <- 1 - rss/tss
rsq_adj_arima <- 1 - (((1-rsq_arima)*(9357-1))/(9357-6-1))

mse_arima <- mean(resi^2)
mape_arima <- MAPE(resi,as.vector(luft_linear))
mad_arima <- mad(resi)

```

### 7.4 Bilaga D

I bilaga D innehåller R-kod som har använts för att anpassa Holt-Winter modellen.

```

#-----
#-----HOLT-WINTER-----
#-----

ETS_MAM_damped <- as_tsibble(df) %>%
  model(ETS(value ~ error(method = "M") + trend(method = "Ad") + season(method = "M")))

ETS_MAM_damped_aicc <- ETS_MAM_damped[[1]][[1]]$fit$fit$AICc

fit <- ETS_MAM_damped

# Tar ut de anpassade värdena och residualerna
fitted <- fit[[1]][[1]]$fit$est$.fitted
resi <- fitted - as.vector(luft_linear)

# Räknar ut de förklarande mätten för modellen
rss <- sum((fitted-as.vector(luft_linear))^2)
tss <- sum((as.vector(luft_linear)-mean(as.vector(luft_linear)))^2)
rsq_holt <- 1 - rss/tss
rsq_adj_holt <- 1 - (((1-rsq_holt)*(9357-1))/(9357-3-1))

mse_holt <- mean(resi^2)
mape_holt <- MAPE(resi, as.vector(luft_linear))
mad_holt <- mad(resi)

```