Time Series Analysis: Project 1

Due on Aprl 11, 2021

Professor Tianwei Yu

Peng Deng

Contents

1 Problem statement

Use dataset "chicagoNMMAPS" from the "dlnm" package in R. The data set contains daily mortality, weather, and pollution data for Chicago from 1987 to 2000. We would like to use the data to do some training and forecast by implementing some timeseries models.

2 Dataset overview

Firstly, we can have a quick look at the dataset, the meanings of the headers in the dataset are explained as follow:

- date: Date in the period 1987-2000
- time: The sequence of observations
- year: Year
- month: Month (numeric)
- doy: Day of the year
- dow: Day of the week (factor)
- death: Counts of all cause mortality excluding accident
- cvd: Cardiovascular Deaths
- resp: Respiratory Deaths
- temp: Mean temperature (in Celsius degrees)
- dptp: Dew point temperature
- rhum: Mean relative humidity
- pm10: PM10
- o3: Ozone

3 Data preprocessing

First, combine the daily data into monthly data. For data with missing values, such as daily PM10, replace the NA with average over non-NA values over the month before taking monthly average. We can see that this preprocessing equals to just take average on non-NA values. So our preprocessing is just to take average on non-NA values in every month and separate the data into three segments:

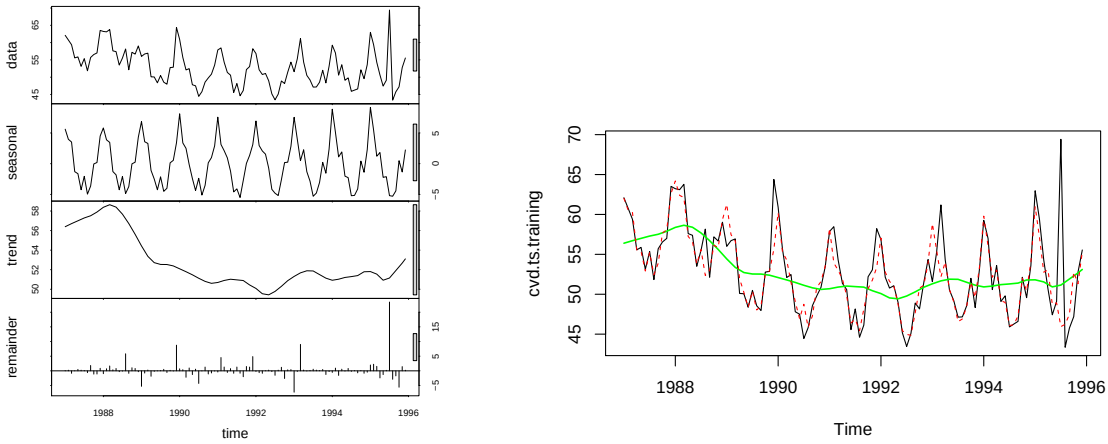
- Training: 1987.1-1995.12
- Validation: 1996.1-1997.12
- Testing: 1998.1-2000.12

4 Univariate analysis of the cardiovascular death(cvd)

4.1 STL and EST model

We know that stl (Seasonal and Trend decomposition using Loess) is a very useful tool to decompose data. Firstly, we just use the following parameters of stl to do a decomposition of the training dataset.

```
cvd.stl = stl(cvd.ts.training, s.window=4, t.window=12, robust=TRUE)
```



(a) The decomposition result by stl (b) Holt-Winters filtering

Figure 1 stl decomposition and fitted figure

Then, we can have a quik look of the decomposition result as Figure 1. In Figure 1(b), the black line is the original training data, the green line is the trend, and the red dashed line is the fitted value.
In order to see if the decomposition is good, we can plot the acf of the decomposition remainder as Figure 2. As we can see in Figure 2, there is a little acf, so that it is not suitable to do forecast just use stl. Thus, we implement ETS as an complement to do forecast.

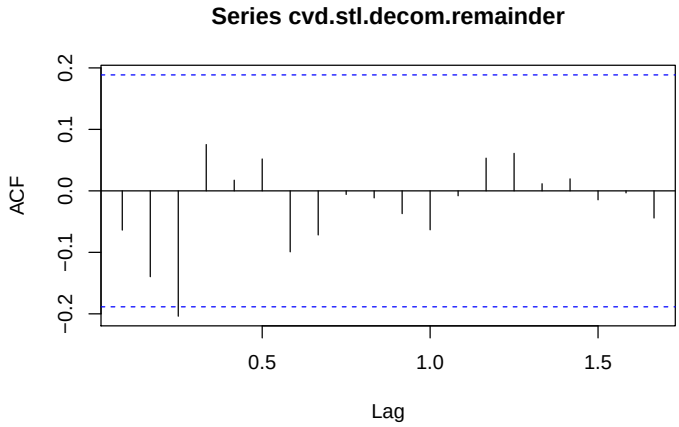


Figure 2 The ACF of residuals

The forecast result is showed as Figure 3. As we can see in Figure 3, the dashed line is the true value of validation dataset, the blue line is the forcasted value of Validation dataset. Then, we calculated the Root Mean Squard Error (RMSE) on the validation dataset.
$\star$ RMSE $= 5.61$ $\star$

4.2 Holt-Winters model

As we know that there are holt-winters models with additive seasonal component and with multiplicative seasonal component. As we can see, there is no evident multiplicative seasonal component in the time series, so we just use the holt-winters model with additive seasonal component. The fitted figures are as Figure 4.

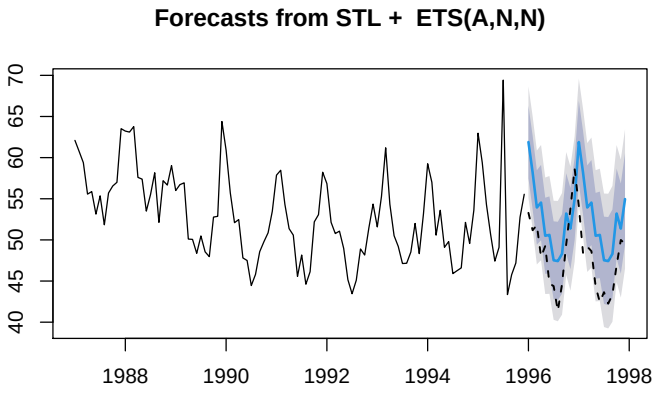
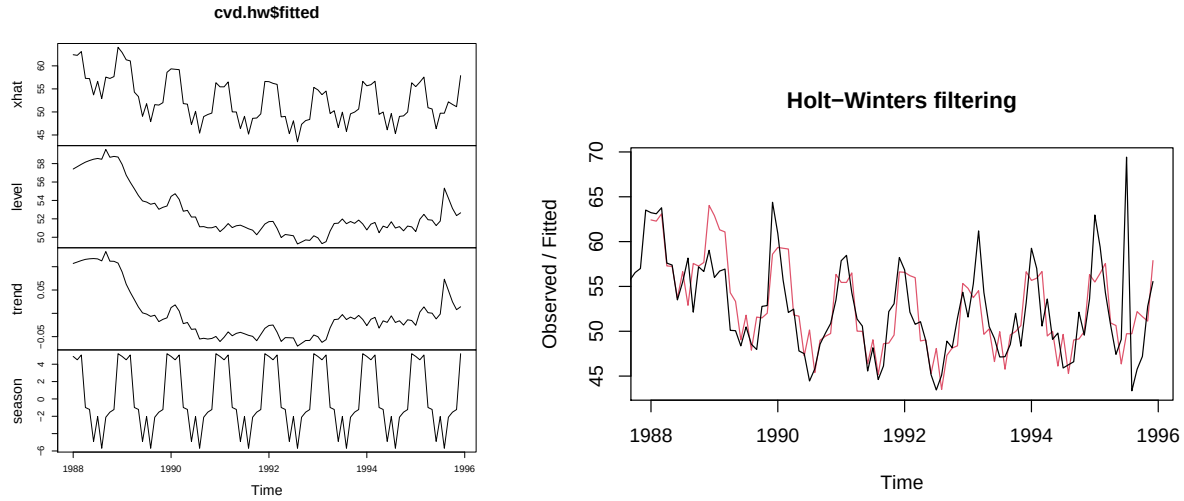


Figure 3 The prediction on validation dataset with STL and ETS



(a) Holt-Winters decomposition

(b) Holt-Winters filtering

Figure 4 Holt-winters model fitted figure

Then, we plot the ACF figure of the residuals according to the model result as Figure 5. As we can see, There is no significant auto-correlation in the residuals, which means our model is somehow good to fit the data.

Then, we would like to implement the fitted model on the validation dataset(1996.1-1997.12) in order to test the model's prediction accuracy. As we can see in Figure 6, the dashed line is the true value of validation dataset, the blue line is the forcasted value of validation dataset. Then, we calculated the Root Mean Squard Error (RMSE) on the validation dataset.

$\star$ RMSE = 5.08 $\star$

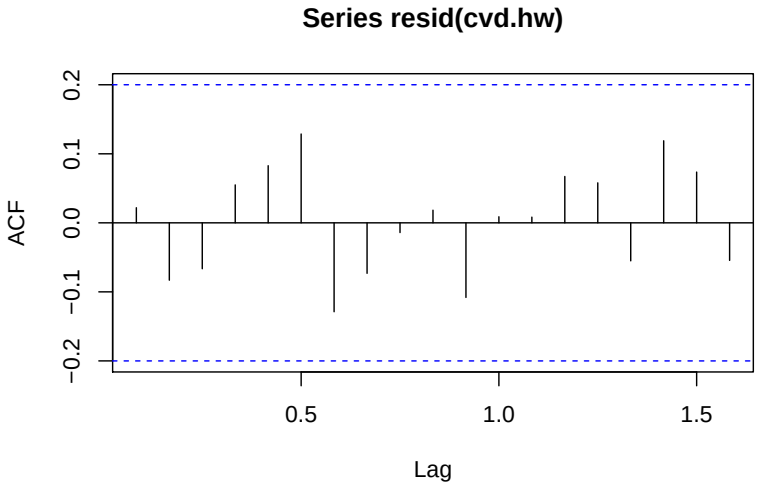


Figure 5 The ACF of residuals

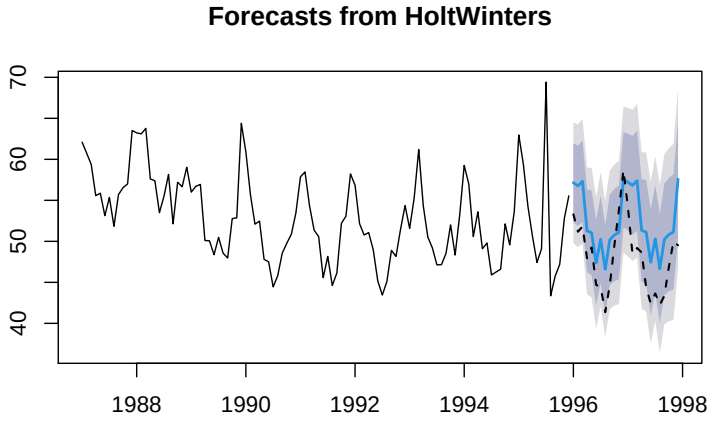


Figure 6 The prediction on validation dataset with holt-winters model

4.3 Seasonal ARIMA model

As we can see, there is evident seasonal component, so that we would like to use Seasonal ARIMA instead of just ARIMA. In order to determine the parameters in seasonal ARIMA, we call "auto.arima" in R as follow.

```
cvd.arima = auto.arima(cvd.ts.training)
```

Then, we get the parameters of Seasonal ARIMA is: ARIMA(1,1,1)(0,0,2)[12]. The fitted figures are as Figure 7.

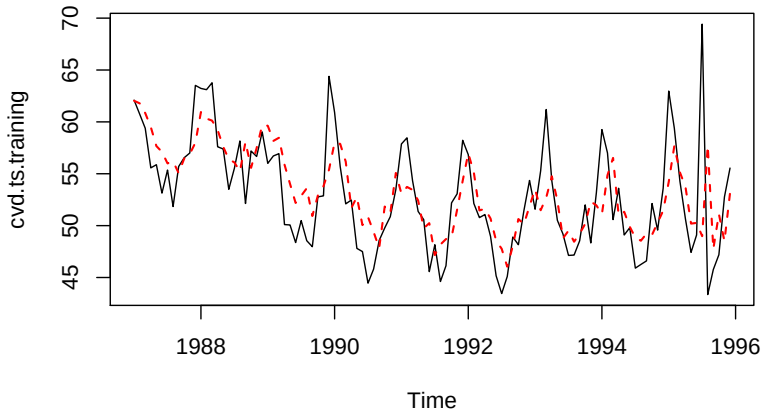


Figure 7 Seasonal ARIMA model fitted figure

Then, we plot the residuals and the ACF figure of the residuals according to the Seasonal ARIMA model result as Figure 8, which means we are going to check residuals.

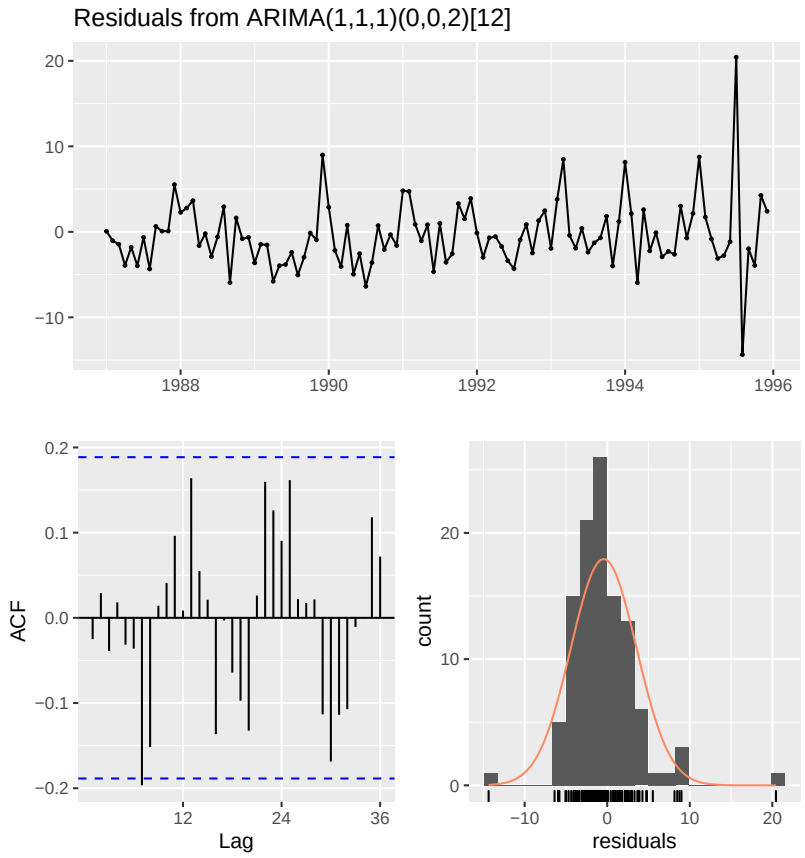


Figure 8 Check residuals of Seasonal ARIMA

As we can see from Figure 8, there is significant acf, so that the model is not very good. However, we can do the Ljung-Box test of the residuals to see if the residuals exhibit serial correlation, the result is as follow:

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)(0,0,2)[12]
Q* = 23.289, df = 18, p-value = 0.1797

Model df: 4.   Total lags used: 22
```

As we can see from the Ljung-Box test, the p-value equals 0.1797, which is larger than 0.05, so that we should not reject the null hypothesis, which means the residuals does not exhibit serial correlation. Thus, we can believe that the Seasonal ARIMA model is suitable.
Then, we would like to implement the fitted Seasonal ARIMA model on the validation dataset(1996.1-1997.12) in order to test the model's prediction accuracy. As we can see in Figure 9, the dashed line is the true value of validation dataset, the blue line is the forcasted value of Validation dataset. Then, we calculated the Root Mean Squard Error (RMSE) on the validation dataset.
$\star$ RMSE = 6.35 $\star$

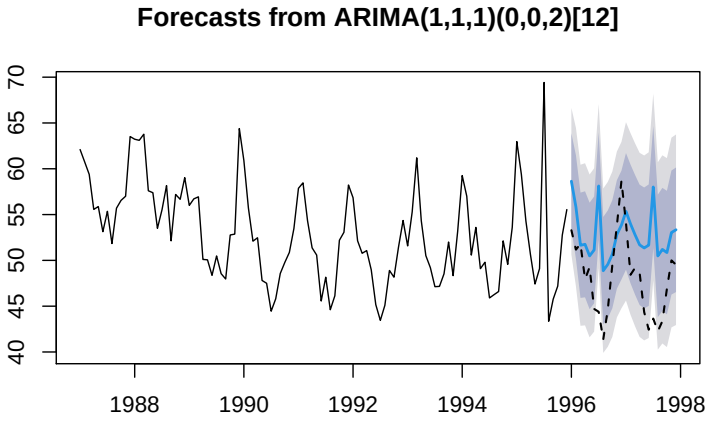
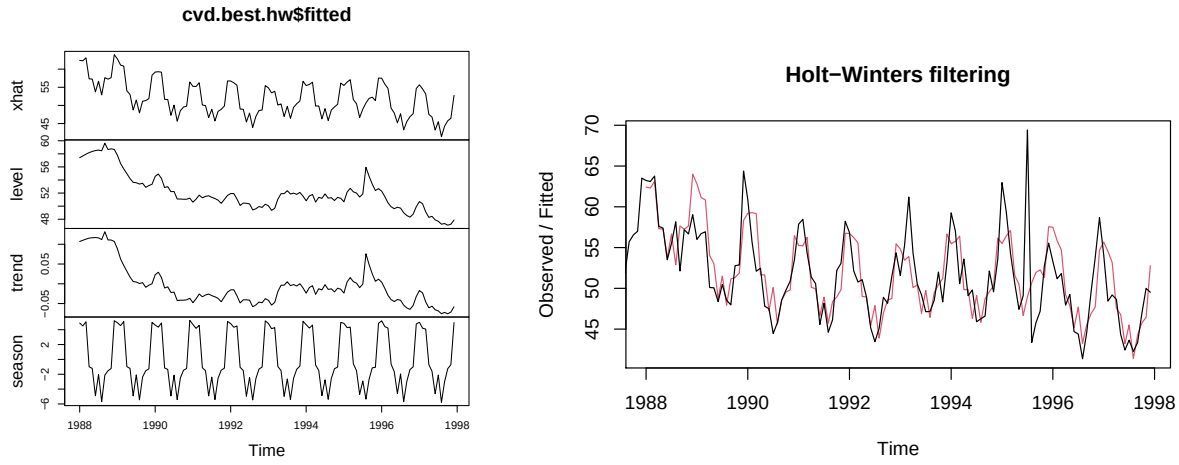


Figure 9 The prediction on validation dataset with Seasonal ARIMA model

4.4 Run the best model

According to the RMSE value of the three models as follow, we can find that the **holt-winters model** is the bset. Thus, we would like to run holt-winters in the following.
- STL+ETS: RMSE = 5.61
- Holt-winters: RMSE = 5.08 $\star$
- Seasonal ARIMA: RMSE = 6.35

Firstly, we would like to run the best-performing model (holt-winters) on the 1987.1-1997.12 data (training dataset and validation dataset). We just use the holt-winters model with additive seasonal component. The fitted figures are as Figure 10.



(a) Holt-Winters decomposition

(b) Holt-Winters filtering

Figure 10 Best model fitted figure

Then, we plot the ACF figure of the residuals according to the model result as Figure 11.

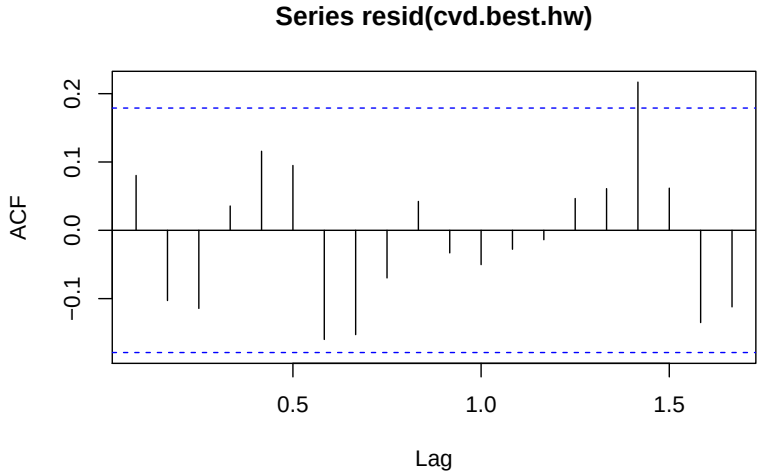


Figure 11 The ACF of residuals with the best model

As we can see from Figure 11, there is evident acf. Thus, we have to do Ljung-Box test of the residuals to see if the residuals exhibit serial correlation, the result is as follow:

```
        Box-Ljung test

data:  resid(cvd.best.hw)
X-squared = 0.79242, df = 1, p-value = 0.3734
```

As we can see from the Ljung-Box test, the p-value equals 0.3734, which is larger than 0.05, so that we should not reject the null hypothesis, which means the residuals does not exhibit serial correlation. Thus, we can believe that the best model (holt-winters) is suitable.

Then, we would like to implement the fitted model on the test dataset(1998.1-2000.12) in order to do forecast.

As we can see in Figure 12, the dashed line is the true value of test dataset, the blue line is the forcasted value of test dataset.

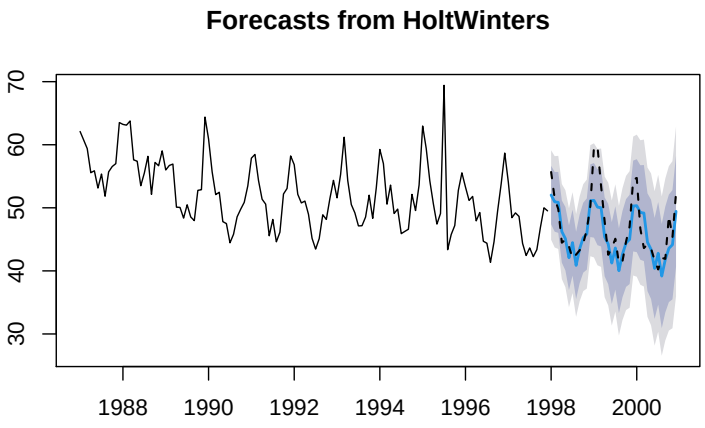


Figure 12 The prediction on test dataset with best model (holt-winters)

Then, we calculated the Root Mean Squard Error (RMSE) on the test dataset.
$\star$ RMSE = 3.16 $\star$
We can see that the forecast is pretty good, it is very near to the true value. Besides, the RSME equals 3.16, which is relatively small.

Above all, we can conclude that the holt-winters model performs very good.

5 Multivariate analysis of the cardiovascular death(cvd)

Now consider also the temperature (temp) and the pollution variables (PM10 and o3), which could help predict mortality of some diseases. Use the cardiovascular death as outcome as well.

5.1 ARIMA with external variables

Firsly, lets have a quik look of the value of the four variables (cvd, temp, pm10 and o3). their value on training data set is showed as Figure 13.

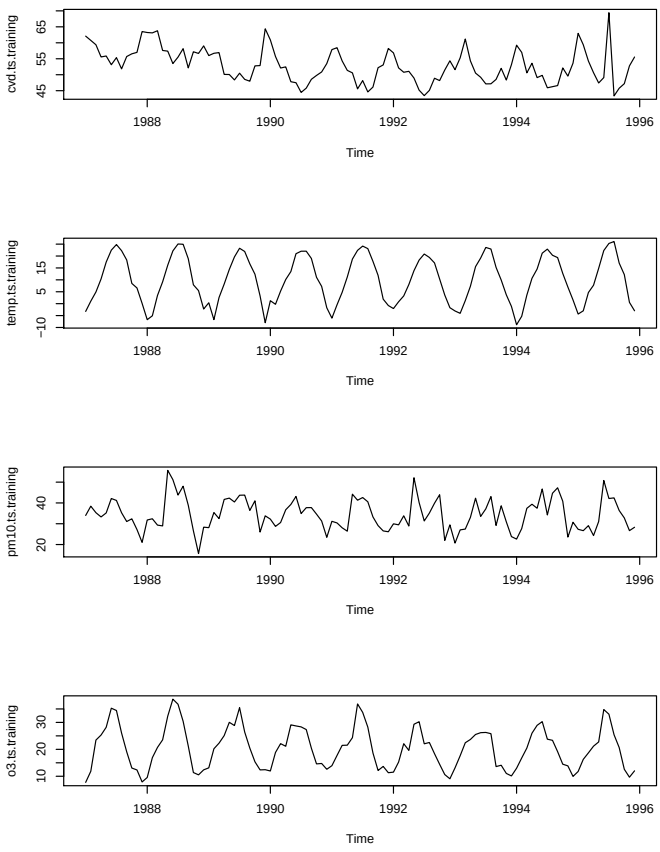


Figure 13 The value of four variables on training dataset

As we can see from Figure 13, there may have some relationships between cvd and external variables. Then, we implement auto.arima to choose the parameters of "ARIMA with external variables" model.

```
multi.arima.external = auto.arima(cvd.ts.training,
            xreg=cbind(temp.ts.training, pm10.ts.training, o3.ts.training))
```

Then, we can see the result as follow

```
Series: cvd.ts.training
Regression with ARIMA(0,1,1) errors

Coefficients:
         ma1  temp.ts.training  pm10.ts.training  o3.ts.training
      -0.8198           -0.4280            0.0409          0.1229
s.e.   0.0602            0.0533            0.0634          0.0709

sigma^2 estimated as 12.57:  log likelihood=-285.76
AIC=581.52   AICc=582.11   BIC=594.88
```

Thus, we get the parameters of ARIMA with external variables is: the error η_t is ARIMA(0,1,1). The fitted figure is as Figure 14.

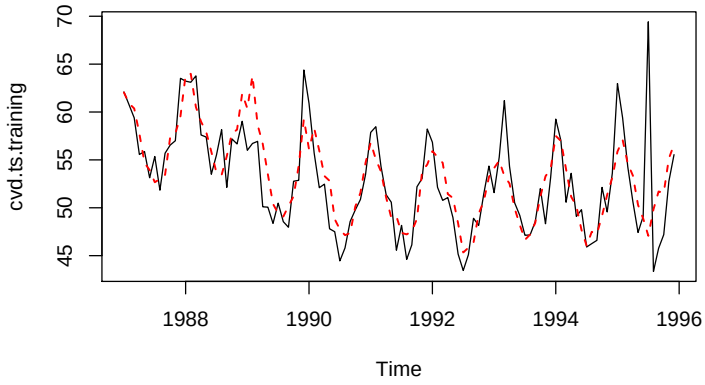


Figure 14 ARIMA with external variables model fitted figure

Then, we plot the residuals and the ACF figure of the residuals according to the the model result as Figure 15, which means we are going to check residuals.

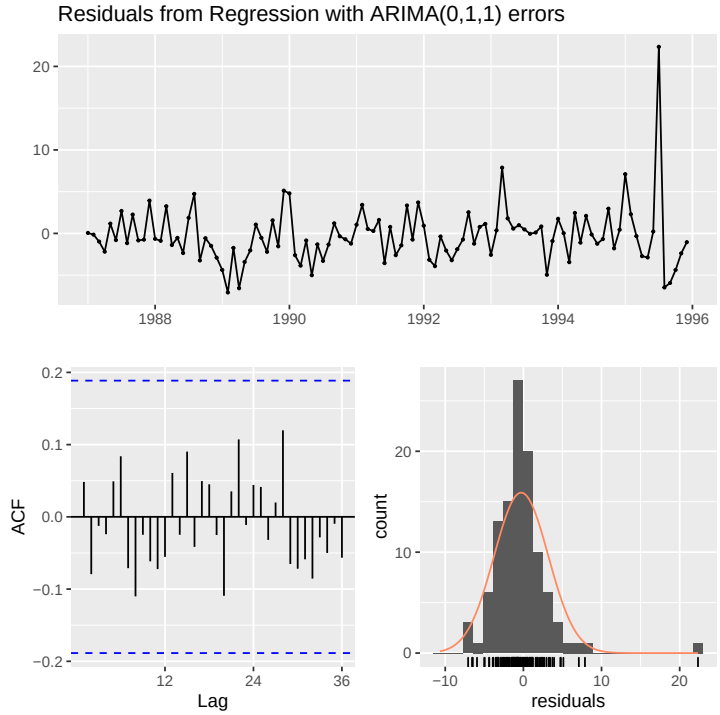


Figure 15 Check residuals of ARIMA with external variables

As we can see from Figure 15, there is no significant acf, so that the model is somehow very good.

Then, we would like to implement the fitted ARIMA with external variables model on the validation dataset(1996.1-1997.12) in order to test the model's prediction accuracy. As we can see in Figure 16, the dashed line is the true value of validation dataset, the blue line is the forcasted value of validation dataset. Then, we calculated the Root Mean Squard Error (RMSE) on the validation dataset.
$\star$ RMSE = 4.83 $\star$

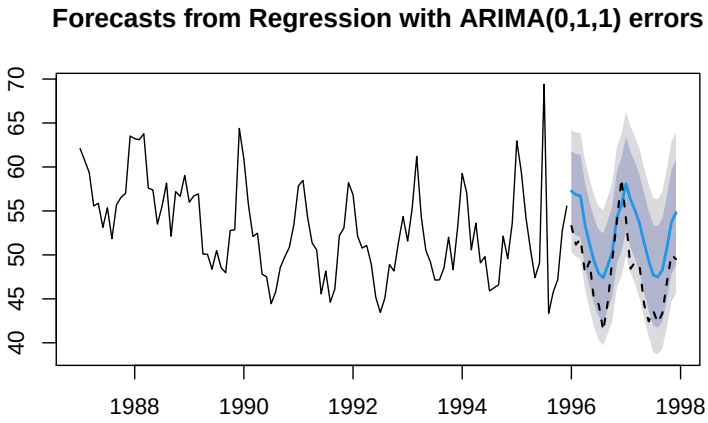


Figure 16 The prediction on validation dataset with ARIMA with external variables model

5.2 Vector AR model

In this section, we would like to implement Vector AR model. In order to avoid too many parameters (overfitting), we would like to control the lag.max=2. We run the following code to select the suitable lag order:

```
VARselect(multi.training, lag.max=2, type="const")
```

The result is as follow:

```
selection
AIC(n)  HQ(n)  SC(n) FPE(n)
      2       2       1       2

$criteria
                1           2
AIC(n)    10.83243    10.44461
HQ(n)     11.03611    10.81123
SC(n)     11.33496    11.34917
FPE(n) 50650.55661 34415.04660
```

Thus, we would like to use lag oder=2 in this vector AR model. The fitted figure is as Figure 14. As we can see, the fitting seems very good, because there are $3 + 2 \times 3^2 = 21$ parameters in this model, which means it is very possible to overfit.

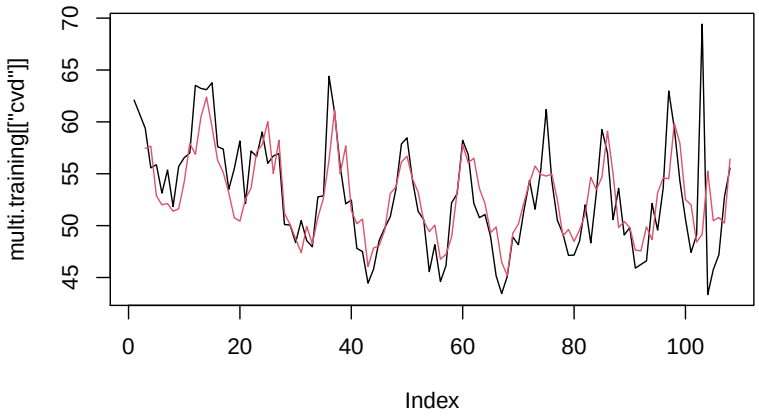


Figure 17 Vector AR model fitted figure

Then, we plot the residuals and the ACF figure of the residuals according to the the model result as Figure 15, which means we are going to check residuals of cvd.

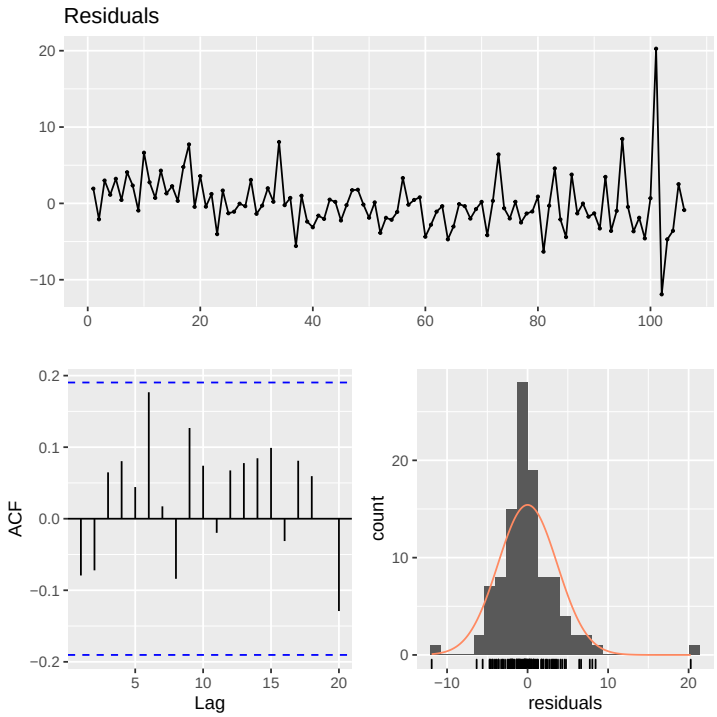


Figure 18 Check residuals of vector AR model

As we can see from Figure 18, there is no significant acf, so that the model is somehow very good.

Then, we would like to implement the fitted vector AR model on the validation dataset(1996.1-1997.12) in order to test the model's prediction accuracy. Firstly, we would like to see the fitted parameters as follow:

```
VAR Estimation Results:
=======================
```

```
Estimated coefficients for equation cvd:
=======================================
Call:
cvd = cvd.l1 + temp.l1 + pm10.l1 + o3.l1 + cvd.l2 + temp.l2 +
        pm10.l2 + o3.l2 + const

    cvd.l1      temp.l1      pm10.l1        o3.l1      cvd.l2      temp.l2
    0.36411984 -0.29789871 -0.04243912  0.13608390  0.20698752  0.38811791
    pm10.l2        o3.l2        const
    -0.01432600 -0.28986066 26.80201828
```

Then, we can do forecast. As we can see in Figure 19, the dashed line is the true value of validation dataset, the blue line is the forcasted value of validation dataset. Then, we calculated the Root Mean Squard Error (RMSE) on the validation dataset. We can see that the RMSE is little larger, so that there might be overfitting in this model.
⋆ RMSE = 5.67 ⋆

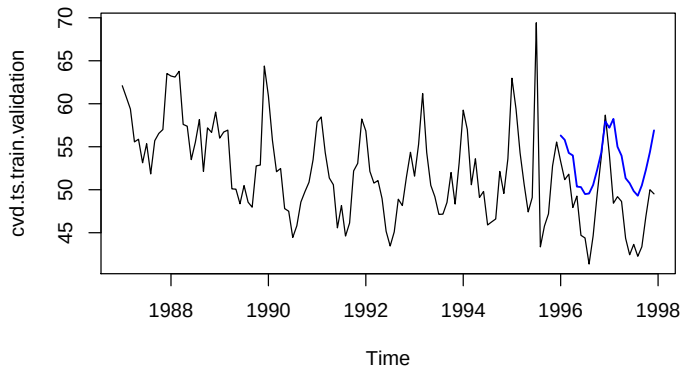


Figure 19 Check residuals of vector AR model

5.3 Run the best model

According to the RMSE value of the three models as follow (including the best univariate model-holt winters), we can find that the **ARIMA with external variables** is the bset. Thus, we would like to run ARIMA with external variables in the following.

- Holt-winters: RMSE = 5.08
- ARIMA with external variables: RMSE = 3.16 ⋆
- Vector AR: RMSE = 5.67

Firstly, we would like to run the best-performing model (ARIMA with external variables) on the 1987.1-1997.12 data (training dataset and validation dataset). We implement auto.arima to choose the parameters of "ARIMA with external variables" model.

```
arima.external.best = auto.arima(cvd.ts.train.validation,
        xreg=cbind(temp.ts.train.val, pm10.ts.train.val, o3.ts.train.val))
```

Then, we can see the result as follow

```
Series: cvd.ts.train.validation
Regression with ARIMA(0,1,2) errors
```

```
Coefficients:
         ma1      ma2   drift  temp.ts.train.val  pm10.ts.train.val  o3.ts.train.val
      -0.7093  -0.1441  -0.0841            -0.3952             0.0411           0.0635
s.e.  0.0884   0.0914   0.0447             0.0498             0.0563           0.0663

sigma^2 estimated as 11.38:  log likelihood=-342.71
AIC=699.43   AICc=700.34   BIC=719.55
```

Thus, we get the parameters of ARIMA with external variables is: the error η_t is ARIMA(0,1,2). The fitted figure is as Figure 20.

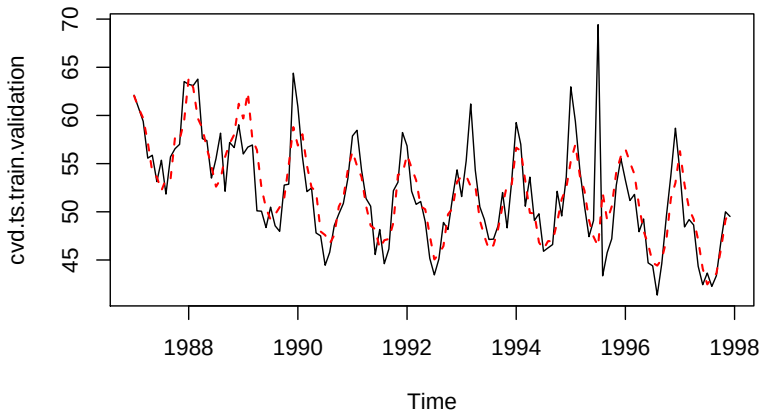


Figure 20 AIRMA with external variables model fitted figure

Then, we plot the residuals and the ACF figure of the residuals according to the the model result as Figure 21, which means we are going to check residuals.

Figure 21 Check residuals of ARIMA with external variables

As we can see from Figure 21, there is evident acf. Thus, we have to do Ljung-Box test of the residuals to

see if the residuals exhibit serial correlation, the result is as follow:

```
Ljung-Box test

data:  Residuals from Regression with ARIMA(0,1,2) errors
Q* = 18.836, df = 18, p-value = 0.402

Model df: 6.   Total lags used: 24
```

As we can see from the Ljung-Box test, the p-value equals 0.402, which is larger than 0.05, so that we should not reject the null hypothesis, which means the residuals does not exhibit serial correlation. Thus, we can believe that the ARIMA with external variables model is suitable.

Then, we would like to implement the fitted ARIMA with external variables model on the test dataset(1998.1-2000.12) in order to do forecast. As we can see in Figure 22, the dashed line is the true value of test dataset, the blue line is the forcasted value of test dataset. Then, we calculated the Root Mean Squard Error (RMSE) on the test dataset.
⋆ RMSE = 3.42 ⋆

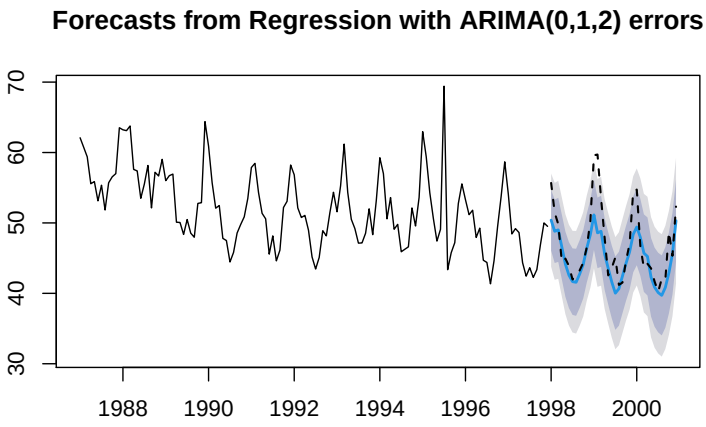


Figure 22 The prediction on test dataset with ARIMA with external variables model

6 Conclusion

After running the univariate and multivariate models, we finally find that the best model is ARIMA with external variables model, which is a multivariate model.

The conclusion makes sense bacause we know that the cardiovascular death should be correlated with the temperature (temp) and pollution (pm10 and o3). Thus, the result is better when we consider the external variables. Above all, the model is good to predict the number of ardiovascular death.