Eindopdracht

Mohammed Al Hor

2022-12-24

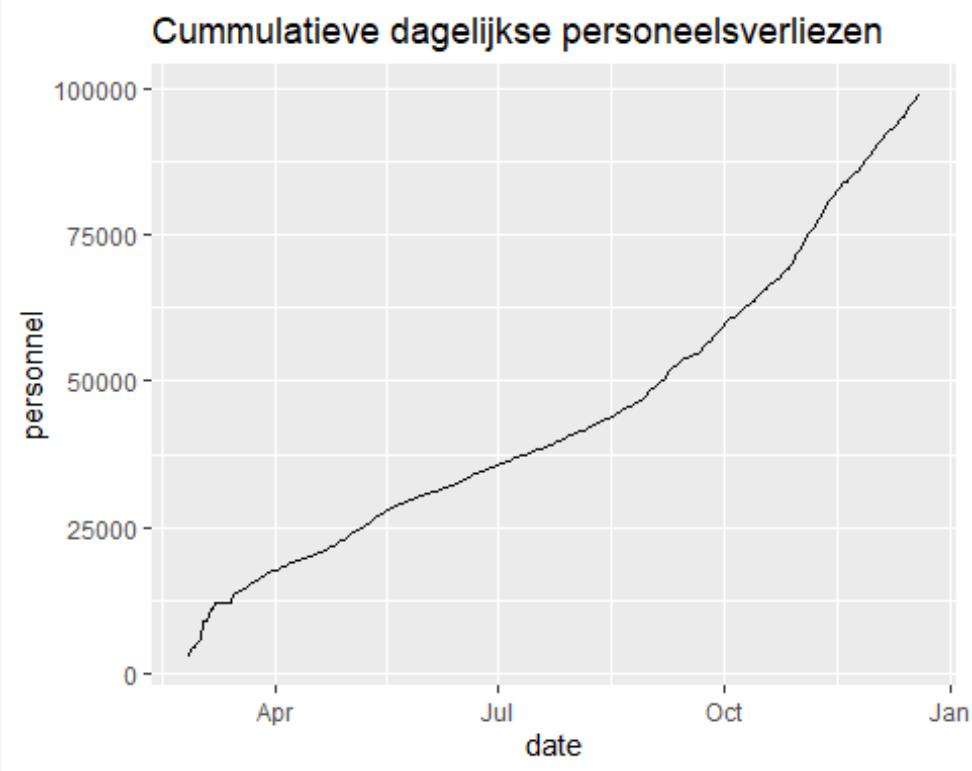
Introductie

Deze zeer relevante dataset bevat de dagelijkse personeelsverliezen aan Russische kant van de oorlog in Ukraine. In dit rapport komen de volgende onderwerpen aan bod: (I) EDA (Exploratory Data Analysis), (II) forecasting waarbij meerdere modellen worden beschouwd. Deze modellen gaan we met elkaar vergelijken om uiteindelijk het beste model uit te kiezen en deze te gebruiken om voorspellingen te doen.

Deelvraag 1 & 2

Onderstaand figuur geeft ons een eerste blik op de data. Omdat we het hier hebben over cumulatieve personeelsverliezen zien we uiteraard een stijgende lijn in dit aantal.

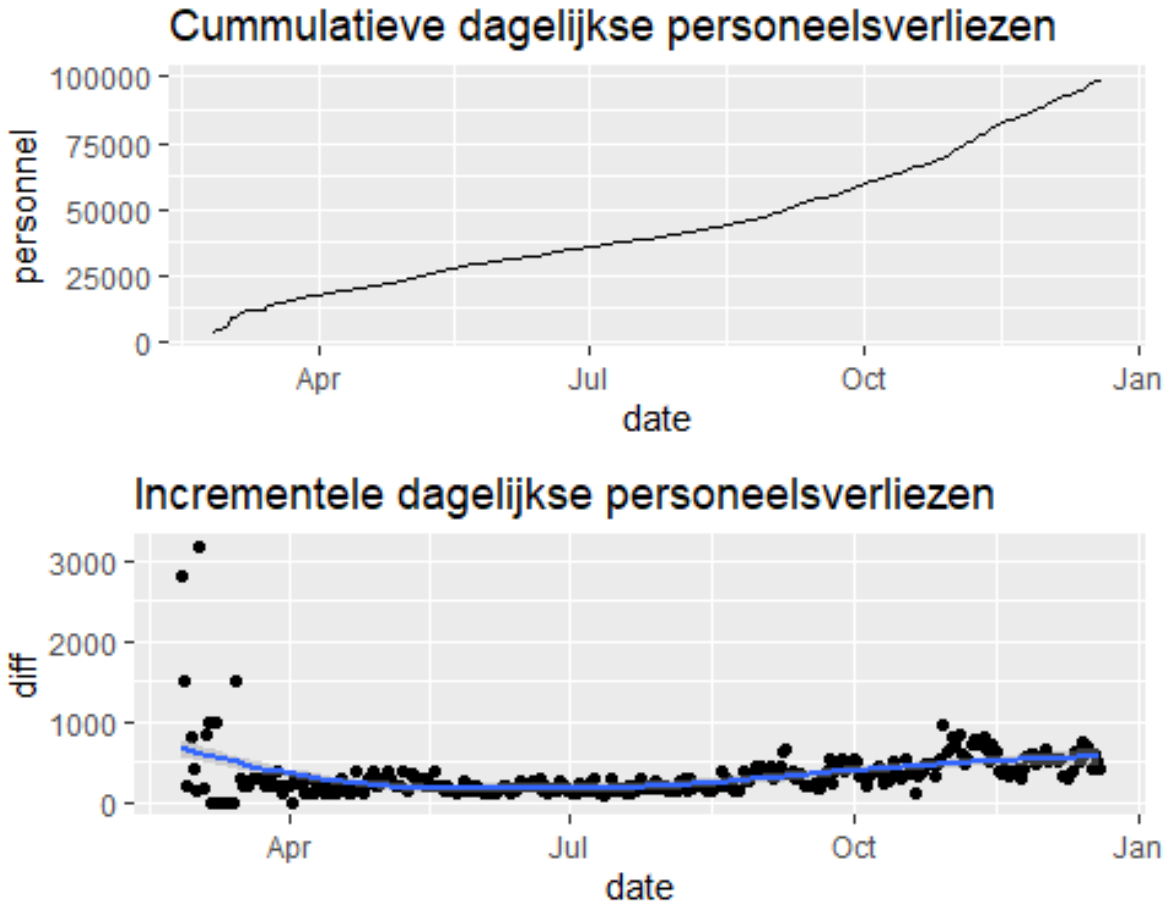
```
plot1 <- ggplot(russia_losses_personnel, aes(x=date, y=personnel)) +geom_line
() + labs(title = "Cummulatieve dagelijkse personeelsverliezen")
plot1
```



De eerste verschillen (*first differences*) worden berekend, zodat we deze kunnen visualiseren.

```
russia_losses_personnel <- russia_losses_personnel %>% mutate(diff = personne
l - lag(personnel))
```

Cumulatieve en dagelijkse/incrementele personeelsverliezen onder elkaar.

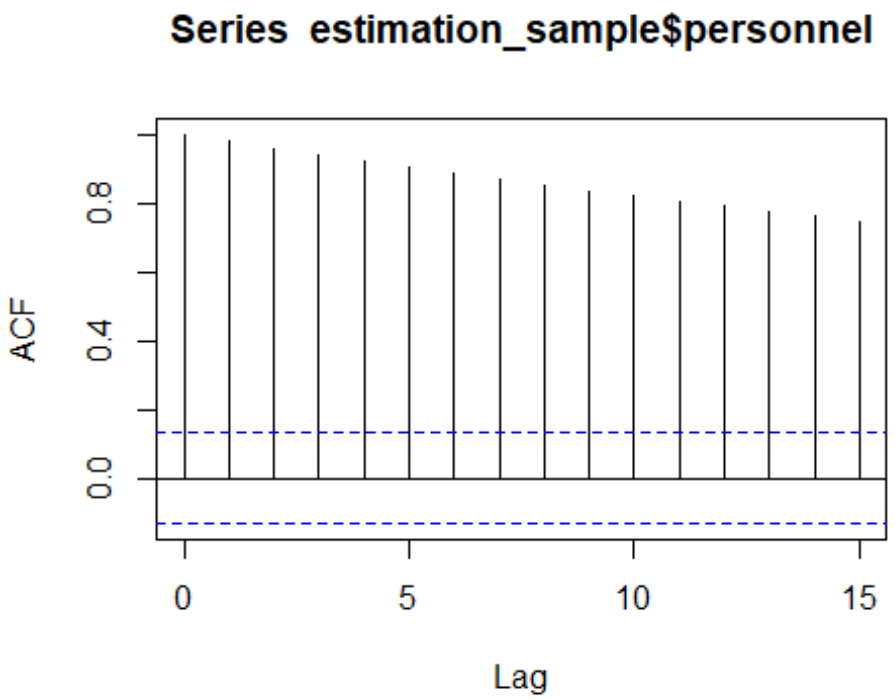


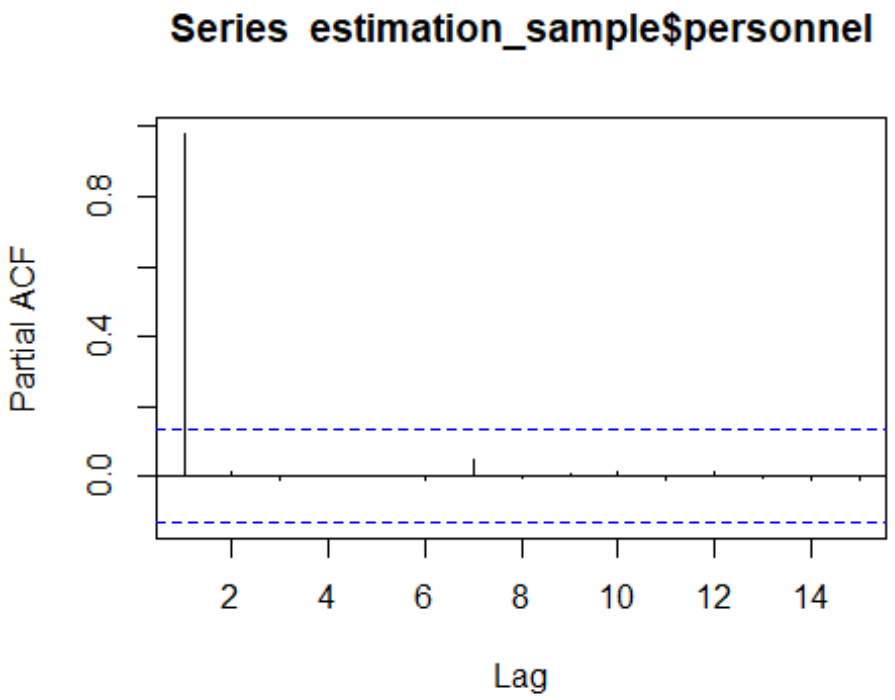
We observeren een aanzienlijke variantie wat betreft personeelsverliezen in de eerste dagen van de oorlog. Een verklaring voor deze variantie is het feit dat deze aan het begin niet dagelijks werden geüpdatet. We zien voor sommige dagen nul personeelsverliezen die in de opvolgende dagen worden overgecompenseerd. Verder zien we dat de trend die we voorheen zagen in de cumulatieve personeelsverliezen grotendeels is verdwenen, de data lijkt stationair. Dit zullen we in de volgende stappen gaan toetsen.

De dataset wordt opgedeeld in een 'estimation sample', die we gaan gebruiken om de modellen te trainen en een 'test sample' die we gaan gebruiken om de modellen te evalueren.

```
estimation_sample <- russia_losses_personnel %>% filter(date <= '2022-10-03')
test_sample <- russia_losses_personnel %>% filter(date > '2022-10-03')
```

We berekenen de (partiële) autocorrelaties voor de cumulatieve personeelsverliezen(niveau):





De ACF of autocorrelatie functie in bovenstaand figuur geeft de correlatie weer over de tijd tussen de verschillende momenten. We observeren dat de eerste 15 momenten significant en afnemend zijn. Dit wekt de suggestie dat er een lange termijn trend in de data zit en deze dus niet stationair is. We zouden nog een ADF (Augmented Dickey Fuller) test kunnen doen om dit te bevestigen.

```
adf.test(estimation_sample$personnel)

##
##  Augmented Dickey-Fuller Test
##
## data:  estimation_sample$personnel
## Dickey-Fuller = 1.0182, Lag order = 6, p-value = 0.99
## alternative hypothesis: stationary
```

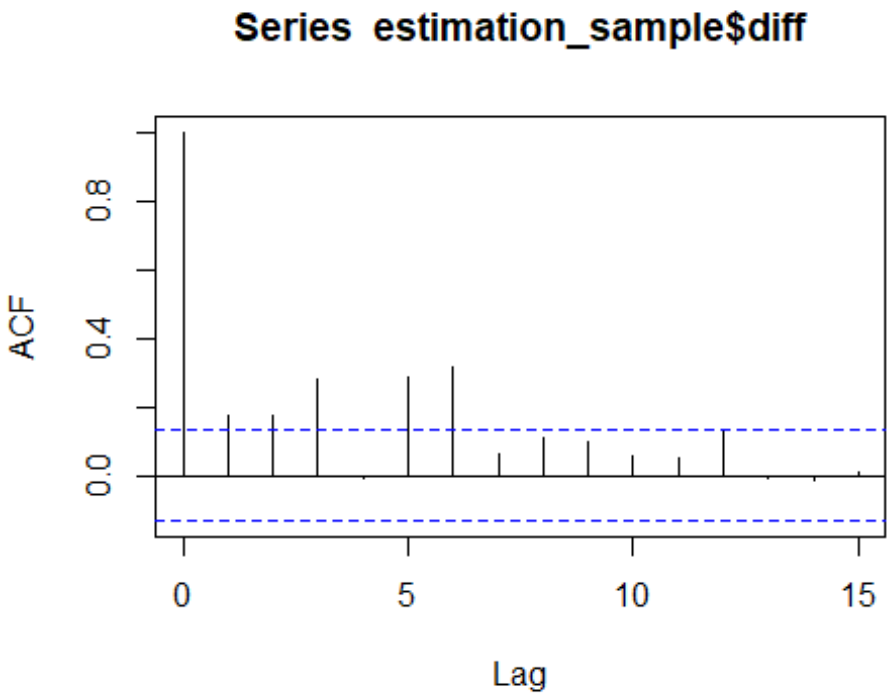
De hoge p-waarde maakt het dat we de null-hypothese niet kunnen verwerpen (p>0.05). Conform onze verwachtingen zijn de cumulatieve personeelsverliezen niet stationair.

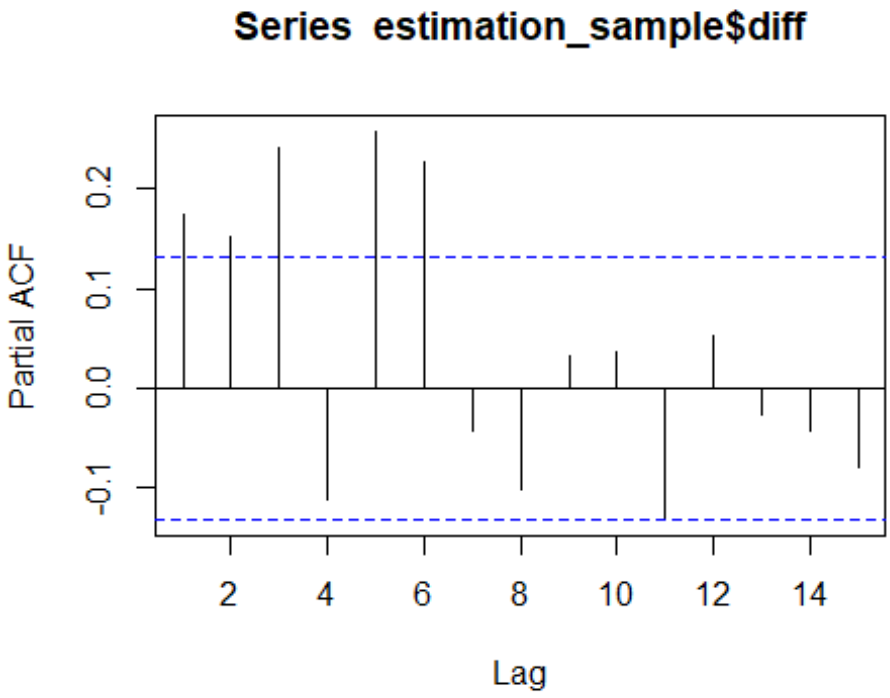
Deelvraag 3

Omdat cumulatieve personeelsverliezen niet stationair zijn gaan we door met de verschillen. We passen hiervoor een methode toe die 'detrending' of 'first differencing' wordt genoemd. Dit stukje feature engineering hebben we in de vorige stappen al gedaan.

Deelvraag 4

Laten we nu gaan kijken naar de (partiële) autocorrelaties voor incrementele/dagelijkse personeelsverliezen.





De ACF toont dat de 1ste, 2de, 3de, 5de en 6de momenten significant zijn. De partiële ACF toont dat de 1ste, 2de, 3de, 5de en 6de momenten significant zijn. Deze informatie kunnen we gebruiken bij het modelleren van deze data.

Deelvraag 5

Voordat we met het leuke werk gaan beginnen moeten we eerst een check doen of onze data stationair is. Dit houdt in dat de mean en de variantie constant moeten zijn en er geen seizoen effecten in mogen zitten. Hiervoor gebruiken we de ADF, ofwel 'Augmented Dickey Fuller' test.

```
adf.test(estimation_sample$diff)

##
##  Augmented Dickey-Fuller Test
##
## data:  estimation_sample$diff
## Dickey-Fuller = -8.3393, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

P-waarde is kleiner dan 0.05, we kunnen de null-hypothese verwerpen en aannemen dat de data stationair is.

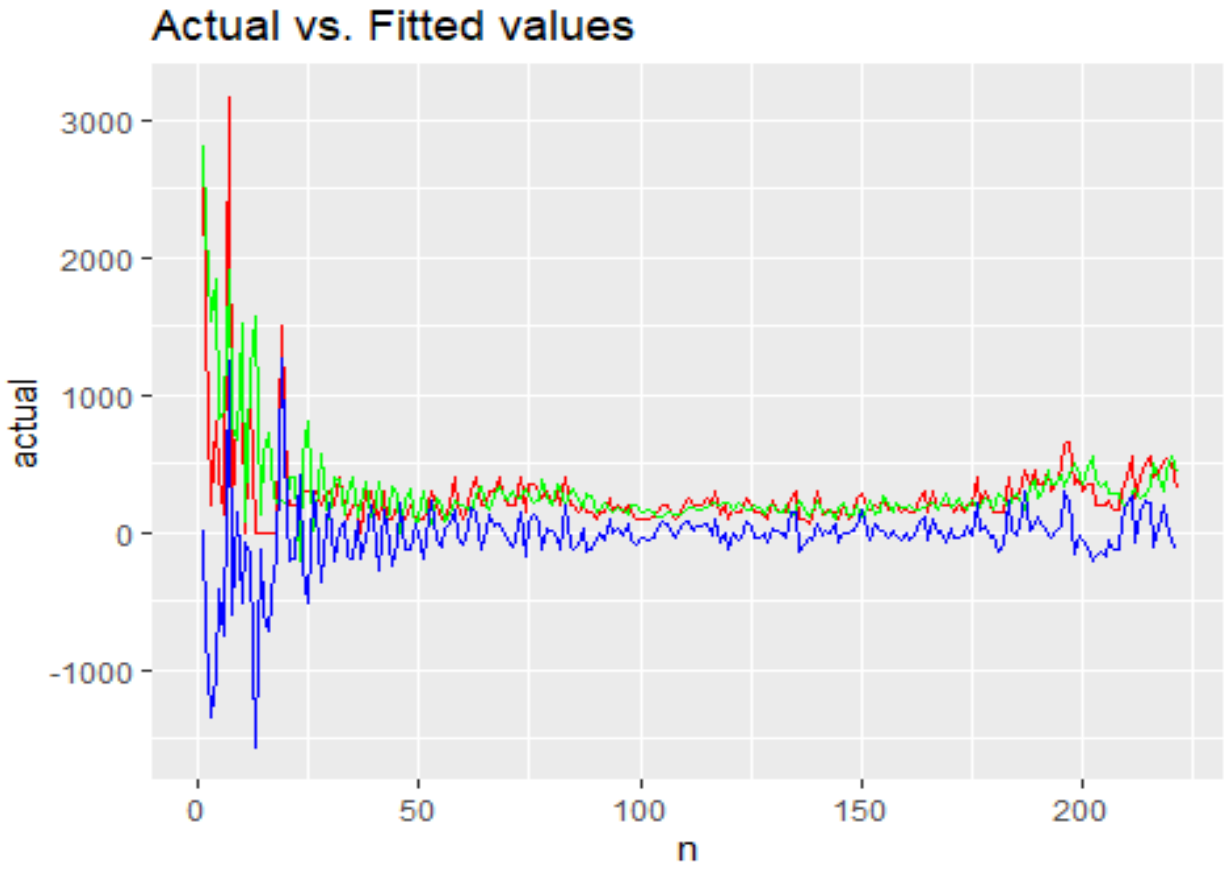
Omdat er geen 'cut off' point zit in beide grafieken kunnen we geen definitieve keuze maken tussen een AR of MA model. Ons eerste model wordt dus een combinatie van deze twee (ARMA of ARIMA model). De parameters voor dit model gaan we schatten door middel van het 'forecast' package. De 'auto.arima' functie in dit package geeft het beste model terug op basis van AIC. Het beste model (laagste AIC) is een **ARIMA(5,1,3)** model. De parameters worden weergegeven in de onderstaande output:

```
## Series: estimation_sample$diff
## ARIMA(5,1,3)
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ma1     ma2      ma3
##       -0.7927  -0.9521  -0.3568  -0.5517  -0.2080  -0.1195  0.3986  -0.4424
## s.e.   0.1651   0.1556   0.2049   0.1425   0.1325   0.1511  0.1005   0.1102
##
## sigma^2 = 67946:  log likelihood = -1533.72
## AIC=3085.44   AICc=3086.29   BIC=3115.98
##
## Training set error measures:
##                      ME     RMSE     MAE  MPE MAPE      MASE      ACF1
## Training set -27.39629 255.3026 136.376 -Inf  Inf 0.9468162 0.1532348
```

Dit ARIMA(5,1,3) model wekt de suggestie dat de 'diff' variabele toch niet stationair is. Ondanks de resultaten van de ADF toets, wordt in het ARIMA model een trend gevonden en hiervoor gecorrigeerd.

Deelvraag 6

In dit onderdeel gaan we een 'deep-dive' doen op dit model, eerst kijken we naar de residuen t.o.v. van de werkelijke waardes en zetten we deze in een grafiek.



Bovenstaand figuur geeft in het rood de werkelijke waardes terug, in het groen de geschatte en in het blauw de residuen. We observeren hogere residuen aan het begin van de oorlog, dit is in lijn met het feit dat aan het begin de personeelsverliezen niet dagelijks werden gerapporteerd.

Laten we hierop wat diagnostische testen doen. We beginnen met de 'Ljung Box' test, omdat we geïnteresseerd zijn in de autocorrelatie tussen elke lag. De null-hypothese van is dat de data onafhankelijk zijn verdeeld.

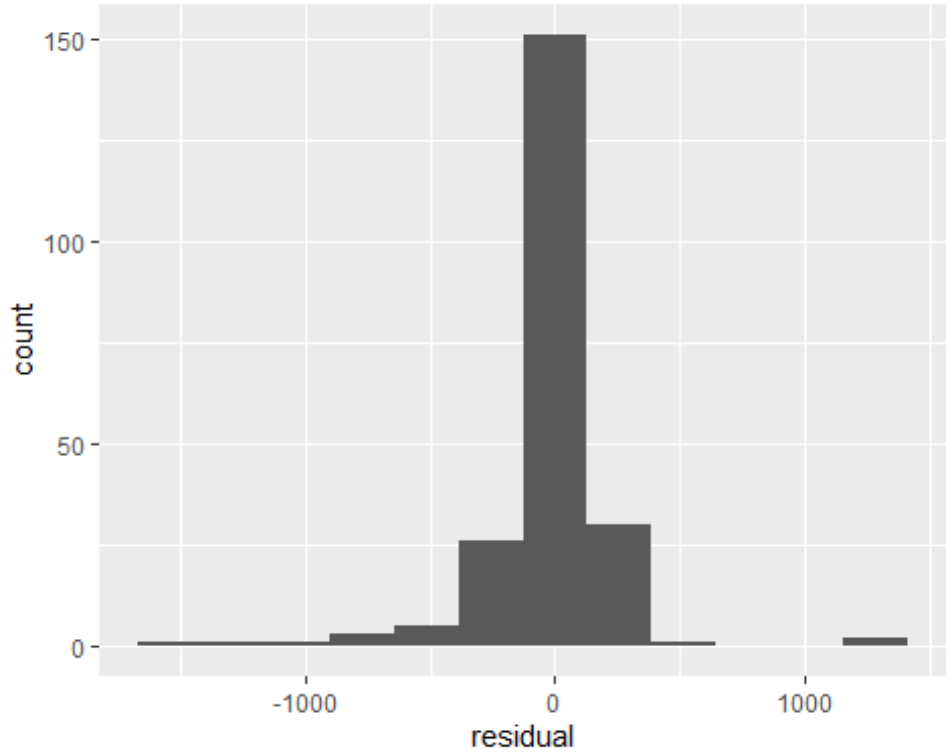
H0: The data are independently distributed

Ha: The data are not independently distributed; they exhibit serial correlation.

```
##
##  Box-Ljung test
##
## data:  actuals_predict$residual
## X-squared = 5.26, df = 8, p-value = 0.7295
```

Omdat we deze toets toepassen op de residuen van een ARIMA model is de hyptohese die eigenlijk toetsen dat de residuen van het ARIMA model geen aurocorrelatie vertonen. Met een p-waarde van 0.7295 kunnen we de null-hypothese niet verwerpen. De residuen vertonen dus geen autocorrelatie.

We kunnen aan de hand van wat getallen en grafieken een kijkje nemen naar de normaliteit van de residuen.



In dit histogram zien we wat uitschieters in de residuen, maar de verdeling ziet er redelijk normaal uit.

Summary statistics:

```
summary(actuals_predict$residual)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1563.987   -86.435    -7.557   -27.396    64.726  1260.667
```

Standaard deviatie:

```
sd(actuals_predict$residual)

## [1] 254.4046
```

We zien een standaard deviatie van ongeveer 254.

Skewness:

```
skewness(actuals_predict$residual)

## [1] -1.113406
```

We zien een negatieve skewness van -1.11, dit is relatief hoog (skewness > 1). De uitschieters in bovenstaande histogram zijn hier waarschijnlijk de oorzaak van.

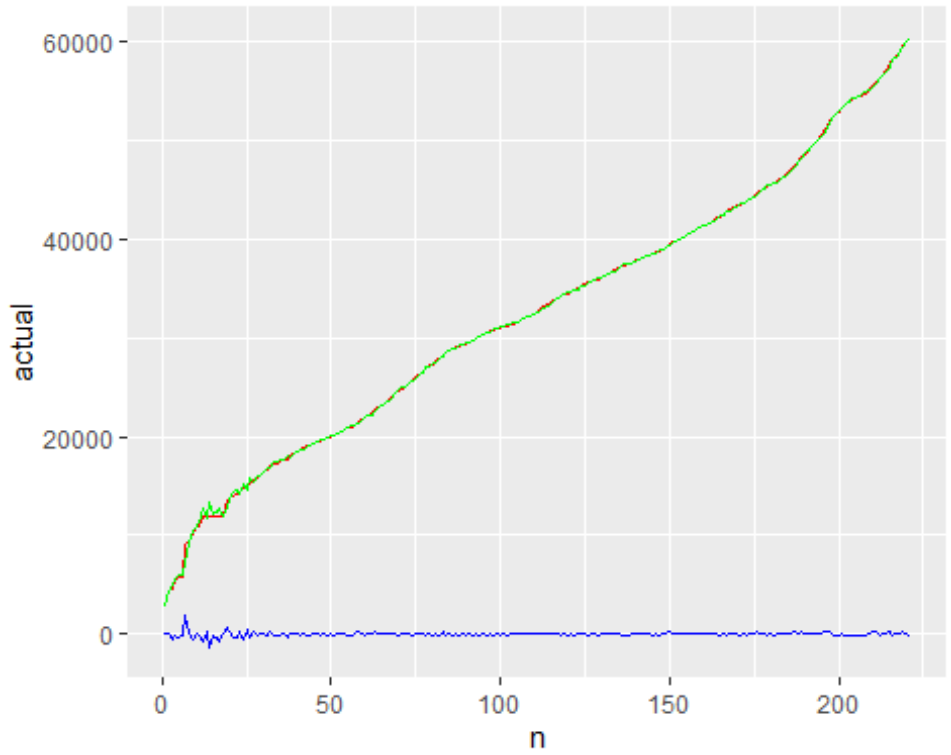
Deelvraag 7

In dit onderdeel gaan we twee alternatieve modellen bekijken. Allereerst beschouwen we een ARIMA model op cumulatieve personeelsverliezen. Wederom wordt de 'auto.arima' functie gebruikt om het optimale model te bepalen. Let op: cumulatieve personeelsverliezen zijn niet stationair, dit hebben we hiervoor getoetst. Om deze reden wordt dit als argument meegegeven. In de onderstaande output staan de parameters die horen bij dit model:

```
stepwise_fit_2 <- auto.arima(estimation_sample$personnel, stationary = FALSE)
summary(stepwise_fit_2)

## Series: estimation_sample$personnel
## ARIMA(2,2,4)
##
## Coefficients:
##           ar1      ar2     ma1      ma2      ma3      ma4
##       -1.6883  -0.9761  0.8713  -0.4918  -0.7473  -0.1329
## s.e.   0.0472   0.0264  0.0955   0.0917   0.0898   0.0901
##
## sigma^2 = 51257:  log likelihood = -1497.62
## AIC=3009.24   AICc=3009.77   BIC=3032.96
##
## Training set error measures:
##                     ME     RMSE     MAE        MPE      MAPE      MASE
## Training set -13.32862 222.2644 122.205 -0.1878078 0.7263966 0.4665121
##                     ACF1
## Training set -0.03305929
```

Het beste model dat wordt teruggegeven door de 'auto.arima' functie is een **ARIMA(2,2,4).** Er is dus twee keer *differencing* toegepast. De AIC van dit model is iets lager dan het vorige model. Hieronder worden de residuen in een tabel gezet en geplot in een figuur.



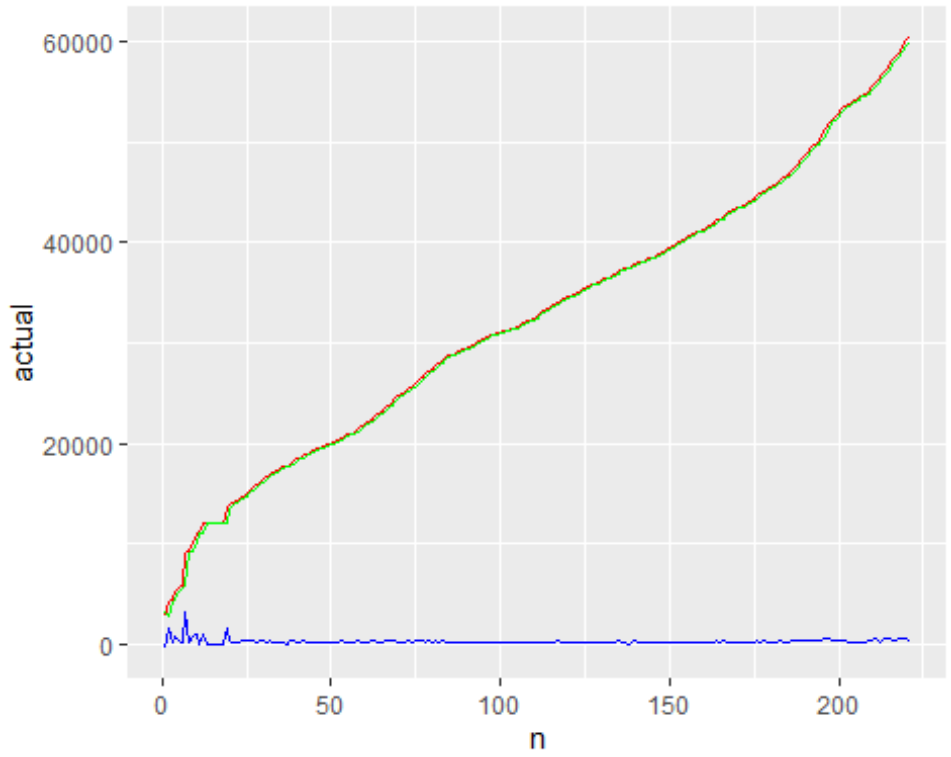
Bovenstaand figuur geeft in het rood de werkelijke waardes terug, in het groen de geschatte en in het blauw de residuen. Wederom zorgt de variantie aan het begin voor hogere residuen.

Laten we ook een simpel autoregressief model beschouwen met 1 lag. In de onderstaande output staan de parameters die horen bij dit model:

```
ar_1 <- arima(estimation_sample$personnel, c(1,0,0))
summary(ar_1)

##
## Call:
## arima(x = estimation_sample$personnel, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.9999   32413.42
## s.e.  0.0001   28169.71
##
## sigma^2 estimated as 143364:  log likelihood = -1629.88,  aic = 3265.76
##
## Training set error measures:
##                    ME     RMSE      MAE      MPE     MAPE     MASE
ACF1
## Training set 258.987 378.6347 262.6989 1.230755 1.359789 1.002842 0.000707
5679
```

Onderstaande grafiek geeft de werkelijke waardes weer in het rood, de geschatte waardes in het groen en de residuen in het blauw.



Laten we de AIC bekijken van de verschillende modellen. Het **AR(1)** model heeft een AIC van 3265, het ARIMA(2,2,4) model een AIC van 3009 en het eerste **ARIMA(5,1,3)** model heeft een AIC van 3085. Dit zit redelijk dicht bij elkaar, maar het **ARMA(2,2,4)** model heeft de laagste AIC en dus de beste fit. We kiezen voor nu dus voor het **ARIMA(2,2,4)** model. De grafiek van het **ARIMA(2,2,4)** model op cumulatieve personeelsverliezen vertoont wel symptomen van *overfitting*, de geschatte waarden zitten heel dicht bij de werkelijke waarden. In het volgende onderdeel gaan we dit toetsen op de test data door te kijken naar de *accuracy measures*.

Deelvraag 8

Hieronder de voorspellingen van de voorgaande modellen op de testdata.

ARIMA(3,0,3):

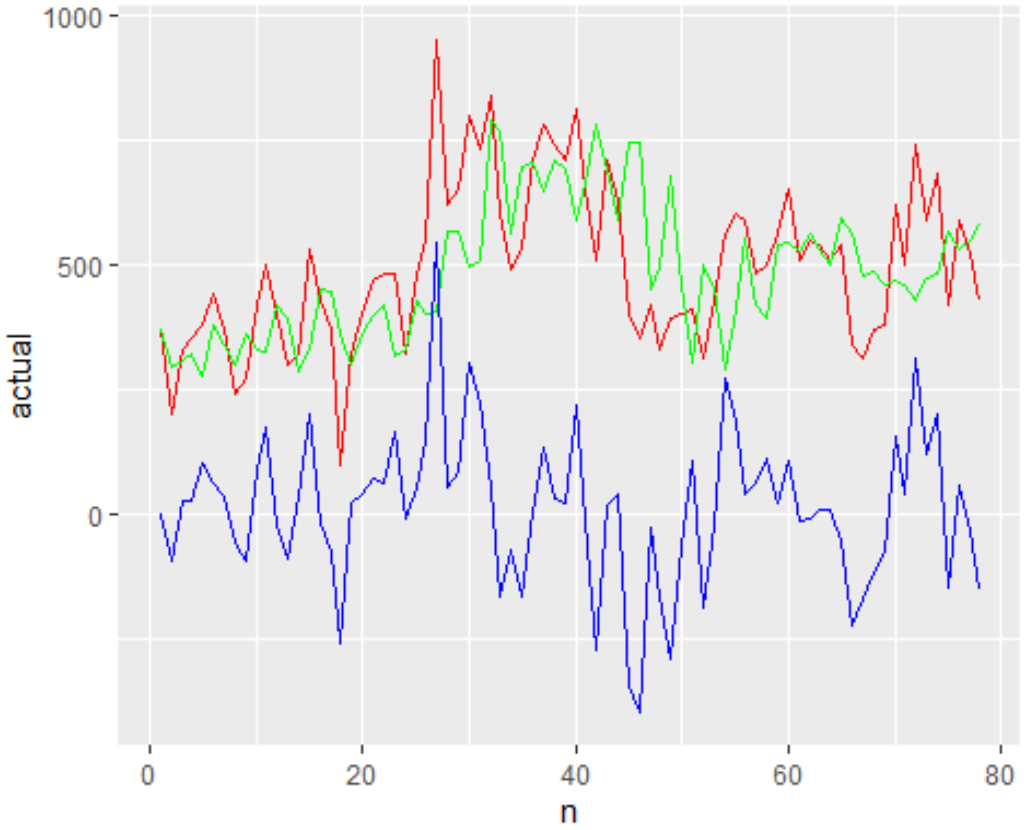
```
## Series: test_sample$diff
## ARIMA(5,1,3)
##
## Coefficients:
##           ar1      ar2      ar3      ar4     ar5      ma1     ma2      ma3
##       -0.7927  -0.9521  -0.3568  -0.5517  -0.208  -0.1195  0.3986  -0.4424
## s.e.   0.0000   0.0000   0.0000   0.0000   0.000   0.0000  0.0000   0.0000
```

```
##
## sigma^2 = 65476:  log likelihood = -498.63
## AIC=999.26   AICc=999.31   BIC=1001.6

accuracy(arima_test_1)

##                     ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
## Training set 12.20749 152.6345 112.037 -5.429648 26.13473 1.057212 0.3536498
```

Onderstaande grafiek geeft de werkelijke waardes weer in het rood, de geschatte waardes in het groen en de residuen in het blauw.



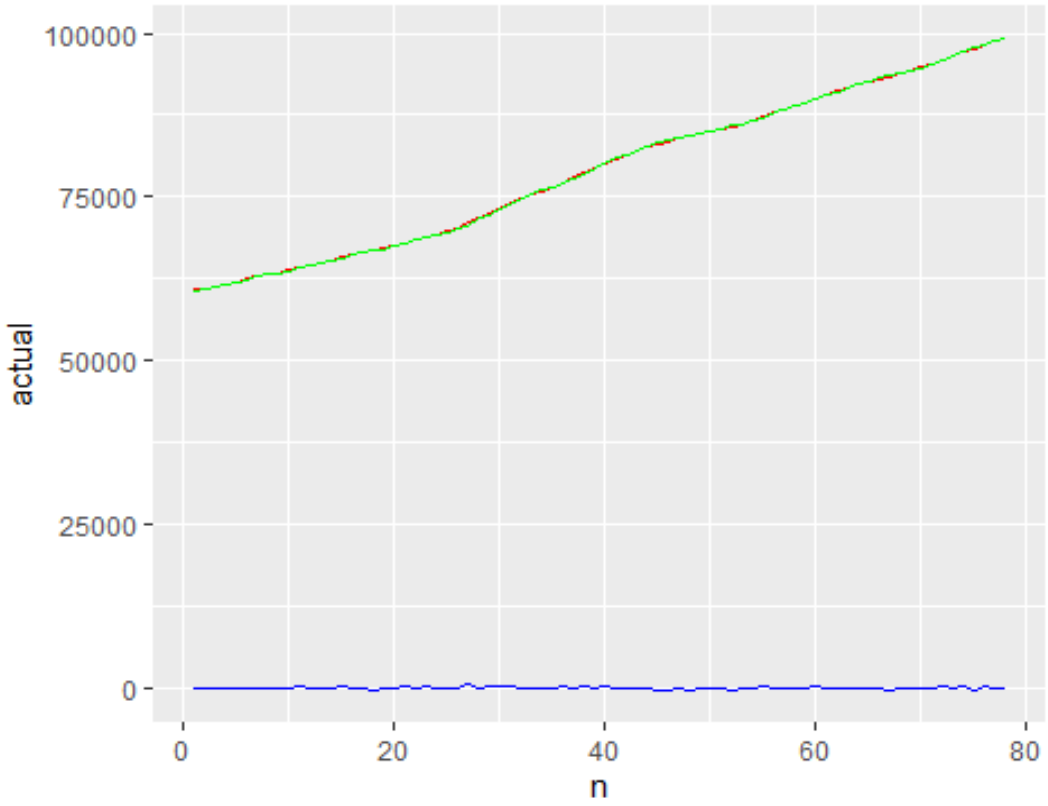
Dit model heeft een AIC van 999 en de *accuracy metrieken* zijn als volgt:

```
##                     ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
## Training set 12.20749 152.6345 112.037 -5.429648 26.13473 1.057212 0.3536498
```

ARIMA(2,2,4):

```
## Series: test_sample$personnel
## ARIMA(2,2,4)
##
## Coefficients:
##           ar1      ar2     ma1      ma2      ma3      ma4
##       -1.6883  -0.9761  0.8713  -0.4918  -0.7473  -0.1329
## s.e.   0.0000   0.0000  0.0000   0.0000   0.0000   0.0000
##
## sigma^2 = 49853:  log likelihood = -489.87
## AIC=981.74   AICc=981.79   BIC=984.07
```

Onderstaande grafiek geeft de werkelijke waardes weer in het rood, de geschatte waardes in het groen en de residuen in het blauw.



Dit model heeft een AIC van 982 en de *accuracy metrieken* zijn als volgt:

```
##                     ME     RMSE      MAE        MPE      MAPE      MASE      ACF1
## Training set 20.41312 146.055 114.243 0.03066989 0.1468885 0.2289022 0.2645644
```

AR(1):

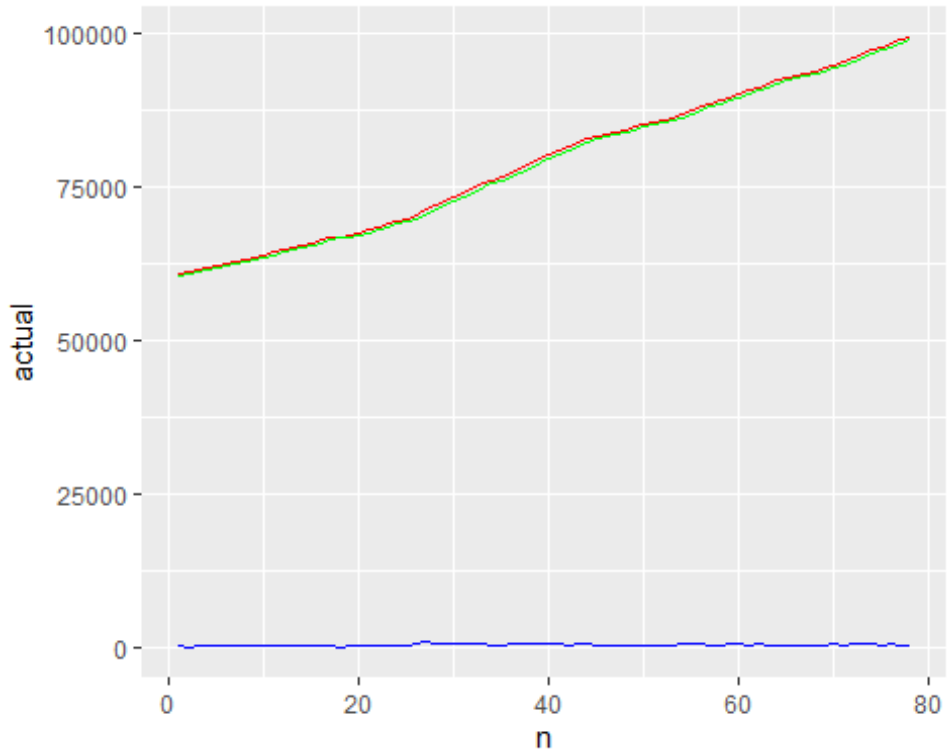
```
## Series: test_sample$personnel
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1      mean
##       0.9999  32413.42
## s.e.  0.0000      0.00
##
## sigma^2 = 143364:  log likelihood = -603.77
## AIC=1209.55   AICc=1209.6   BIC=1211.9

## Series: test_sample$personnel
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1      mean
##       0.9996  79753.06
## s.e.  0.0005  18912.00
##
## sigma^2 = 282002:  log likelihood = -602.69
## AIC=1211.39   AICc=1211.71   BIC=1218.46
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE      MAPE      MASE      ACF1
## Training set 485.6596 524.1858 499.0343 0.6135213 0.6355193 0.9998867 0.5277807
```

Dit model heeft een AIC van 1210 en de *accuracy metrieken* zijn als volgt:

```
##                     ME     RMSE      MAE       MPE      MAPE     MASE      ACF1
## Training set 501.6553 526.5952 501.6553 0.6388958 0.6388958 1.005138 0.6526577
```

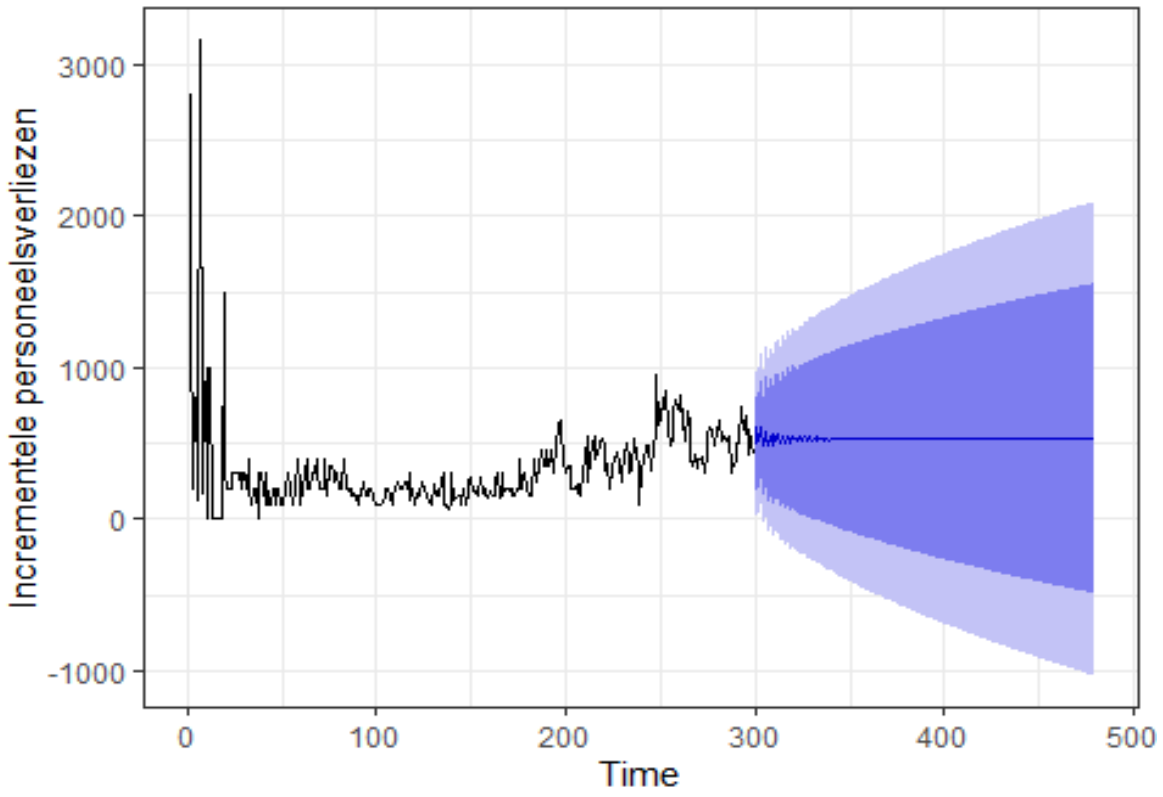
Onderstaande grafiek geeft de werkelijke waardes weer in het rood, de geschatte waardes in het groen en de residuen in het blauw.



Modelkeuze: Het **ARIMA(2,2,4)** en het **ARIMA(5,1,3)** model hebben de laagste AIC, respectievelijk 982 en 999. Ook op de *accuracy metrieken* die we hebben berekend scoort het **ARIMA(2,2,4)** model op cumulatieve personeelsverliezen iets beter. Dit model heeft voor nu dus de voorkeur. Echter zitten deze metrieken dicht bij elkaar, om die reden zullen we in de volgende vraag beide modellen beschouwen.

Deelvraag 9

Het **ARIMA(5,1,3)** model wordt nu opnieuw geschat op de gehele dataset en er wordt een voorspelling gedaan op de opvolgende 180 dagen.



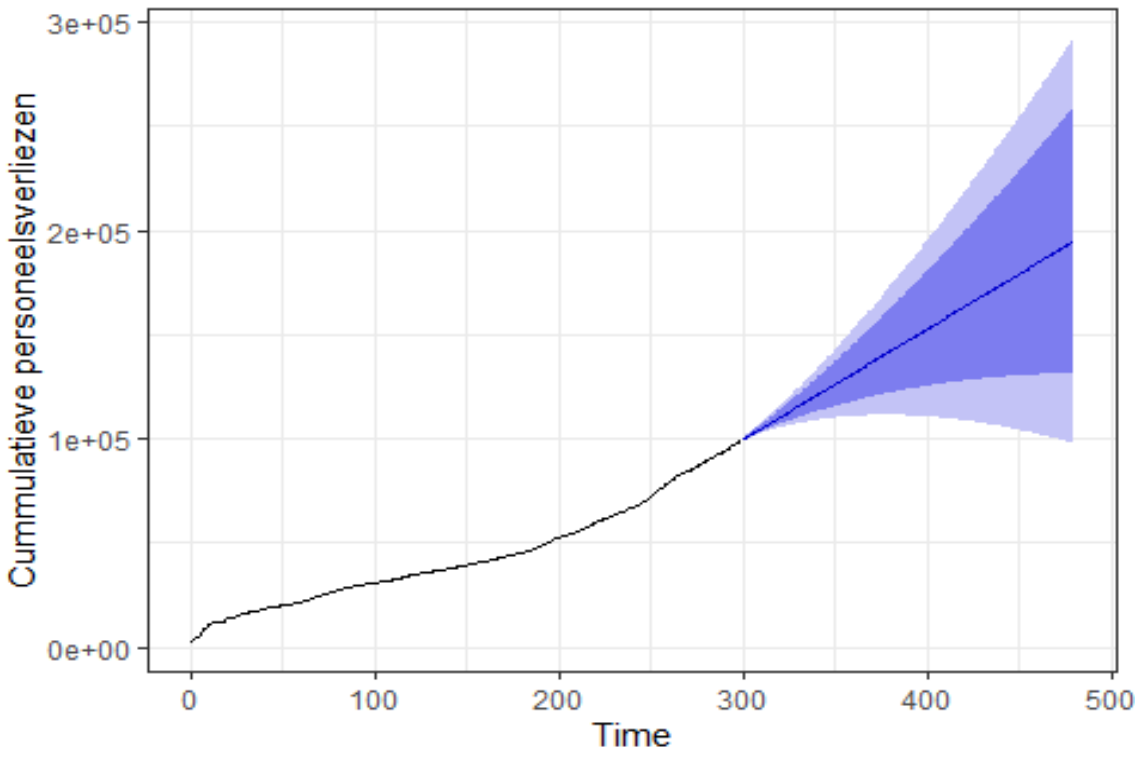
Bovenstaande grafiek toont de voorspellingen voor de opvolgende 180 dagen, met in het donkerblauw de 95% betrouwbaarheids intervallen en in het lichtblauw de 80% betrouwbaarheidsintervallen.

Hieronder de beschrijvende statistieken van het **ARIMA(5,1,3)** model:

```
## Series: russia_losses_personnel$diff
## ARIMA(5,1,3)
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ma1     ma2      ma
3
##       -0.7679  -0.9064  -0.3025  -0.5045  -0.1856  -0.0959  0.3820  -0.417
9
## s.e.   0.1701   0.1503   0.1963   0.1316   0.1236   0.1589  0.0953   0.109
4
##
## sigma^2 = 56017:  log likelihood = -2049.33
## AIC=4116.66   AICc=4117.28   BIC=4149.93
##
```

```
## Training set error measures:
##                      ME     RMSE     MAE  MPE MAPE      MASE      ACF1
## Training set -18.46745 233.0902 127.339 -Inf  Inf 0.9511007 0.1445132
```

Het **ARIMA(2,2,4)** model wordt nu opnieuw geschat op de gehele dataset en er wordt een voorspelling gedaan op de opvolgende 180 dagen.



Bovenstaande grafiek toont de voorspellingen voor de opvolgende 180 dagen, met in het donkerblauw de 95% betrouwbaarheidsintervallen en in het lichtblauw de 80% betrouwbaarheidsintervallen.

Hieronder de beschrijvende statistieken van het **ARIMA(5,1,3)** model:

```
## Series: russia_losses_personnel$personnel
## ARIMA(2,2,4)
##
## Coefficients:
##           ar1      ar2     ma1      ma2      ma3      ma4
##       -1.6940  -0.9789  0.9056  -0.4380  -0.7363  -0.1352
## s.e.   0.0381   0.0227  0.0749   0.0793   0.0790   0.0762
##
## sigma^2 = 43243:  log likelihood = -2005.94
## AIC=4025.87   AICc=4026.26   BIC=4051.73
##
```

```
## Training set error measures:
##                      ME    RMSE      MAE         MPE      MAPE      MASE
## Training set -6.325565 205.149 119.8513 -0.1244479 0.5793198 0.3703796
##                      ACF1
## Training set -0.02029846
```