Honor Project
Xinyue Zhang
B00722712

Introduction

Nowadays, sensible resource allocation has been the most important topic around the world, especially human resource. Rational human resource allocation makes work more efficient and flexible. In hospital, if we can arrange proper doctors' quantity, there would be more efficient work. Therefore, we will analyze the quantity of patients in this project in order to arrange proper quantity of doctors every day. In this project, there are five years' data about the quantity of patients every day in different four hospital sites (CCHC, DGH, HCH, QEII) in Halifax from April 1st, 2011 to March 31st, 2016. We will use first four years and six months' data to fit a model and predict next six months' data in fifth year, then compare the predicted data with the real data. Finally, we can get a model that can predict the quantity of patients in the last six months' data in fifth year, which means we can help hospital to manage doctors in order to avoid waste of human resource.

Method

First, we use first four years and six months' data in one site to do time series plot and to check whether there is a trend or periodic signal. We also use the spectra diagram to decide the most obvious period. From the plot, there are an increasing trend in every year and weekly periodicity. Then, we fit generalized addictive model to smooth the data. We assume that the data follow a passion distribution and the mean is λ.
The gam function is $\log(\lambda) \sim \alpha + \beta * \mathrm{t} + \gamma * \mathrm{week} + \mathrm{s}(\mathrm{day})$, where s(day) is smooth function of variable day and we define week is a categories variable.
After smooth the data, we get the fitted values and use these to predict the fitted value in the last six months in fifth year as predicted variable p1.
Next, we compute the residuals that are calculates as the real values mins the fitted value. Then we fit the Autoregressive Integrated Moving Average model to residuals. The ARIMA(p,q) model is $X_t - \sum_{i=1}^{p} \alpha_i X_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \theta_j X_{t-j}$. Then we use first four years and six months' residuals to predict first week's residuals in the last six months in fifth year by ARMA. Then combining the first four years and six months' residuals and real first week's residuals in the last six months in fifth year, we can predict the second week's residuals in the last six months in fifth years. Using this circulation, we get the predicted residuals in the last six months in fifth years as predicted variable p2.
We add p1 and p2 as our final predicted values, and choose 5% significant level to build up confidence interval. Then we compare our predicted value with the observed values in the fifth year. In the last, we calculate the mean squared prediction error and standard prediction error. Using the formula:

$$MSPE = \frac{\sum_{i=1}^{n}(obseved\ values - predicted\ values)^2}{n}$$

$$SSPE = \sqrt{\frac{\sum_{i=1}^{n}(obseved\ values - predicted\ values)^2}{n}}$$

Finally, we use the log of data to refit this model. Assume log(data) follow a normal distribution and the mean is λ. The gam function is $\mathrm{g}(\lambda) \sim \alpha + \beta * \mathrm{t} + \gamma * \mathrm{week} + \mathrm{s(day)}$, where s(day) is smooth function of variable day and we define week is a categories variable. We also use ARMA model to fit residuals. The process is as same as before. In the last, we also get the predicted values for next year and calculate MSPE and SSPE. Compared with these MSPE and SSPE, we need to check whether there exists some improvement.

Results

<u>For CCHC site:</u>

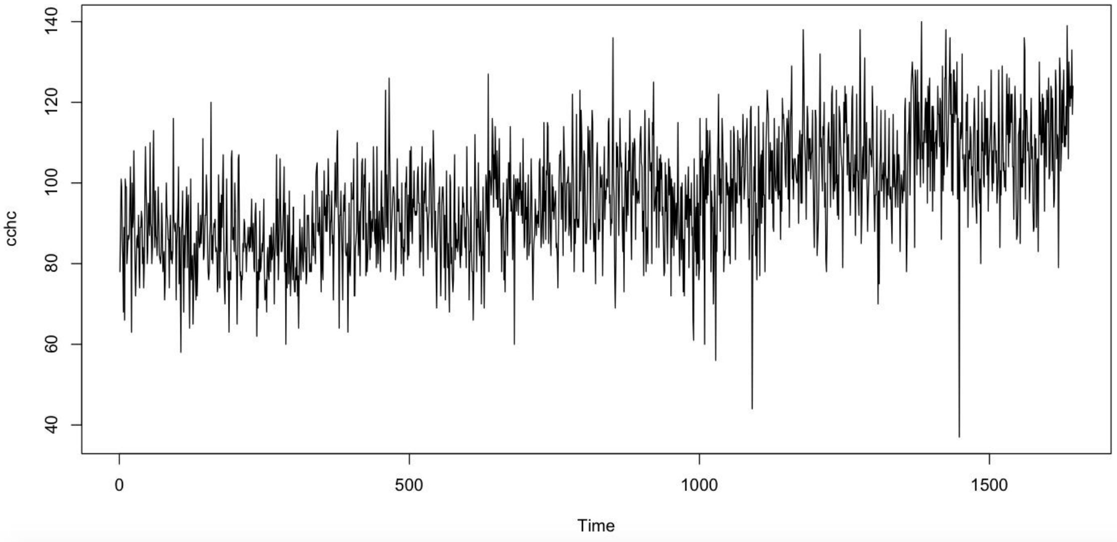


Figure 1. Time series plot of observed data

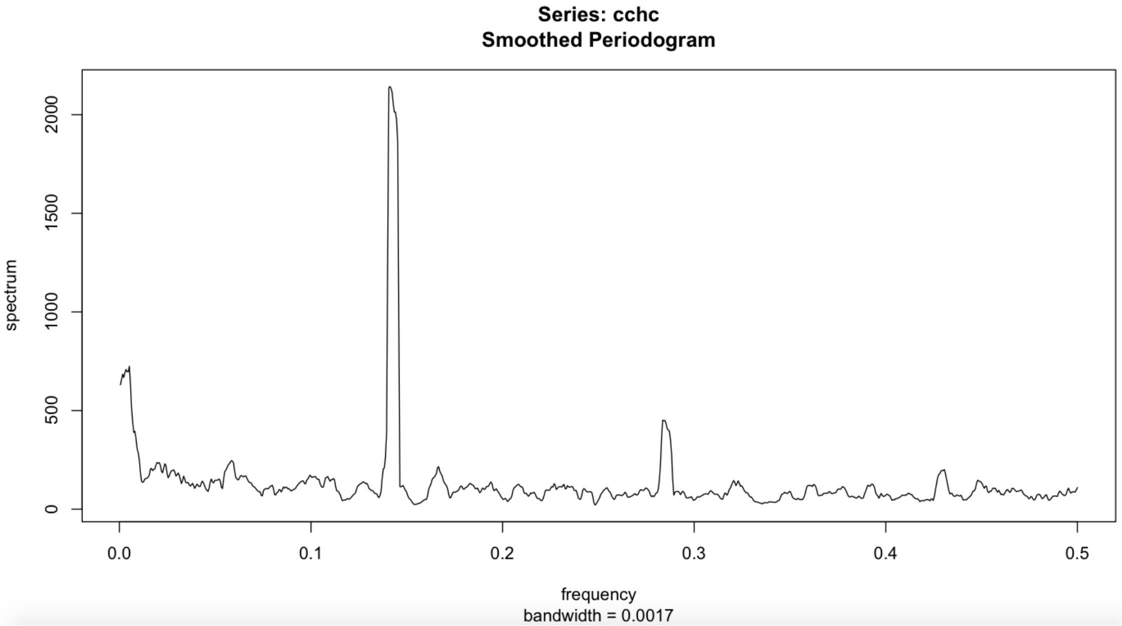


Figure 2. spectra diagram of observed data

From Figure 1, we can see there is periodic signal, and a slightly increasing trend in every year. From Figure 2, the peak is at frequency 0.14, we can infer there is a weekly period. Therefore, we define week as a categories variable to smooth data. And for days in every year, we cannot ensure whether there is linear relation, so we define day is smooth function. For the whole data, we use GAM to smooth and remove periodicity. Then we fit GAM for observed data. The function is $\log(\lambda) \sim \alpha + \beta * \mathrm{t} + \gamma * \mathrm{week} + \mathrm{s}(\mathrm{day})$.

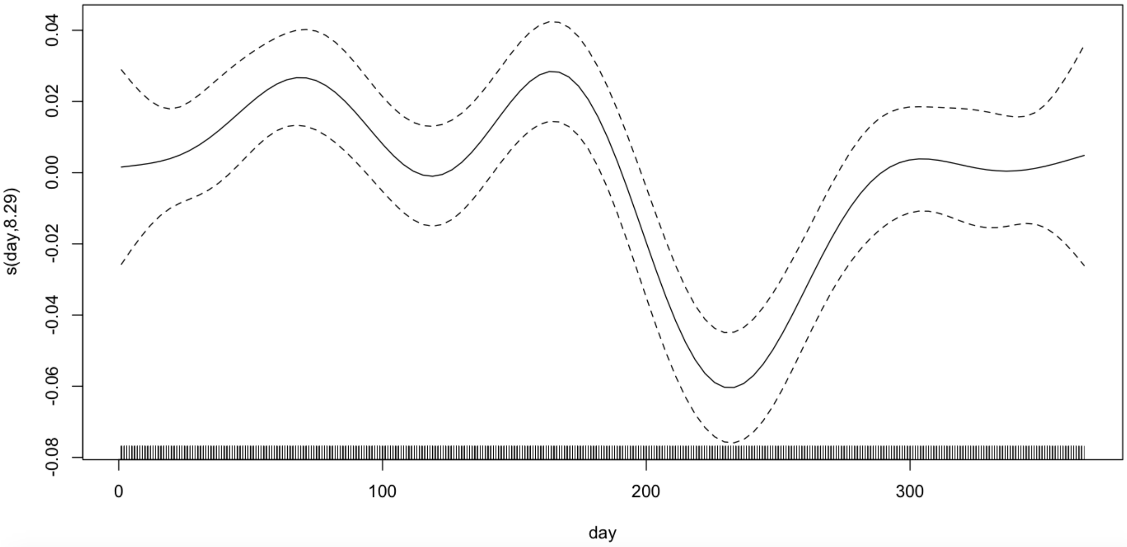


Figure 3. Plot of the smooth components(day) of the fitted GAM.

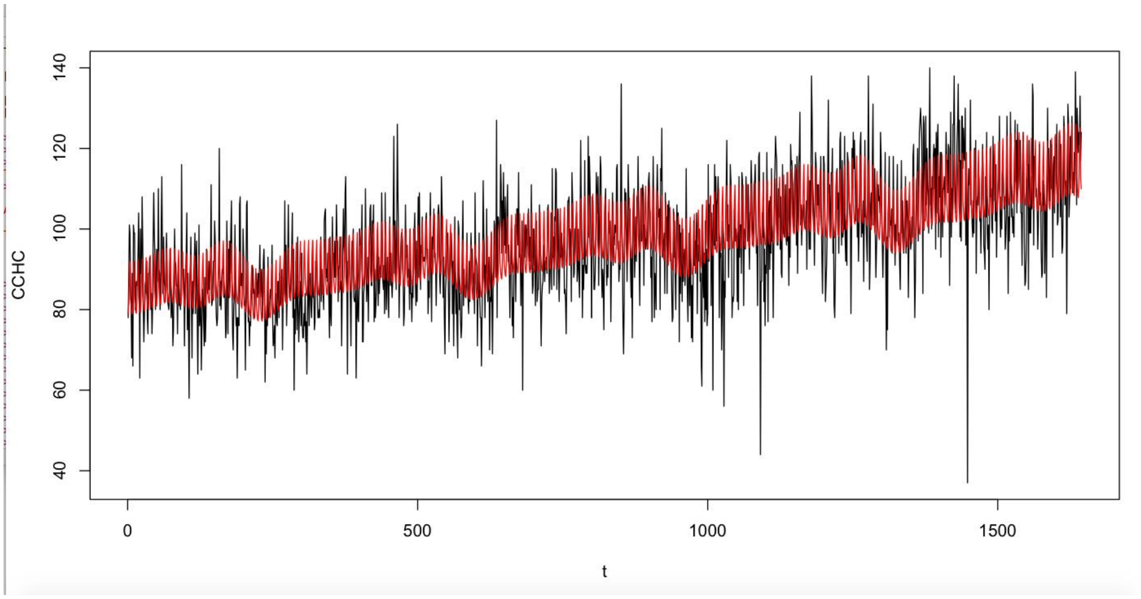
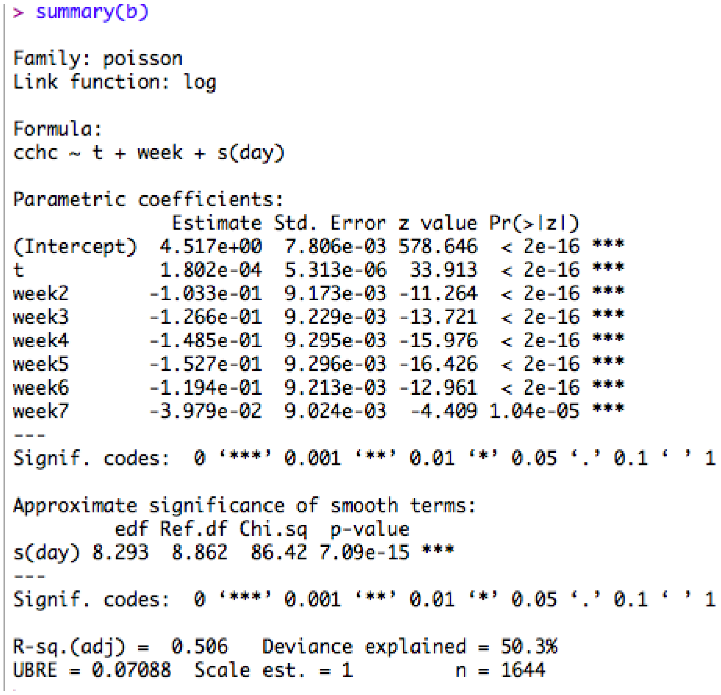
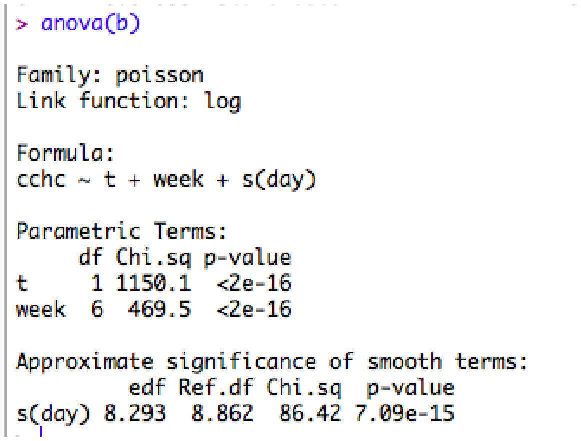


Figure 4. Plot of observed data (black line) versus fitted data (red line).

From Figure 3 and 4, we can see the fitted data versus the original data and check the smooth components behaviors. In Figure 3, it shows the pattern in the year.



```
> summary(b)

Family: poisson
Link function: log

Formula:
cchc ~ t + week + s(day)

Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.517e+00  7.806e-03 578.646  < 2e-16 ***
t            1.802e-04  5.313e-06  33.913  < 2e-16 ***
week2       -1.033e-01  9.173e-03 -11.264  < 2e-16 ***
week3       -1.266e-01  9.229e-03 -13.721  < 2e-16 ***
week4       -1.485e-01  9.295e-03 -15.976  < 2e-16 ***
week5       -1.527e-01  9.296e-03 -16.426  < 2e-16 ***
week6       -1.194e-01  9.213e-03 -12.961  < 2e-16 ***
week7       -3.979e-02  9.024e-03  -4.409 1.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq  p-value
s(day) 8.293  8.862  86.42 7.09e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.506   Deviance explained = 50.3%
UBRE = 0.07088  Scale est. = 1         n = 1644
```



```
> anova(b)

Family: poisson
Link function: log

Formula:
cchc ~ t + week + s(day)

Parametric Terms:
     df Chi.sq p-value
t     1 1150.1  <2e-16
week  6  469.5  <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq  p-value
s(day) 8.293  8.862  86.42 7.09e-15
```

From ANOVA, there are all significant for variables.

Then we predict the GAM for the next year. Firstly, we consider the fitted values prediction for GAM.

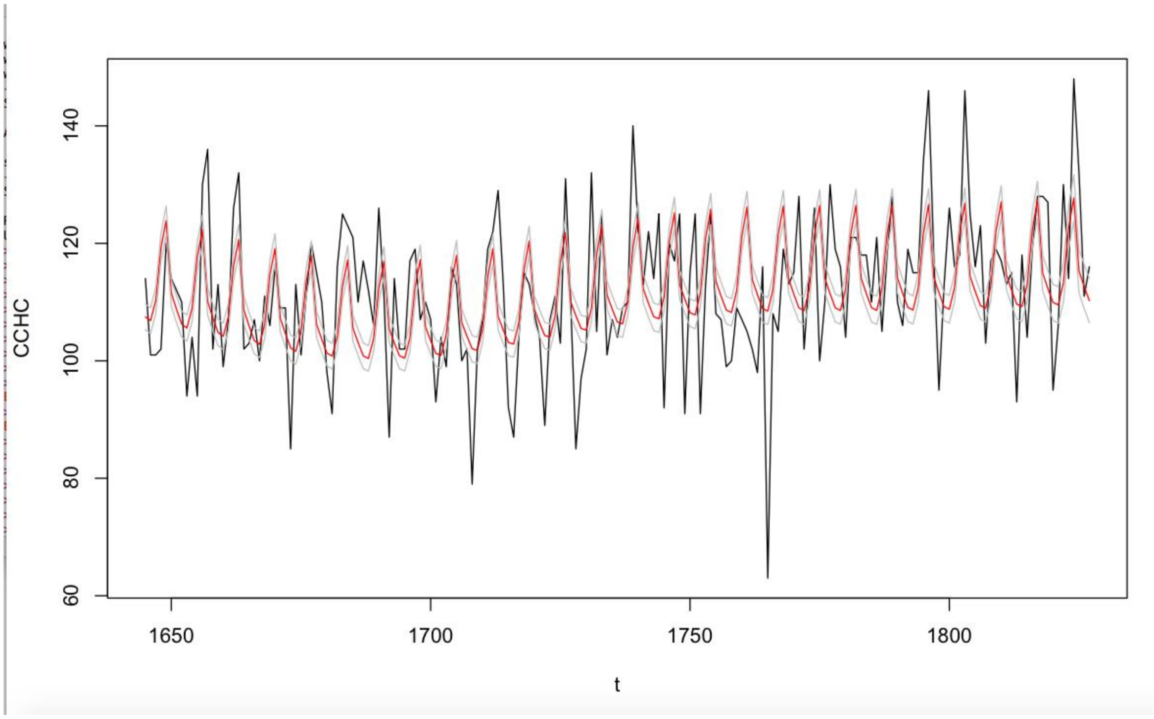
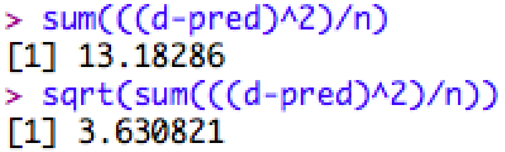


Figure 5. Prediction plot of observed data (black line) versus predicted data (red line), and 95% confidence interval l(grey line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 13.18286
> sqrt(sum(((d-pred)^2)/n))
[1] 3.630821
```

In the case, the MSPE is 13.18286, and the SSPE is 3.630821.

Secondly, we consider the residuals prediction. From Figure 6, we can see the residuals is approximately normal distribution.

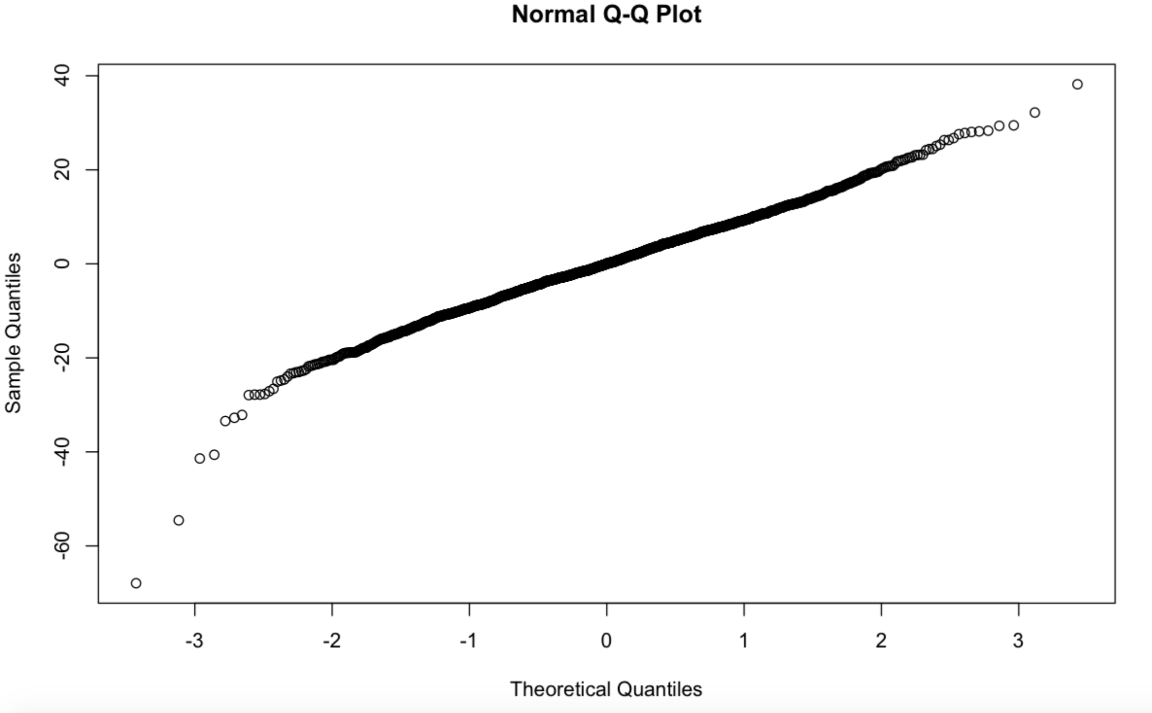
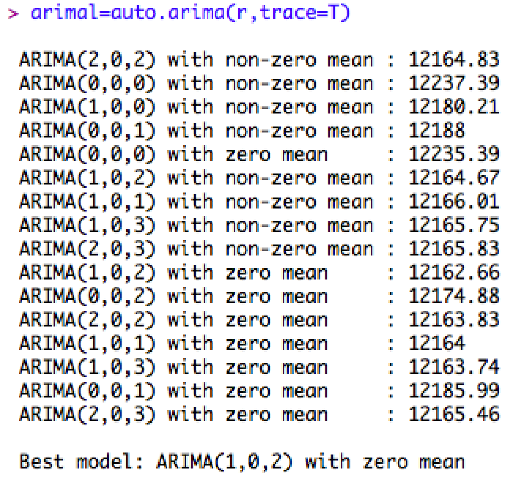


Figure 6. Residuals QQ plot.

Then we use ARMA to fit the residuals. In this case, we choose ARIMA(1,0,2). The ARIMA(1,0,2) is $X_t - \alpha_1 X_{t-1} = \varepsilon_t + \theta_1 X_{t-1} + \theta_2 X_{t-2}$



```
> arima1=auto.arima(r,trace=T)

 ARIMA(2,0,2) with non-zero mean : 12164.83
 ARIMA(0,0,0) with non-zero mean : 12237.39
 ARIMA(1,0,0) with non-zero mean : 12180.21
 ARIMA(0,0,1) with non-zero mean : 12188
 ARIMA(0,0,0) with zero mean     : 12235.39
 ARIMA(1,0,2) with non-zero mean : 12164.67
 ARIMA(1,0,1) with non-zero mean : 12166.01
 ARIMA(1,0,3) with non-zero mean : 12165.75
 ARIMA(2,0,3) with non-zero mean : 12165.83
 ARIMA(1,0,2) with zero mean     : 12162.66
 ARIMA(0,0,2) with zero mean     : 12174.88
 ARIMA(2,0,2) with zero mean     : 12163.83
 ARIMA(1,0,1) with zero mean     : 12164
 ARIMA(1,0,3) with zero mean     : 12163.74
 ARIMA(0,0,1) with zero mean     : 12185.99
 ARIMA(2,0,3) with zero mean     : 12165.46

 Best model: ARIMA(1,0,2) with zero mean
```

Then we predict everyday' residual in next year by circulation.

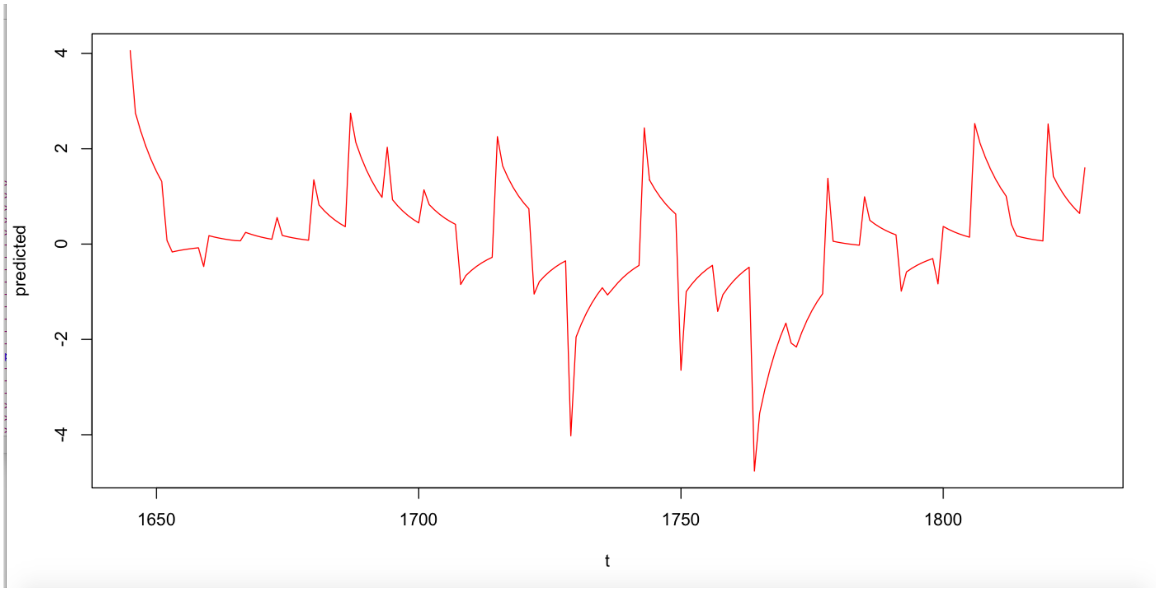


Figure 7. Plot of predicted residuals.

Then we add these two prediction (GAM prediction and residuals prediction), and choose 5% significant level.

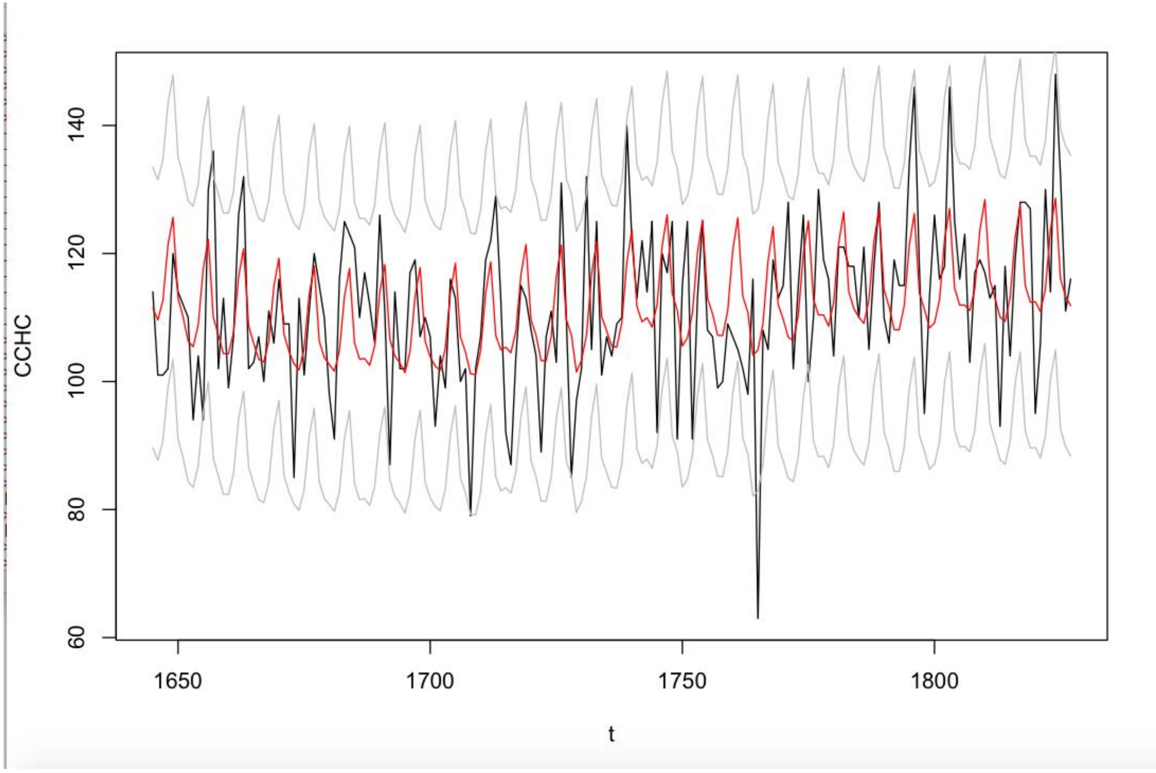
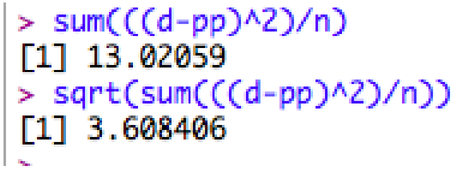


Figure 8. Plot of observed values (black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure8, we can see all original data are in the confidence interval, and for predicted data, the trend is almost the same as the original data. Then we calculate the mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 13.02059
> sqrt(sum(((d-pp)^2)/n))
[1] 3.608406
```

In the case, the MSPE is 13.02059, and the SSPE is 3.608406, which is less than prediction only by GAM effect.

Next, we try to use log of data to refit this model. The process is as same as above. Hence we get these results.

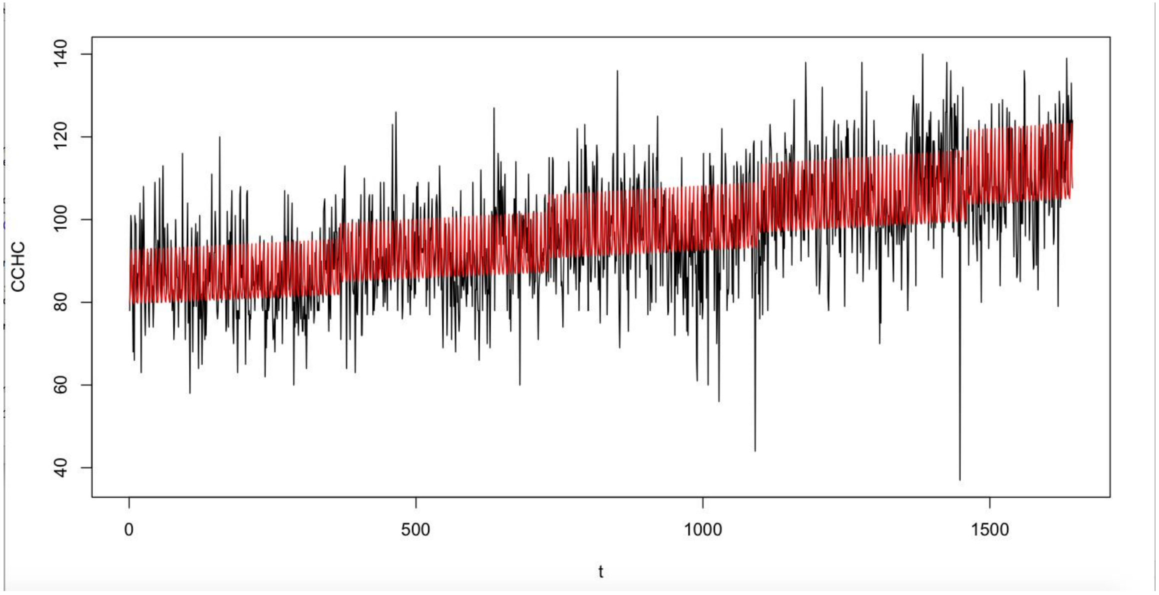


Figure 9. Plot of observed data (black line) versus fitted data (red line).

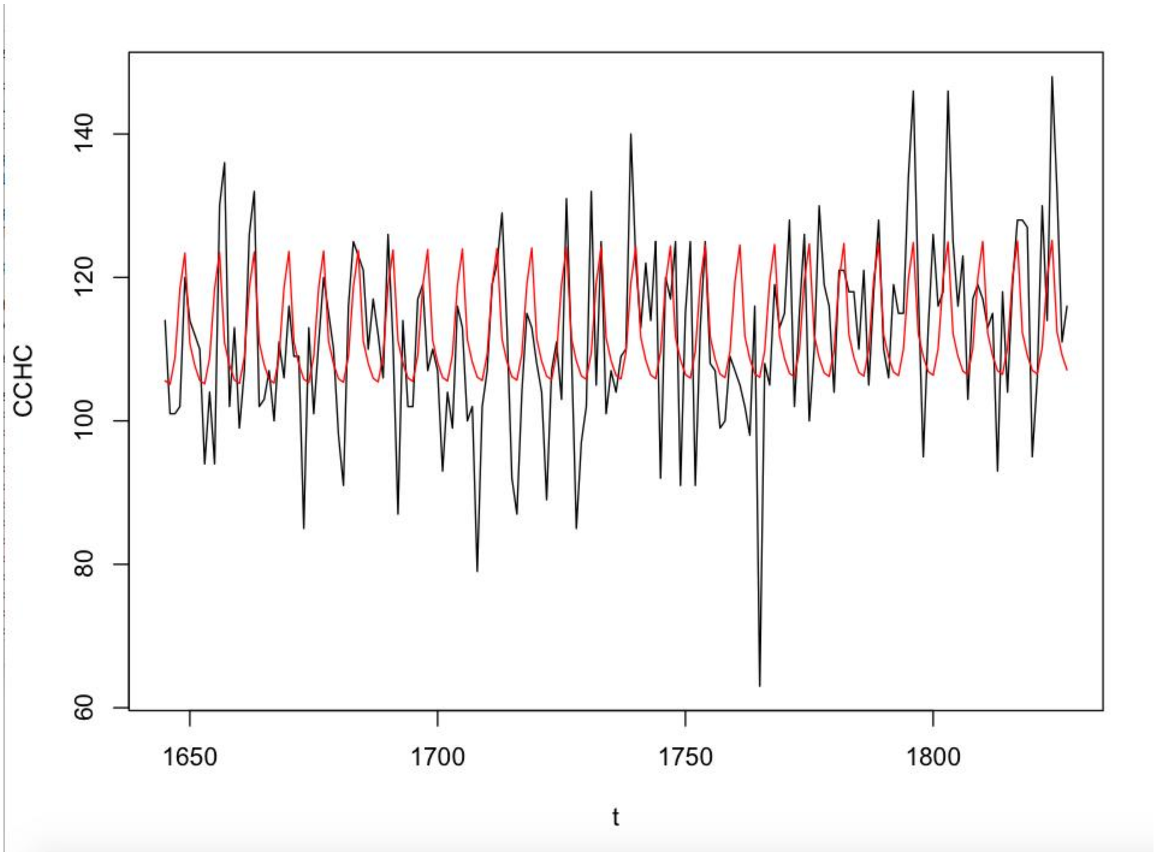
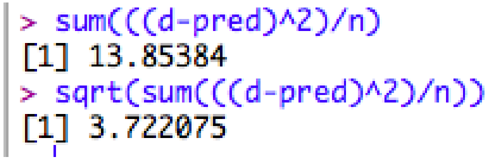


Figure 10. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 13.85384
> sqrt(sum(((d-pred)^2)/n))
[1] 3.722075
```

From Figure 10, the MSPE is 13.85384, and the SSPE is 3.722075.

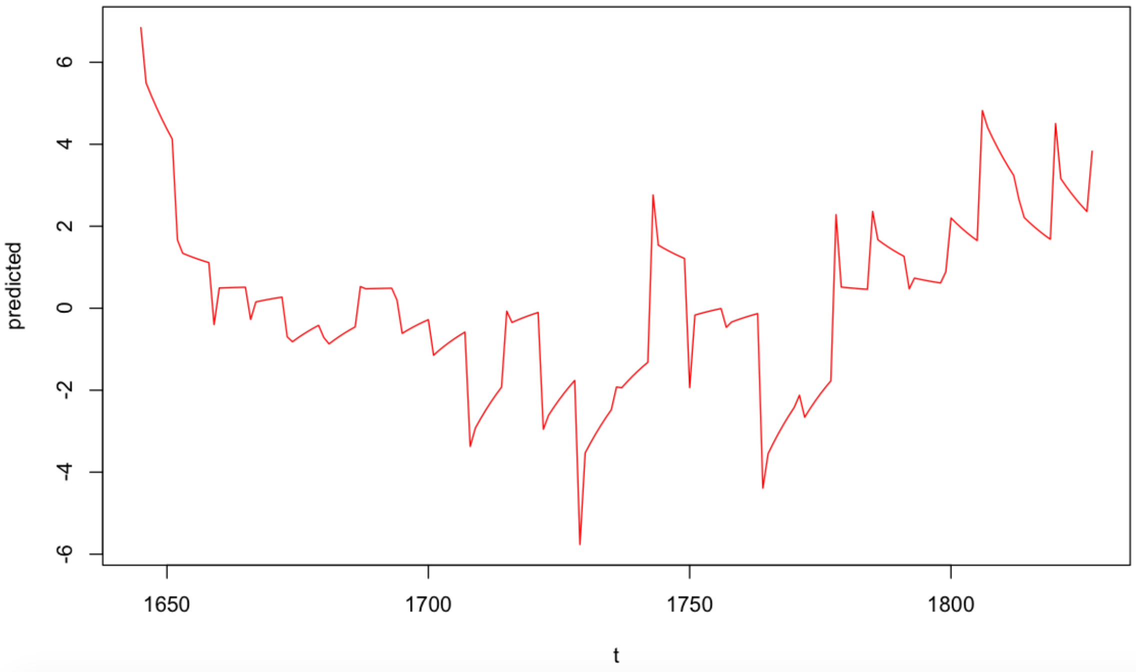


Figure 11. Plot of predicted residuals.

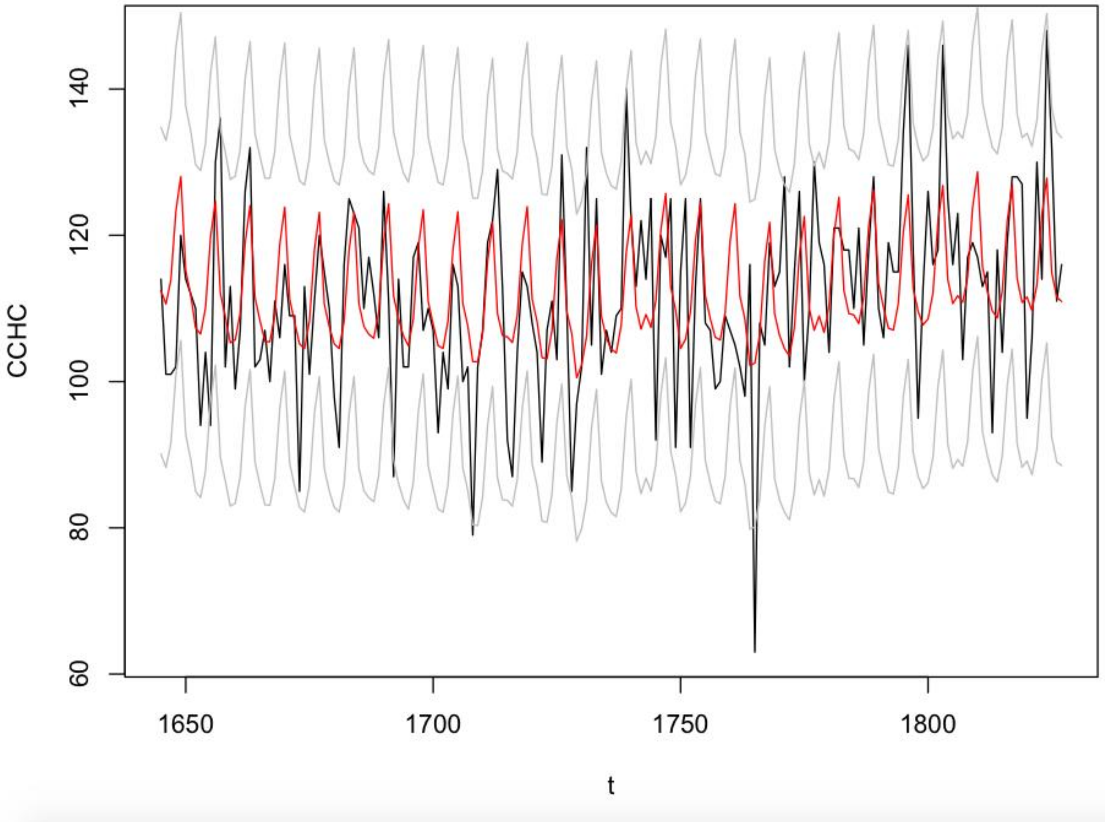
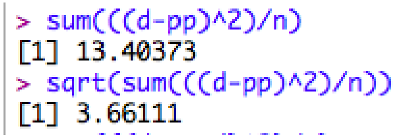


Figure 12. Plot of observed values (black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure 12, the prediction results look roughly same as the Figure 8. We also calculate mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 13.40373
> sqrt(sum(((d-pp)^2)/n))
[1] 3.66111
```

In this case, MSPE is 13.40373, and SSPE is 3.661. We found these are more than prediction by GAM and ARMA effect, which means that improvement cannot be achieved.

For DGH site:

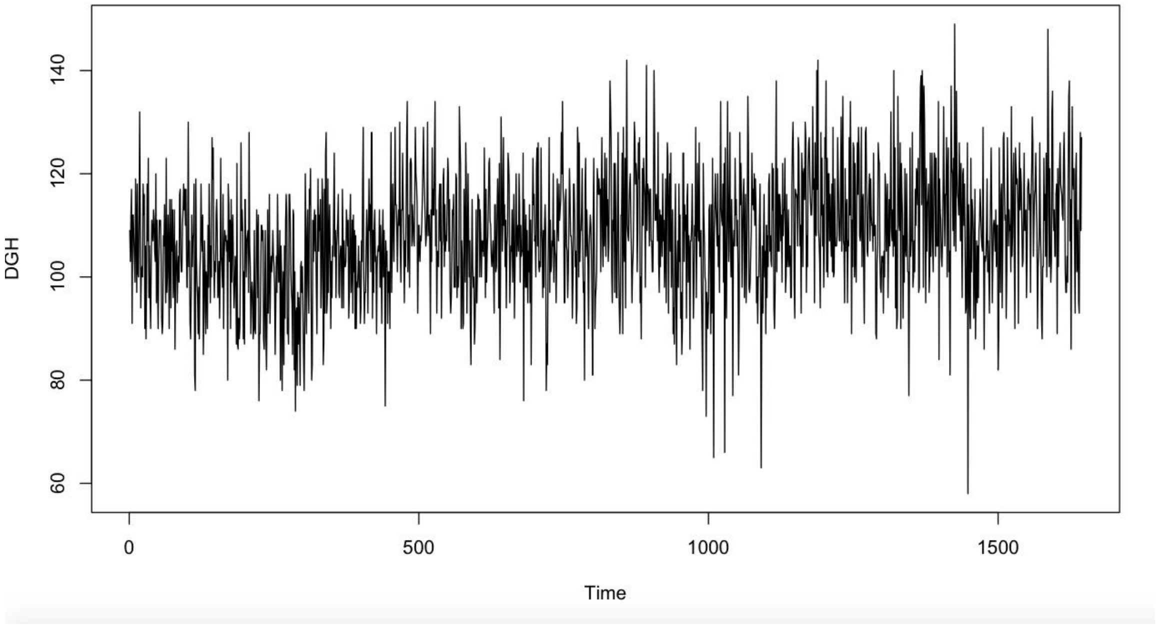


Figure 1. Time series plot of observed data

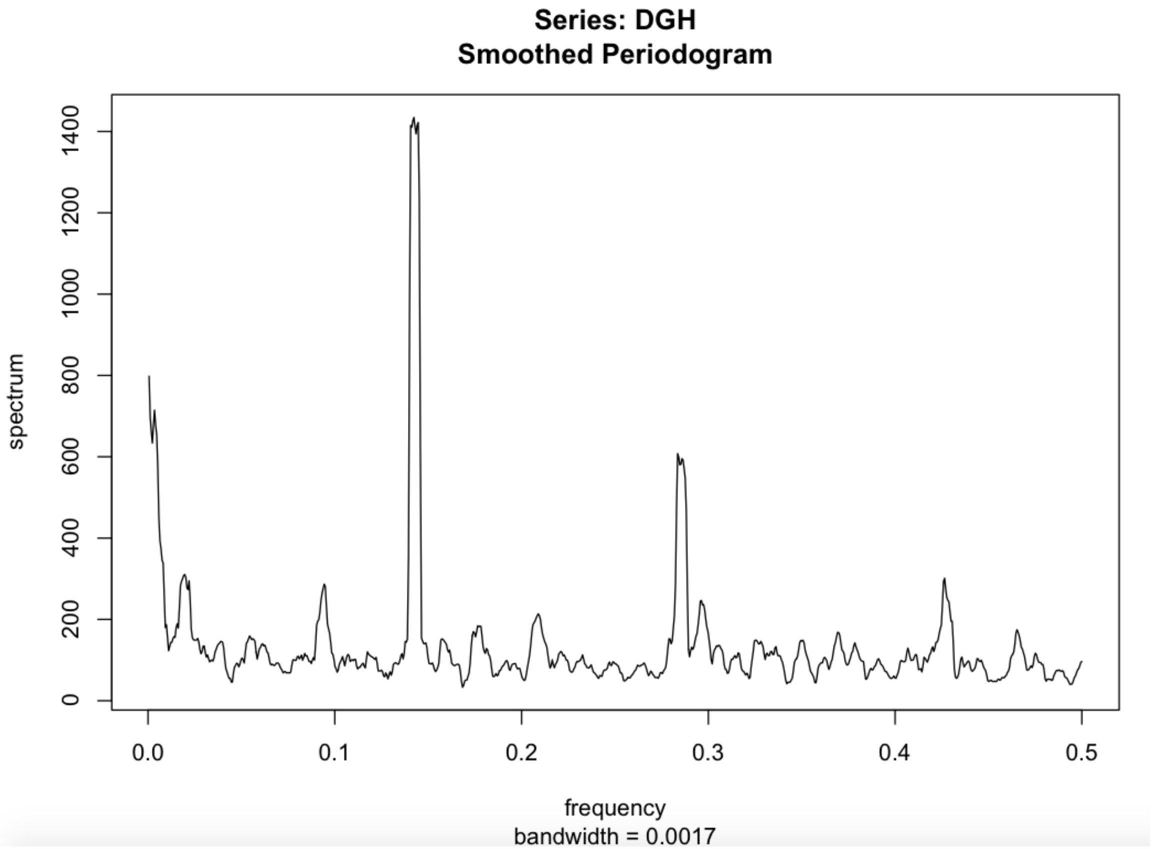


Figure 2. spectra diagram of observed data

From Figure 1, we can see there is periodic signal, and a slightly increasing trend in every year. From Figure 2, the most obvious peak is at frequency 0.14, we can infer there is a weekly period. Therefore, we define week as a categories variable to smooth data. And for days in every year, we cannot ensure whether there is linear relation, so we define day is smooth function. For the whole data, we use GAM to smooth and remove periodicity. Then we fit GAM for observed data. The function is $\log(\lambda) \sim \alpha + \beta * t + \gamma *$

week + s(day).

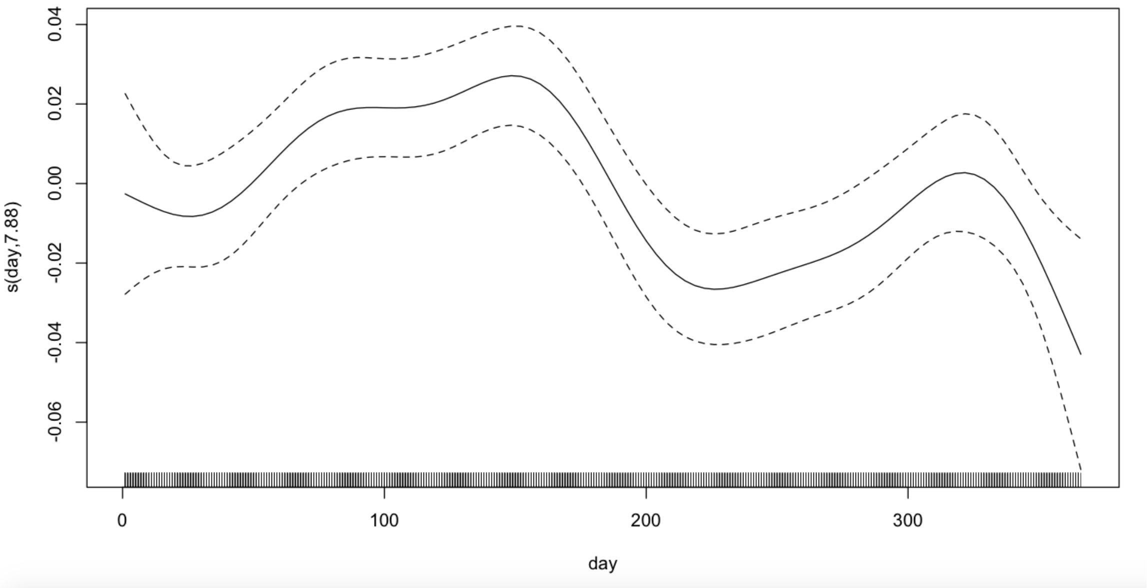


Figure 3. Plot of the smooth components(day) of the fitted GAM.

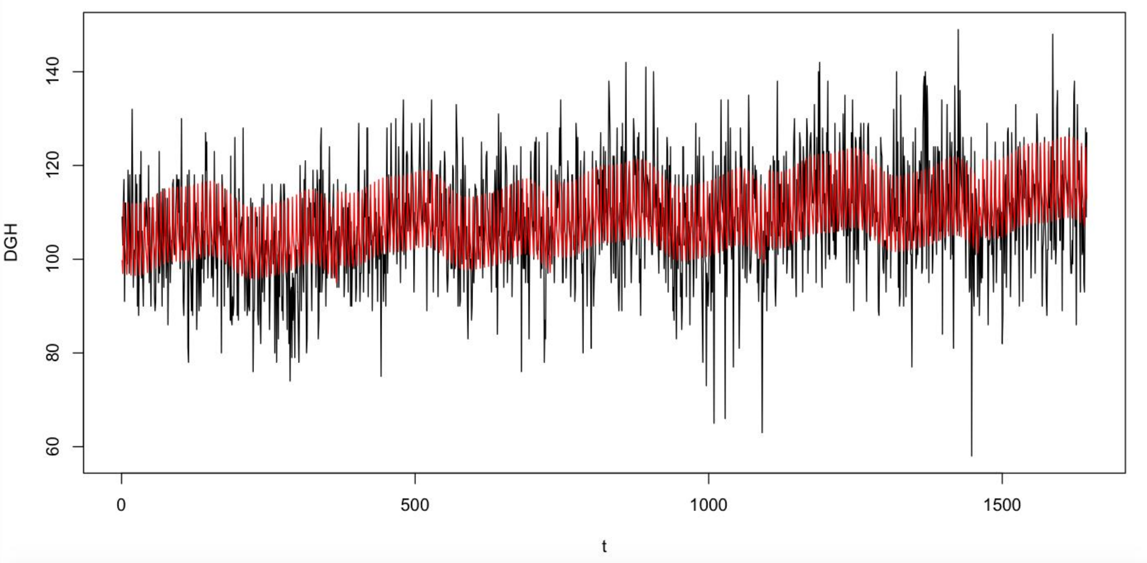
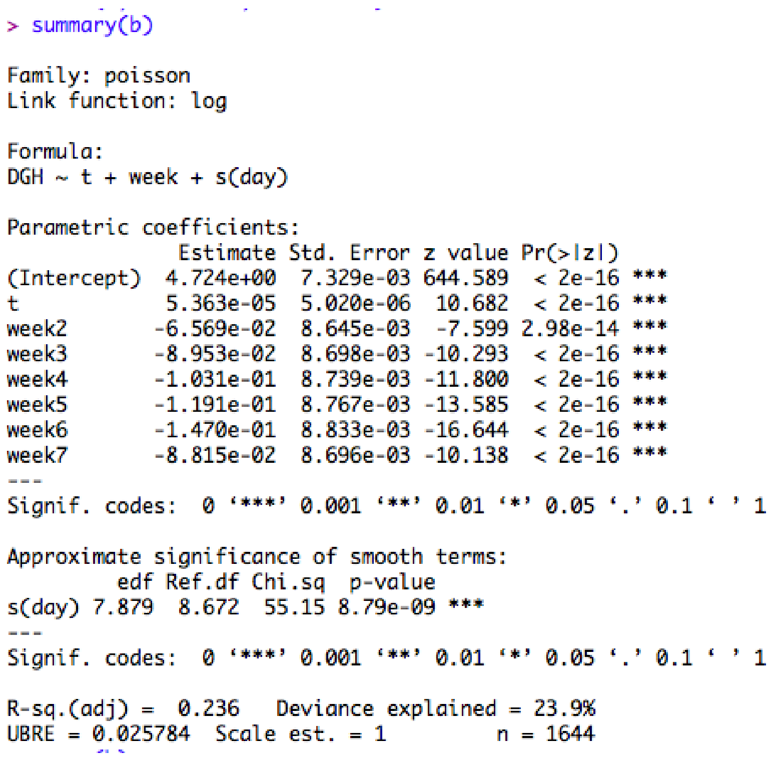
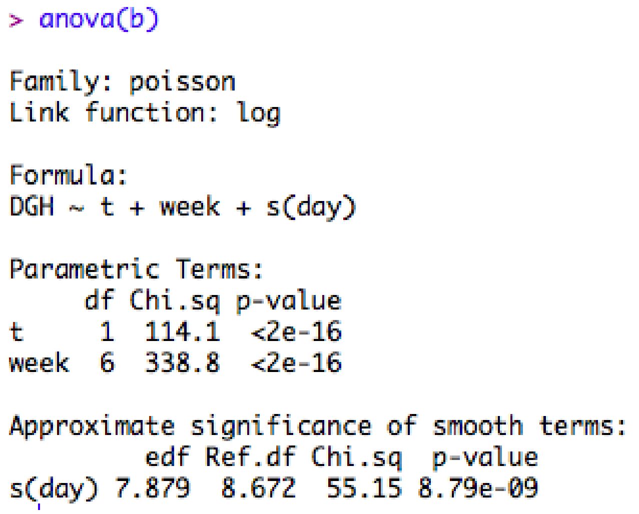


Figure 4. Plot of observed data (black line) versus predicted data (red line).

From Figure 3 and 4, we can see the fitted data versus the original data and check the smooth components behaviors. In Figure 3, it shows the pattern in the year.
From ANOVA, there are all significant for variables.



```
> summary(b)

Family: poisson
Link function: log

Formula:
DGH ~ t + week + s(day)

Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.724e+00  7.329e-03 644.589  < 2e-16 ***
t            5.363e-05  5.020e-06  10.682  < 2e-16 ***
week2       -6.569e-02  8.645e-03  -7.599 2.98e-14 ***
week3       -8.953e-02  8.698e-03 -10.293  < 2e-16 ***
week4       -1.031e-01  8.739e-03 -11.800  < 2e-16 ***
week5       -1.191e-01  8.767e-03 -13.585  < 2e-16 ***
week6       -1.470e-01  8.833e-03 -16.644  < 2e-16 ***
week7       -8.815e-02  8.696e-03 -10.138  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq  p-value
s(day) 7.879  8.672  55.15 8.79e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.236   Deviance explained = 23.9%
UBRE = 0.025784  Scale est. = 1         n = 1644
```



```
> anova(b)

Family: poisson
Link function: log

Formula:
DGH ~ t + week + s(day)

Parametric Terms:
     df Chi.sq p-value
t     1  114.1  <2e-16
week  6  338.8  <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq  p-value
s(day) 7.879  8.672  55.15 8.79e-09
```

From ANOVA, there are all significant for variables.

Then we predict the GAM for the next year. Firstly, we consider the fitted values prediction for GAM.

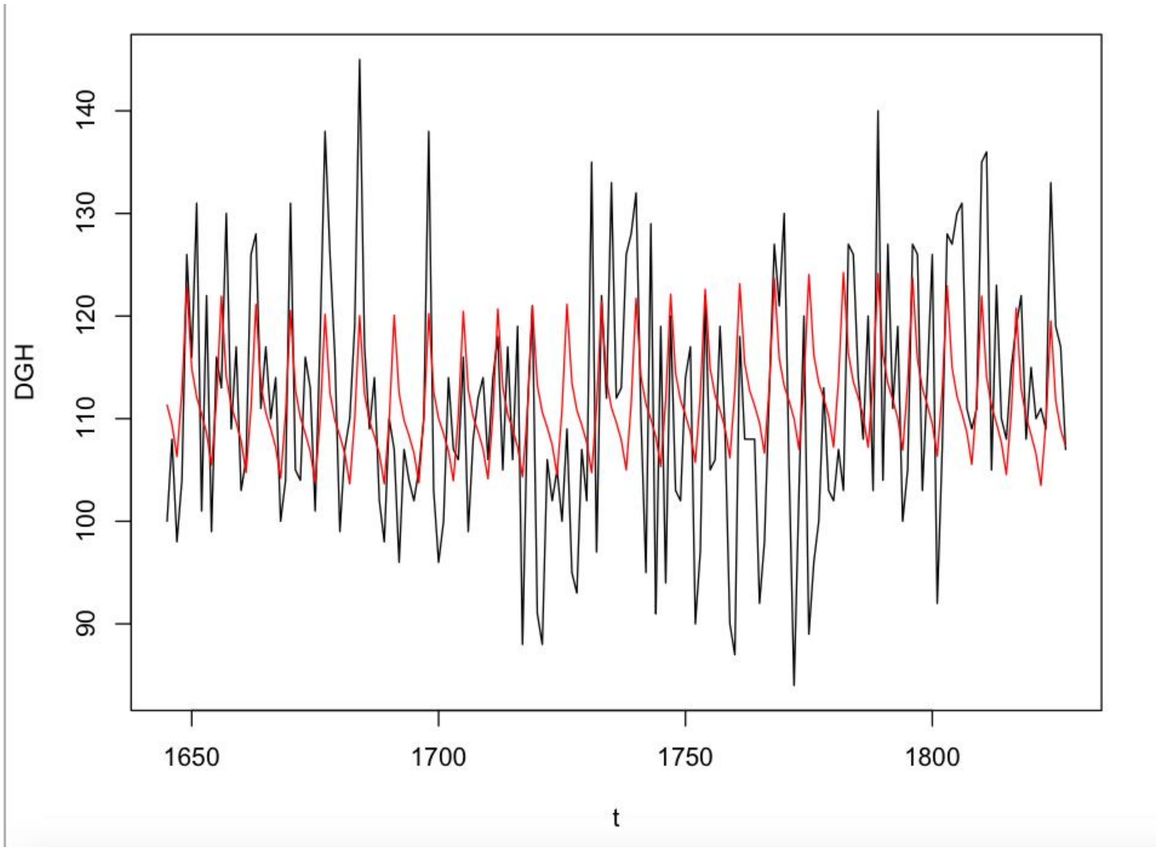
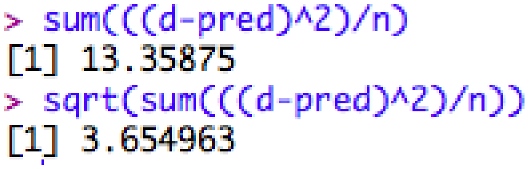


Figure 5. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 13.35875
> sqrt(sum(((d-pred)^2)/n))
[1] 3.654963
```

In the case, the MSPE is 13.35875, and the SSPE is 3.654963.

Secondly, we consider the residuals prediction. From Figure 6, we can see the residuals is approximately normal distribution.

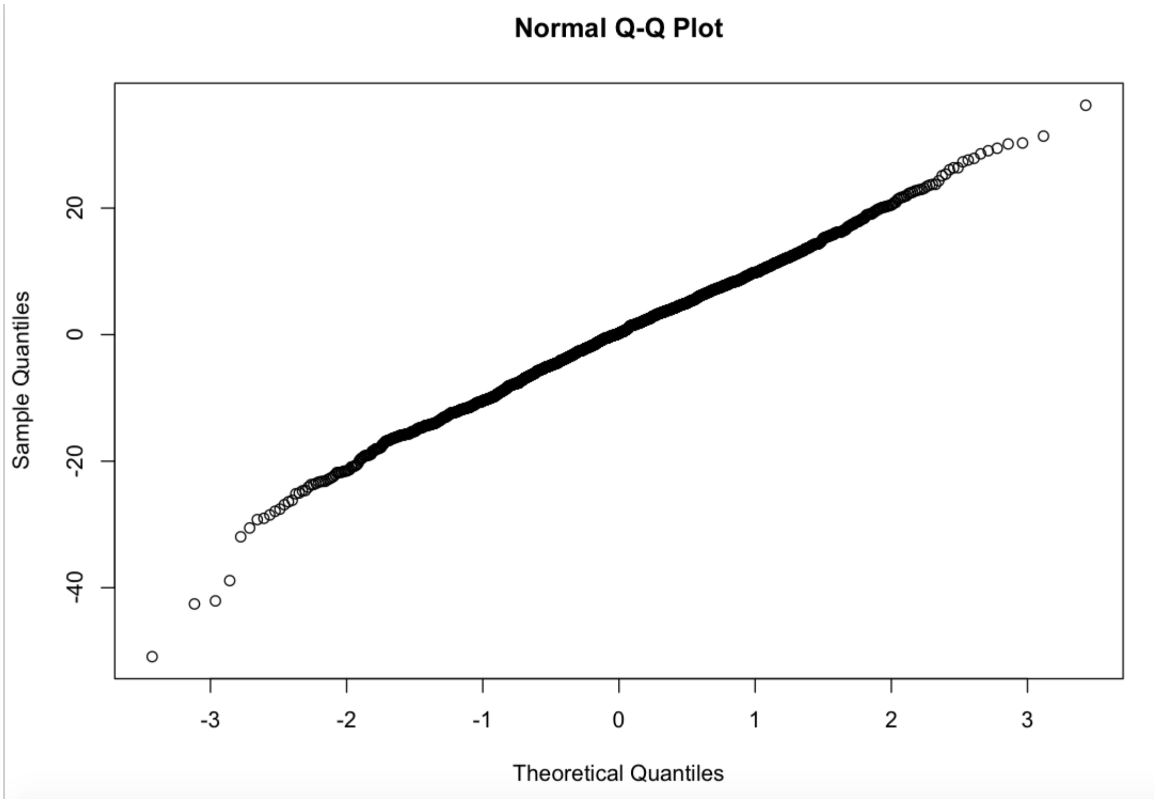
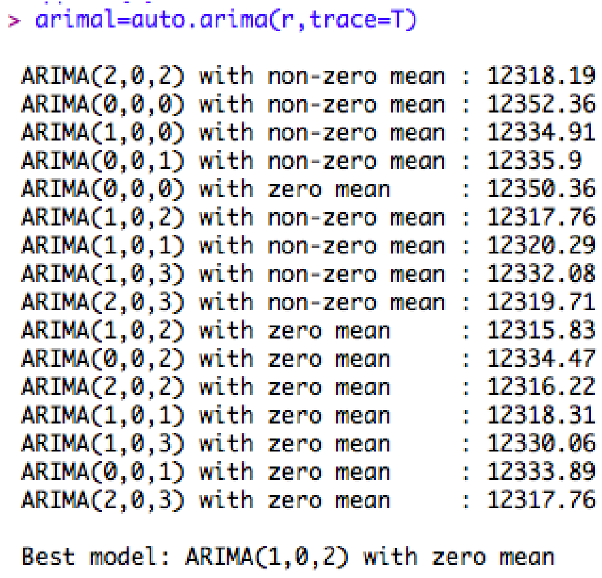


Figure 6. Residuals QQ plot.

Then we use ARMA to fit the residuals. In this case, we choose ARIMA(1,0,2).
The ARMA(1,0,2) is $X_t - \alpha_1 X_{t-1} = \varepsilon_t + \theta_1 X_{t-1} + \theta_2 X_{t-2}$



```
> arima1=auto.arima(r,trace=T)

 ARIMA(2,0,2) with non-zero mean : 12318.19
 ARIMA(0,0,0) with non-zero mean : 12352.36
 ARIMA(1,0,0) with non-zero mean : 12334.91
 ARIMA(0,0,1) with non-zero mean : 12335.9
 ARIMA(0,0,0) with zero mean     : 12350.36
 ARIMA(1,0,2) with non-zero mean : 12317.76
 ARIMA(1,0,1) with non-zero mean : 12320.29
 ARIMA(1,0,3) with non-zero mean : 12332.08
 ARIMA(2,0,3) with non-zero mean : 12319.71
 ARIMA(1,0,2) with zero mean     : 12315.83
 ARIMA(0,0,2) with zero mean     : 12334.47
 ARIMA(2,0,2) with zero mean     : 12316.22
 ARIMA(1,0,1) with zero mean     : 12318.31
 ARIMA(1,0,3) with zero mean     : 12330.06
 ARIMA(0,0,1) with zero mean     : 12333.89
 ARIMA(2,0,3) with zero mean     : 12317.76

 Best model: ARIMA(1,0,2) with zero mean
```

Then we predict everyday’ residual in next year by weekly circulation.

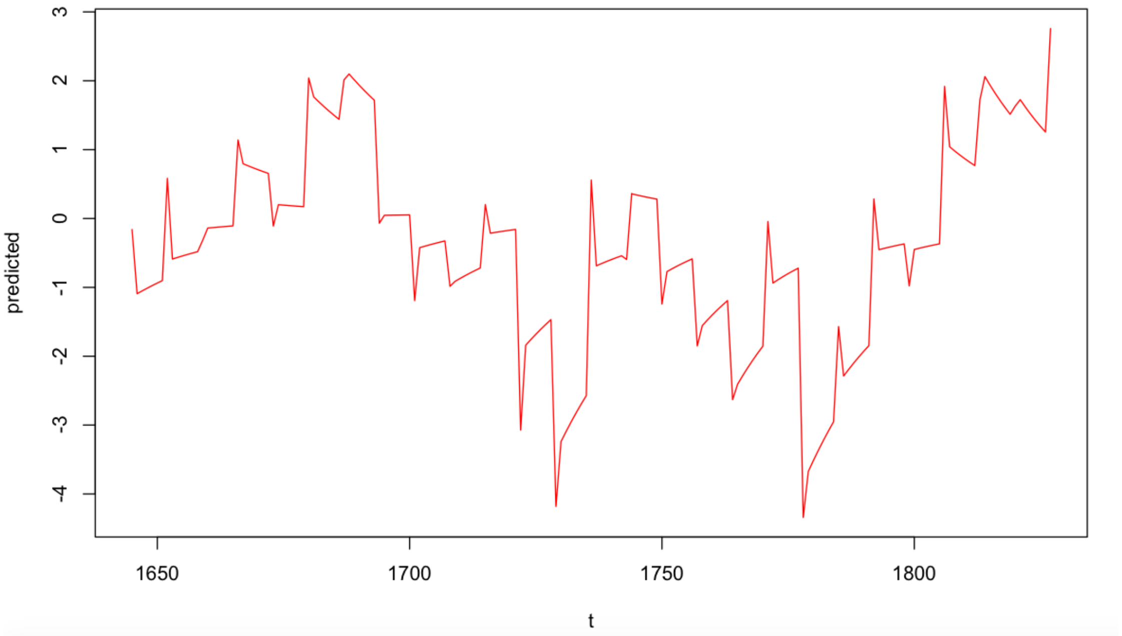


Figure 7. Plot of predicted residuals.

Then we add these two prediction (GAM prediction and residuals prediction), and choose 5% significant level.

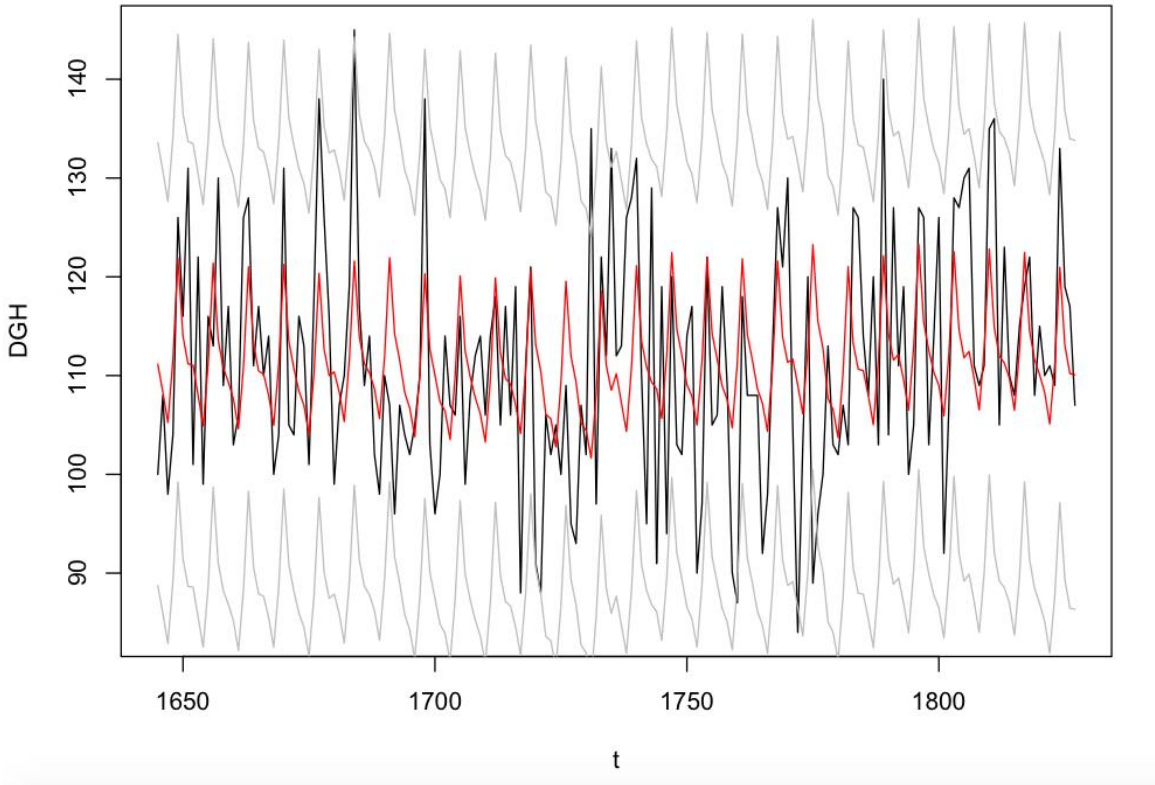
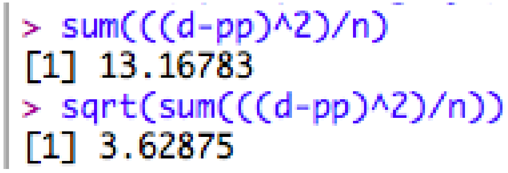


Figure 8. Plot of observed values (black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure8, we can see all the original data are in the confidence interval, and for predicted data, the trend is almost the same as the original data. Then we calculate the mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 13.16783
> sqrt(sum(((d-pp)^2)/n))
[1] 3.62875
```

In the case, the MSPE is 13.16783, and the SSPE is 3.62875, which is less than prediction only by GAM effect.

Next, we try to use log of data to refit this model. The process is as same as above. Hence we get these results.

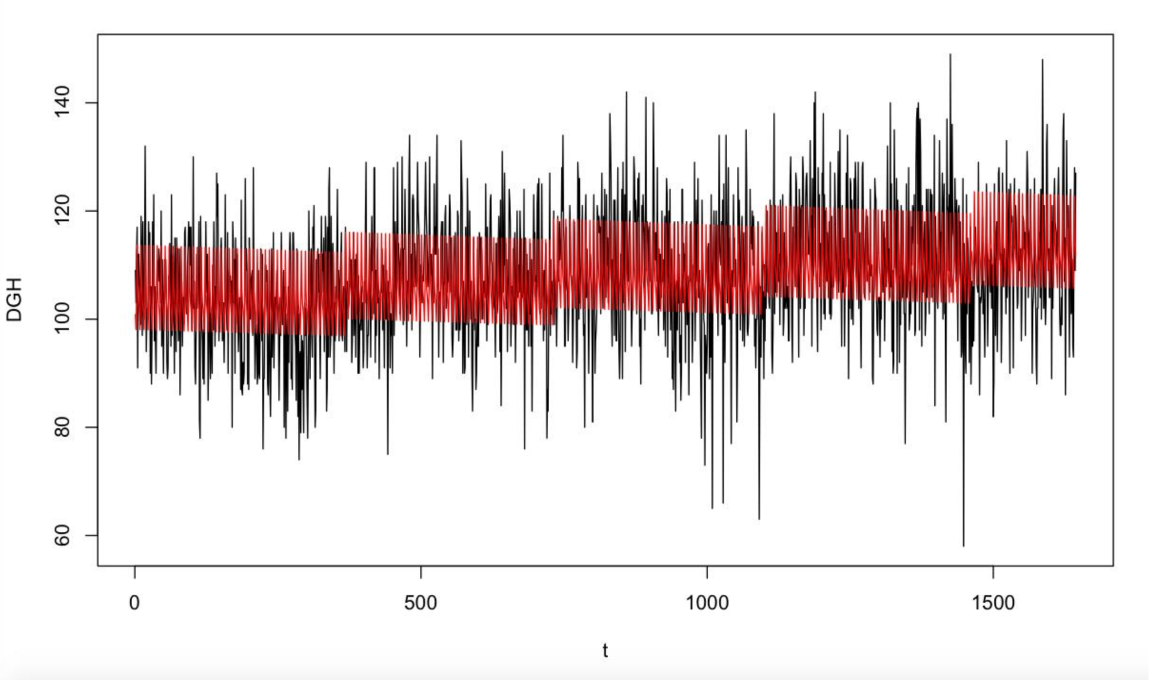


Figure 9. Plot of observed data (black line) versus fitted data (red line).

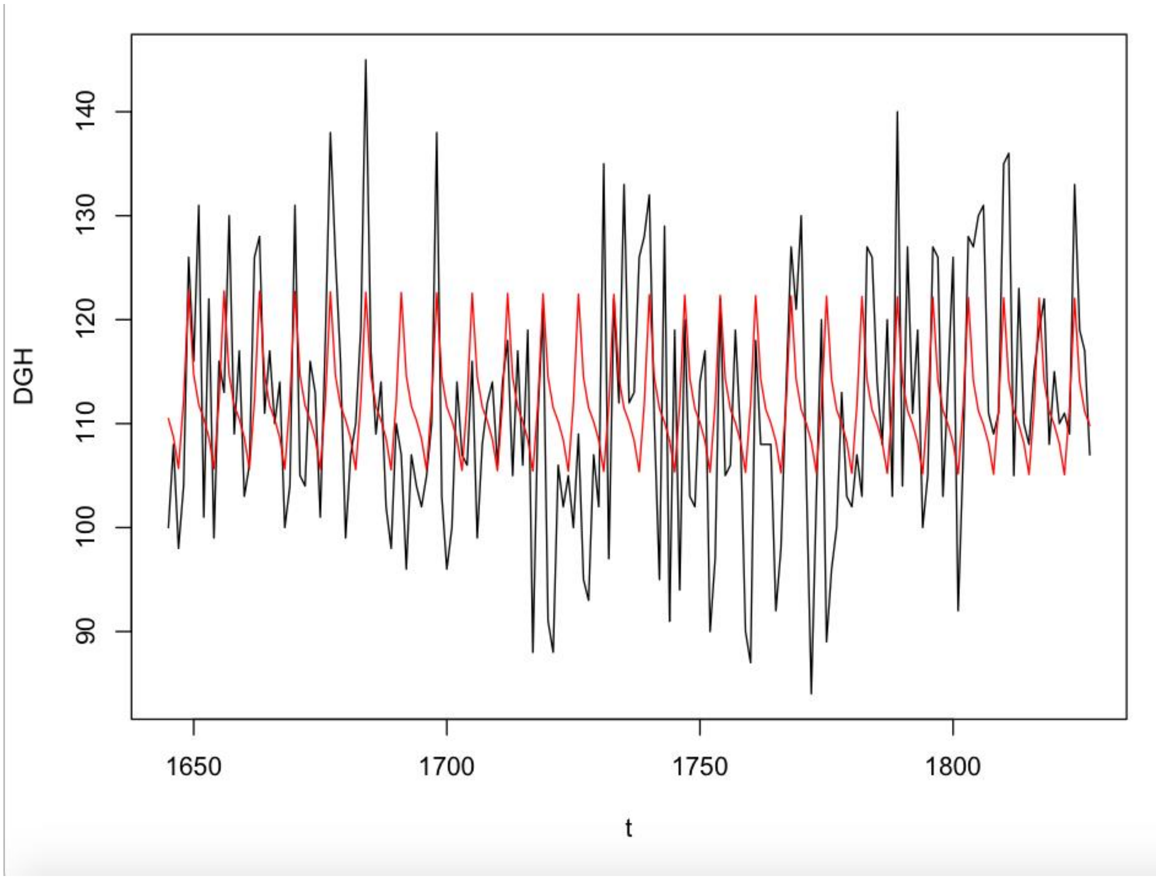
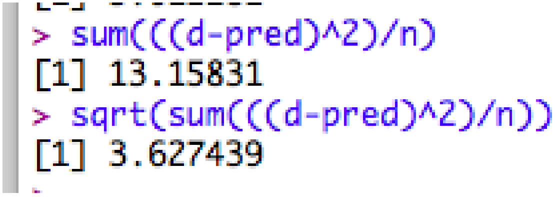


Figure 10. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 13.15831
> sqrt(sum(((d-pred)^2)/n))
[1] 3.627439
```

From Figure 10, the MSPE is 13.15831, and the SSPE is 3.627439.

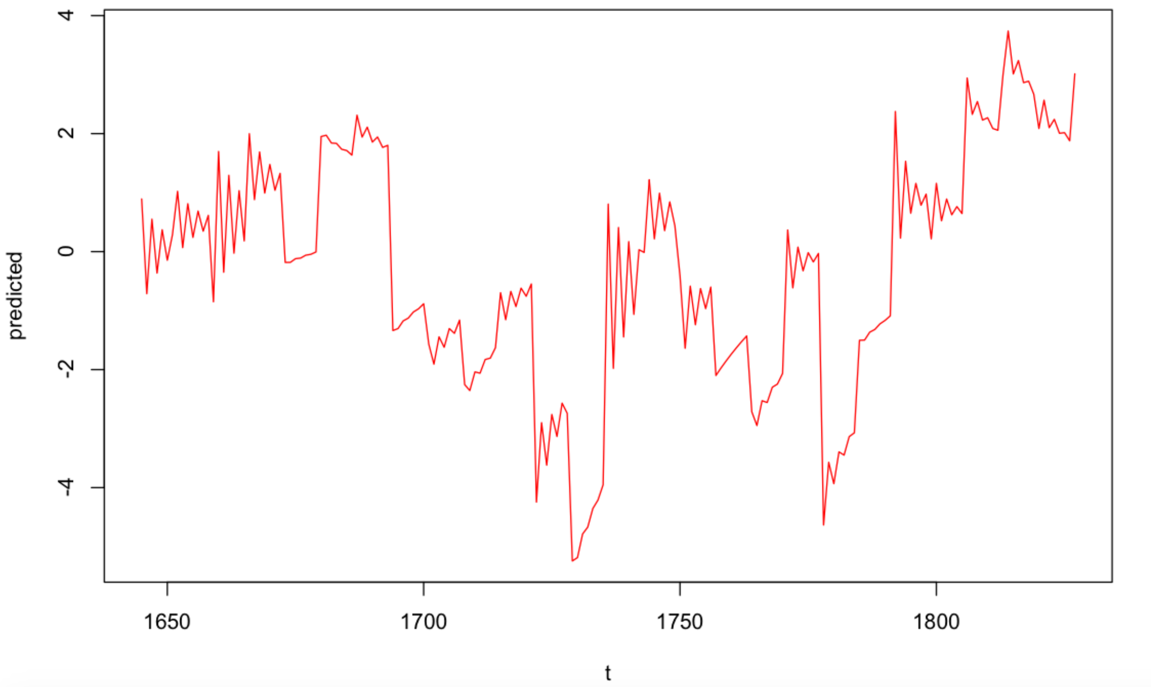


Figure 11. Plot of predicted residuals.

.

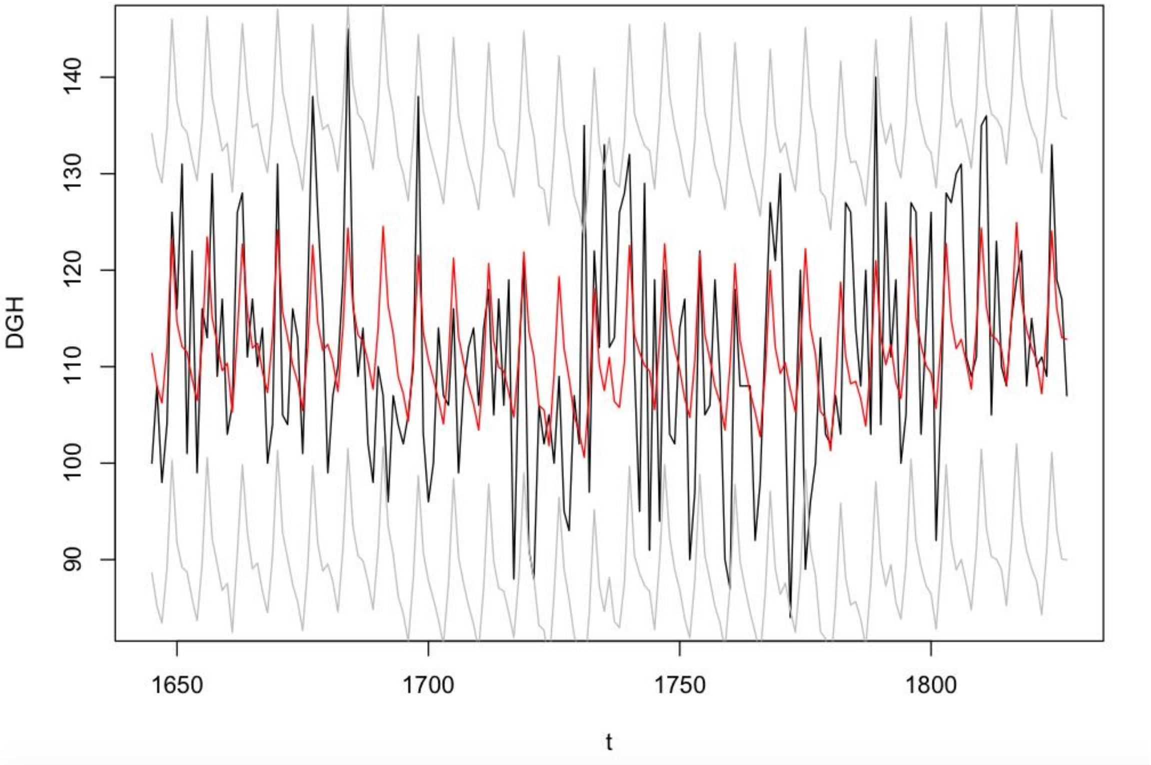
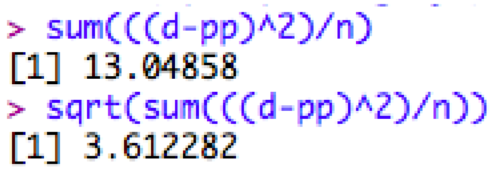


Figure 12. Plot of observed values(black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure 12, the prediction results look roughly same as the Figure 8. We also calculate the calculate the mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 13.04858
> sqrt(sum(((d-pp)^2)/n))
[1] 3.612282
```

In this case, MSPE 13.04858, and SSPE is 3.612282. We found these are less than prediction by GAM and ARMA effect, which means that improvement can be achieved.

For HCH site:

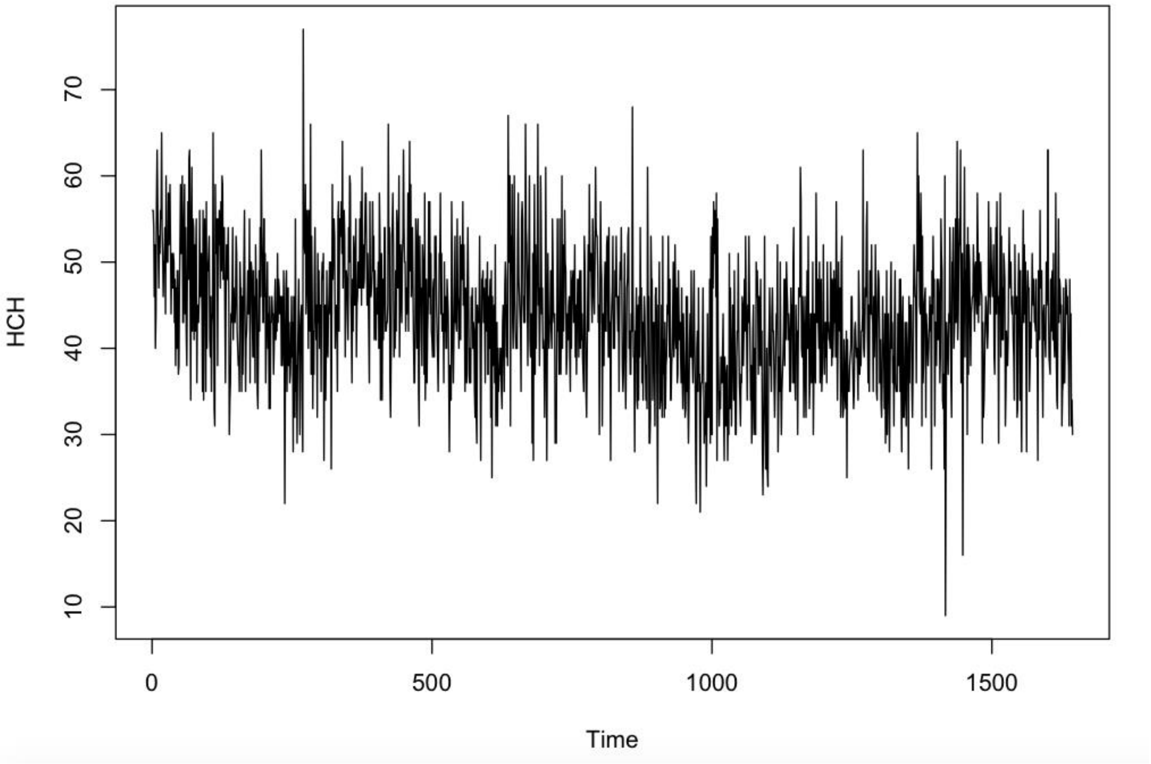


Figure 1. Time series plot of observed data

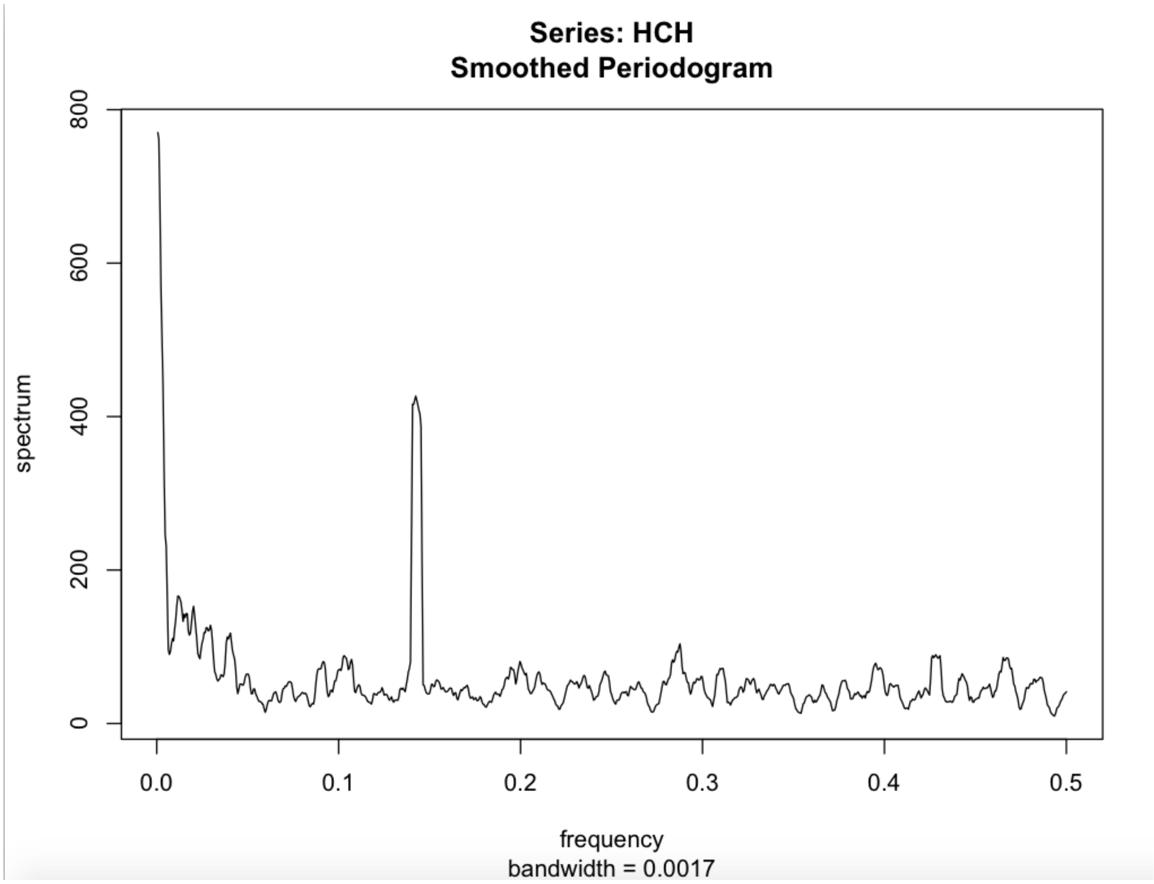


Figure 2. spectra diagram of observed data

From Figure 1, we can see there is periodic signal, and a slightly increasing trend in every year. From Figure 2, the peak is at frequency 0.14, we can infer there is a weekly period. Therefore, we define week as a categories variable to smooth data. And for days in every year, we cannot ensure whether there is linear relation, so we define day is smooth function. For the whole data, we use GAM to smooth and remove periodicity. Then we fit GAM for observed data. The function is $\log(\lambda) \sim \alpha + \beta * \mathrm{t} + \gamma * \mathrm{week} + \mathrm{s}(\mathrm{day})$.

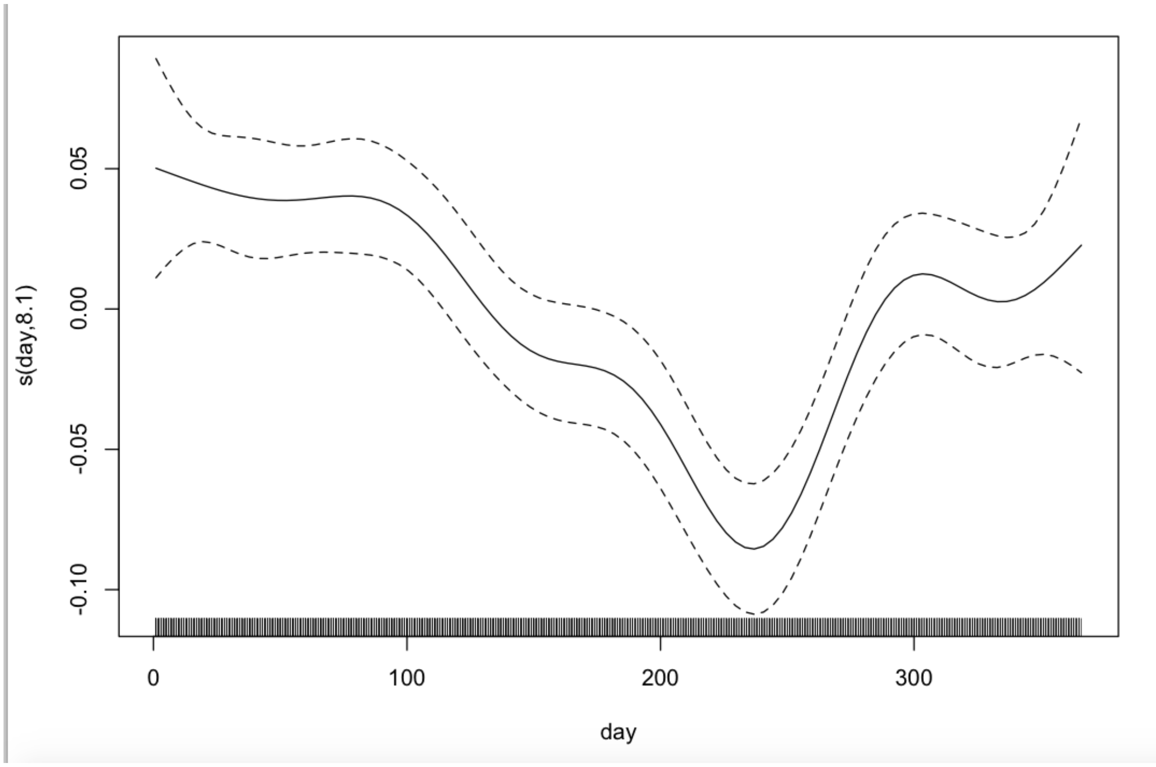


Figure 3. Plot of the smooth components(day) of the fitted GAM.

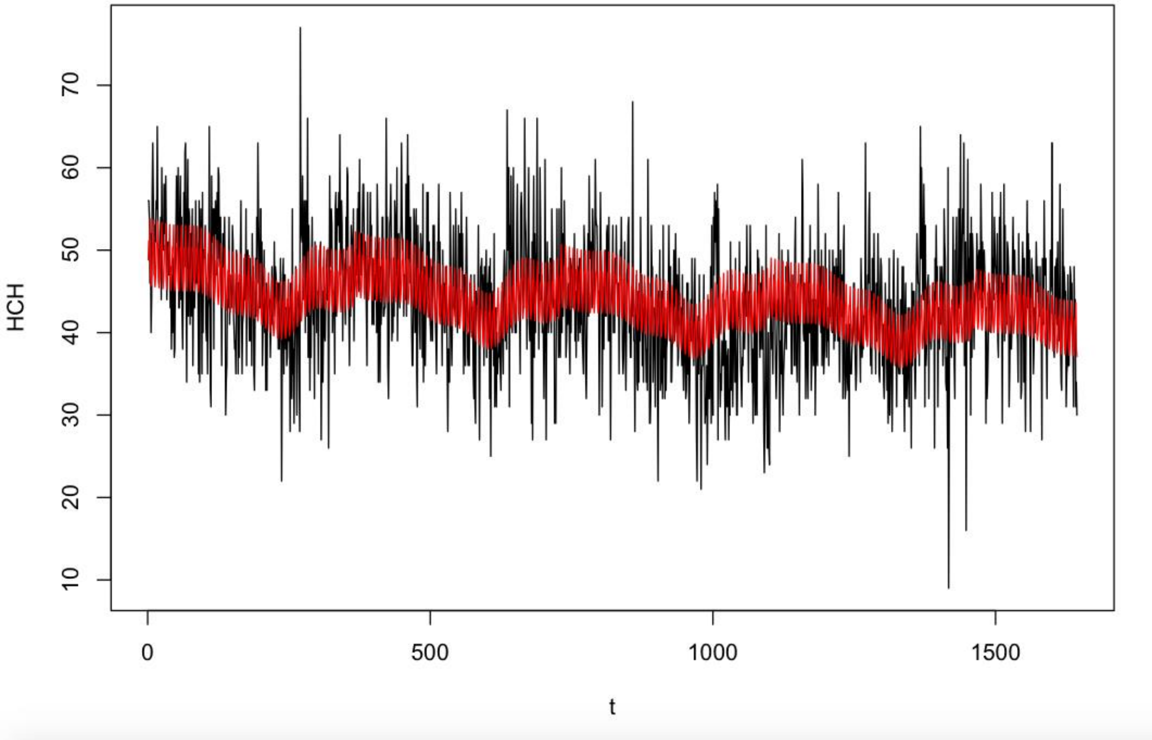
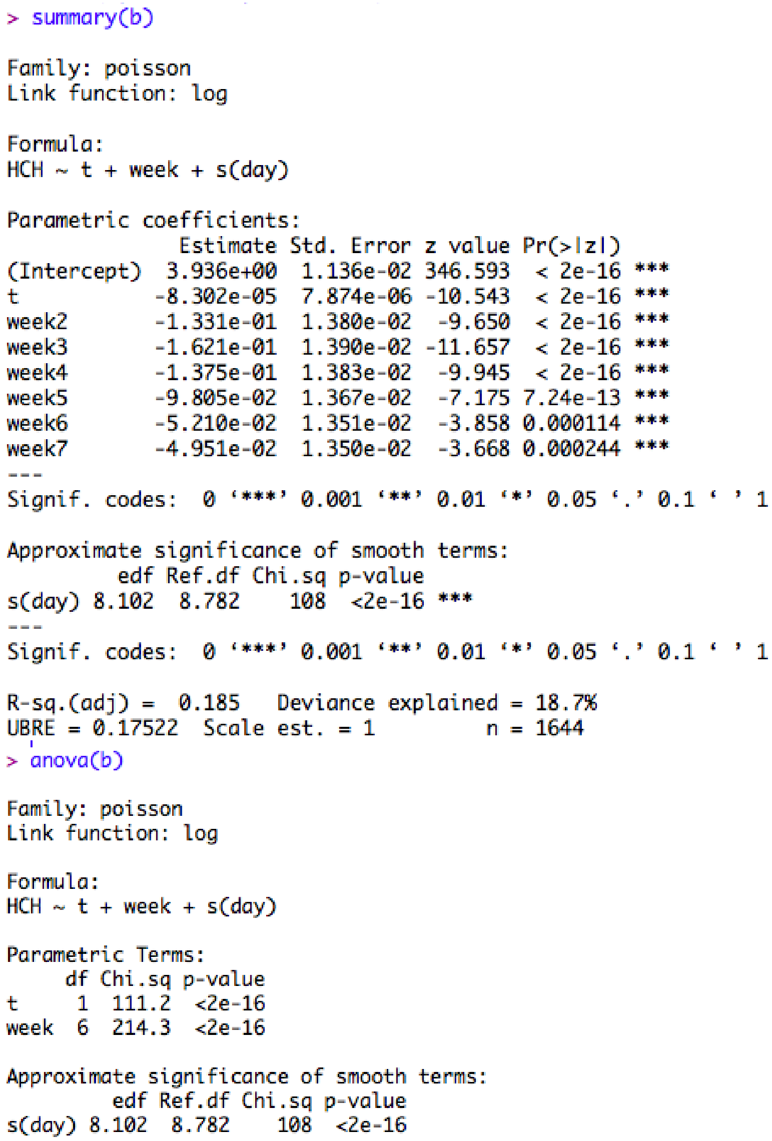


Figure 4. Plot of observed data (black line) versus predicted data (red line).

From Figure 3 and 4, we can see the fitted data versus the original data and check the smooth components behaviors. In Figure 3, it shows the pattern in the year.



```
> summary(b)

Family: poisson
Link function: log

Formula:
HCH ~ t + week + s(day)

Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.936e+00  1.136e-02 346.593  < 2e-16 ***
t           -8.302e-05  7.874e-06 -10.543  < 2e-16 ***
week2       -1.331e-01  1.380e-02  -9.650  < 2e-16 ***
week3       -1.621e-01  1.390e-02 -11.657  < 2e-16 ***
week4       -1.375e-01  1.383e-02  -9.945  < 2e-16 ***
week5       -9.805e-02  1.367e-02  -7.175 7.24e-13 ***
week6       -5.210e-02  1.351e-02  -3.858 0.000114 ***
week7       -4.951e-02  1.350e-02  -3.668 0.000244 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(day) 8.102  8.782    108  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.185   Deviance explained = 18.7%
UBRE = 0.17522  Scale est. = 1         n = 1644
> anova(b)

Family: poisson
Link function: log

Formula:
HCH ~ t + week + s(day)

Parametric Terms:
     df Chi.sq p-value
t     1  111.2  <2e-16
week  6  214.3  <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(day) 8.102  8.782    108  <2e-16
```

From ANOVA, there are all significant for variables.

Then we predict the GAM for the next year. Firstly, we consider the fitted values prediction.

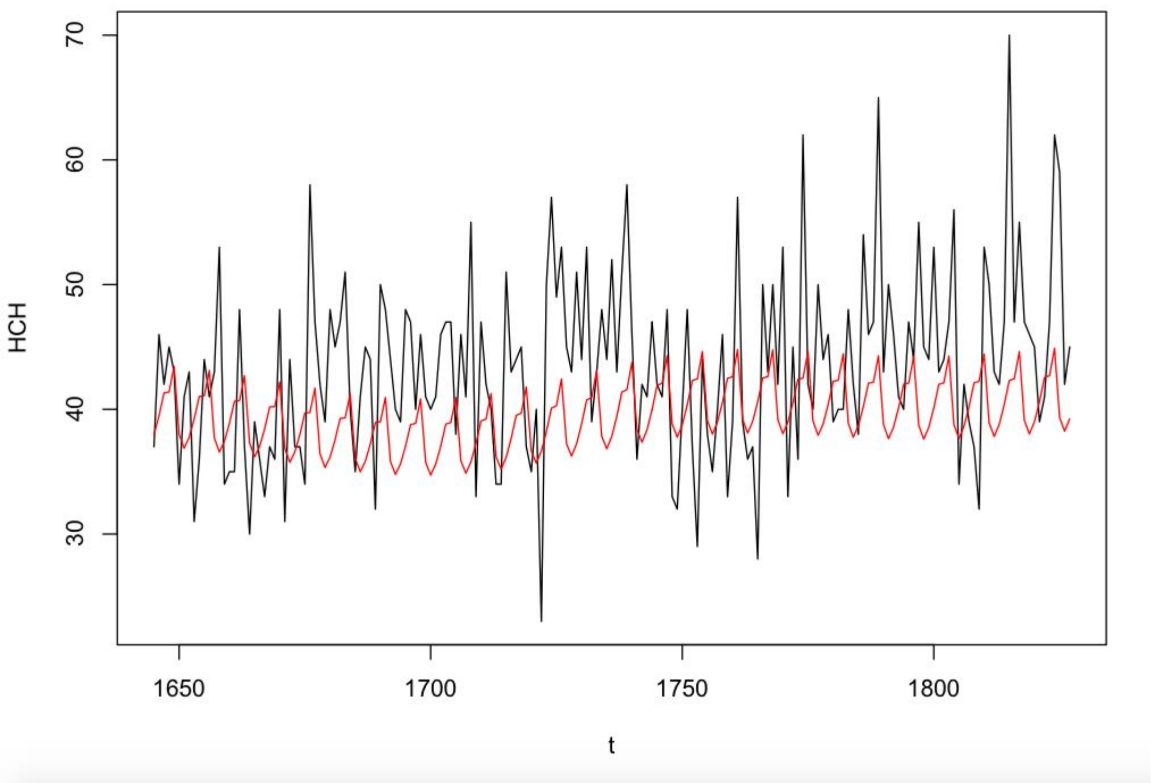
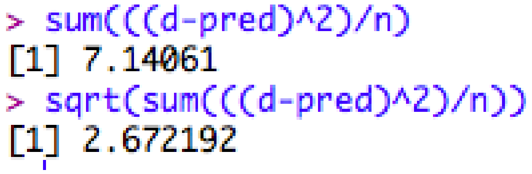


Figure 5. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 7.14061
> sqrt(sum(((d-pred)^2)/n))
[1] 2.672192
```

In the case, the MSPE is 7.14061, and the SSPE is 2.672192

Secondly, we consider the residuals prediction. From Figure 6, we can see the residuals is approximately normal distribution.

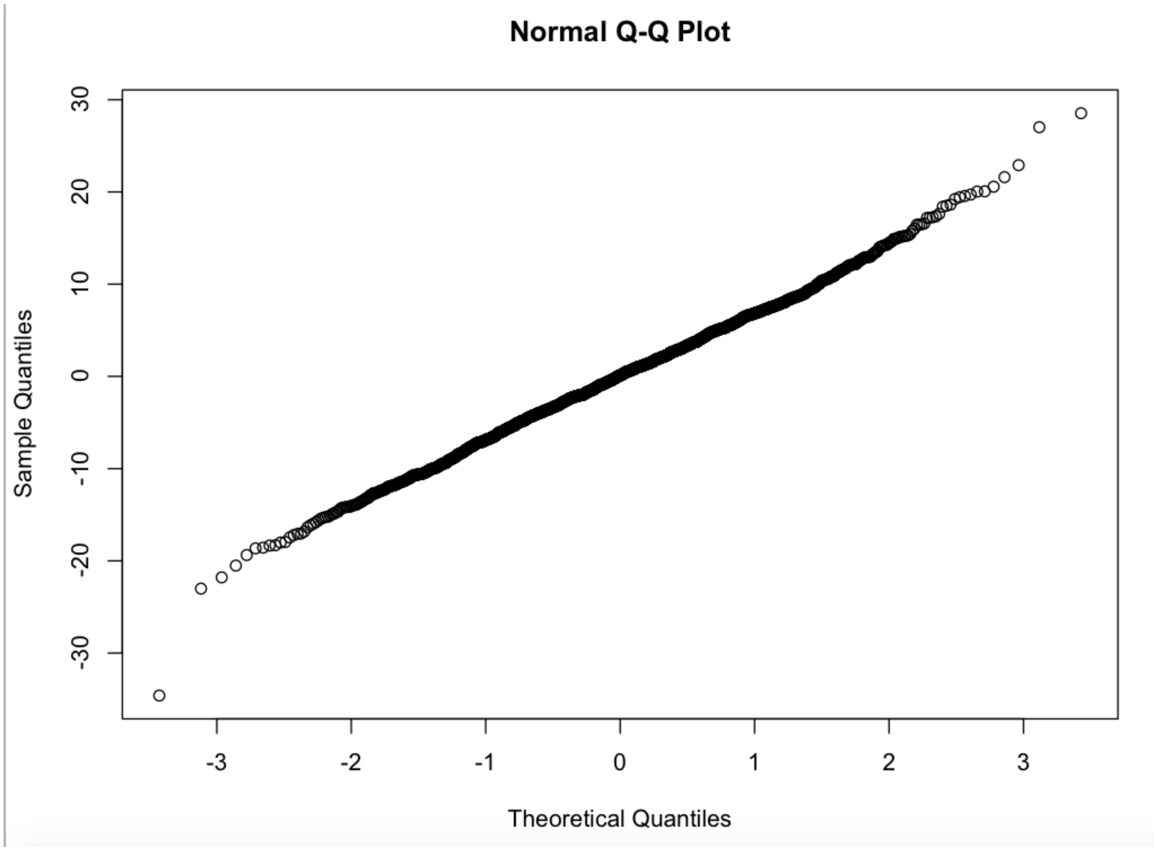
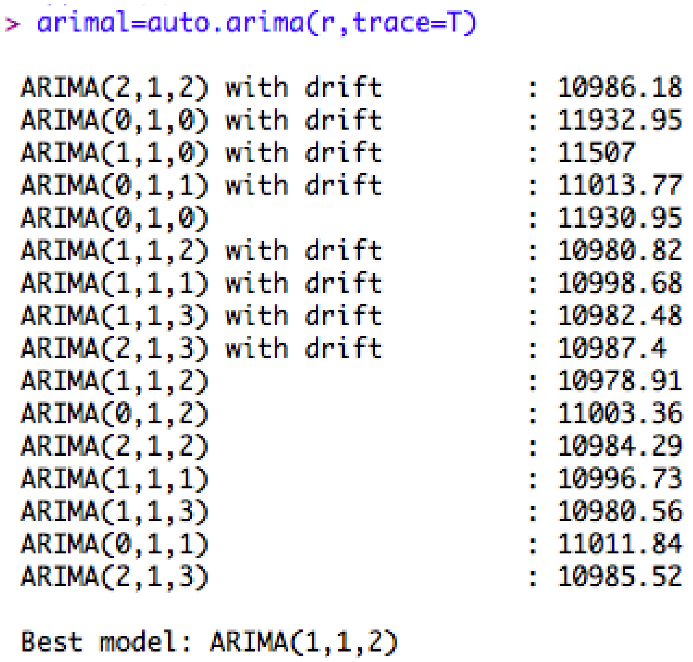


Figure 6. Residuals QQ plot.

Then we use ARMA to fit the residuals. In this case, we choose ARIMA(1,1,2).
The ARIMA(1,1,2) is $X_t - \alpha_1 X_{t-1} = \varepsilon_t + \theta_1 X_{t-1} + \theta_2 X_{t-2}$



```
> arimal=auto.arima(r,trace=T)

 ARIMA(2,1,2) with drift         : 10986.18
 ARIMA(0,1,0) with drift         : 11932.95
 ARIMA(1,1,0) with drift         : 11507
 ARIMA(0,1,1) with drift         : 11013.77
 ARIMA(0,1,0)                    : 11930.95
 ARIMA(1,1,2) with drift         : 10980.82
 ARIMA(1,1,1) with drift         : 10998.68
 ARIMA(1,1,3) with drift         : 10982.48
 ARIMA(2,1,3) with drift         : 10987.4
 ARIMA(1,1,2)                    : 10978.91
 ARIMA(0,1,2)                    : 11003.36
 ARIMA(2,1,2)                    : 10984.29
 ARIMA(1,1,1)                    : 10996.73
 ARIMA(1,1,3)                    : 10980.56
 ARIMA(0,1,1)                    : 11011.84
 ARIMA(2,1,3)                    : 10985.52

 Best model: ARIMA(1,1,2)
```

Then we predict everyday’ residual in next year by weekly circulation.

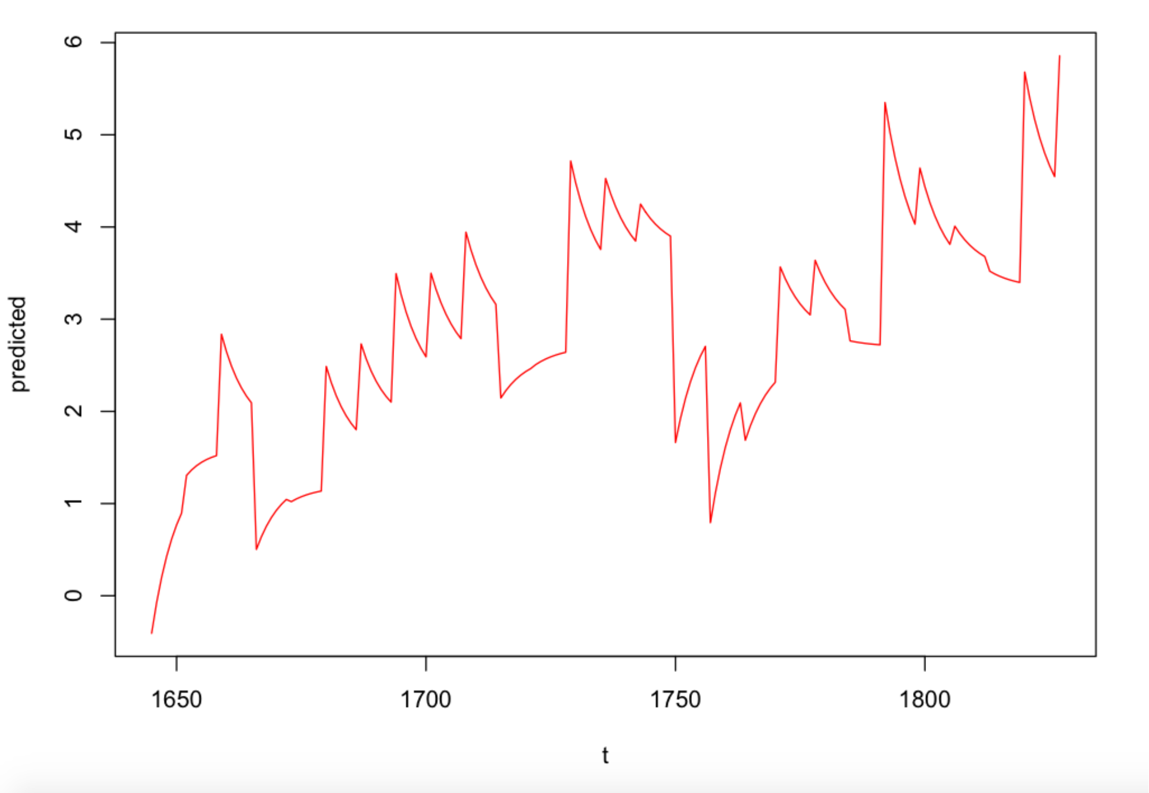


Figure 7. Plot of predicted residuals.

Then we add these two prediction (GAM prediction and residuals prediction), and choose 5% significant level.

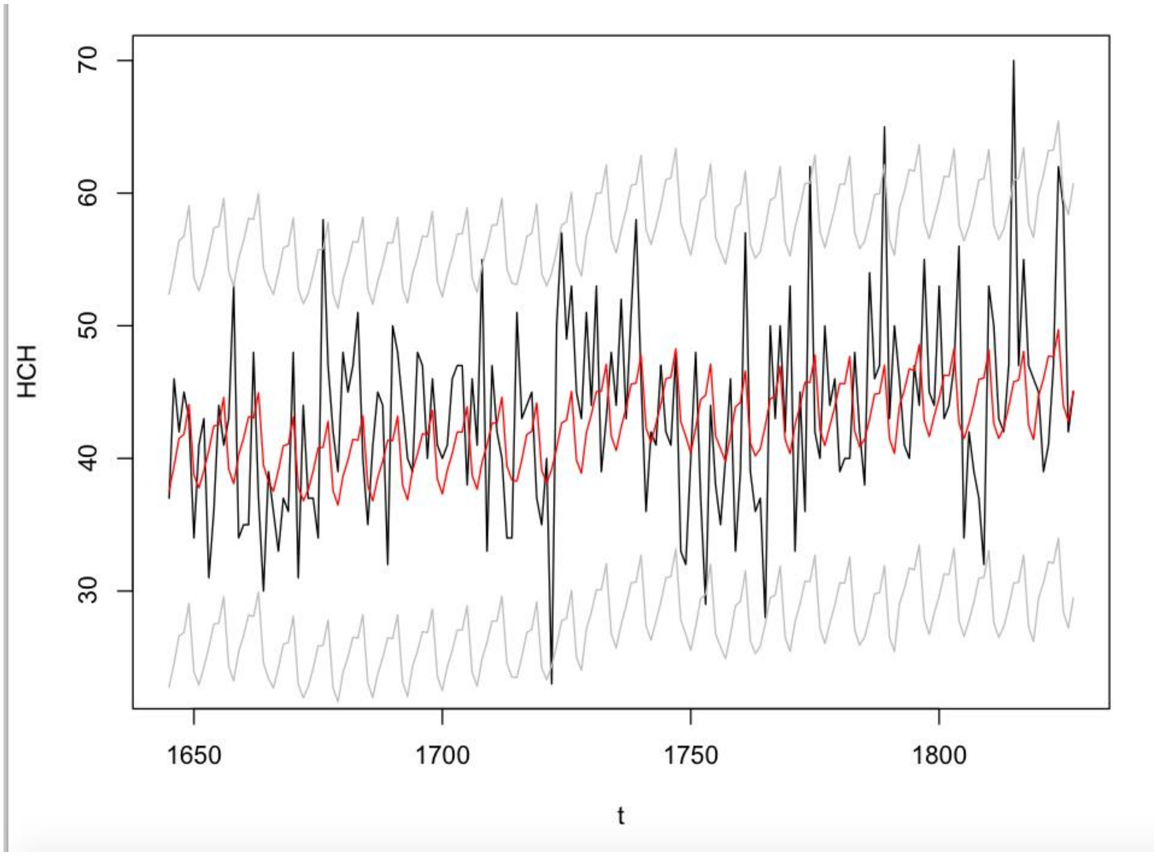
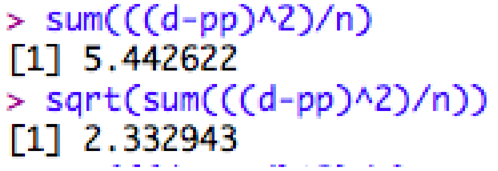


Figure 8. Plot of observed values(black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure8, we can see all the original data are in the confidence interval, and for predicted data, the trend is almost the same as the original data. Then we calculate the mean squared prediction error and standard squared prediction error.



```
> sum(((d-pp)^2)/n)
[1] 5.442622
> sqrt(sum(((d-pp)^2)/n))
[1] 2.332943
```

In the case, the MSPE is 5.442622, and the SSPE is 2.332943, which is less than prediction only by GAM effect.

Next, we try to use log of data to refit this model. The process is as same as above. Hence we get these results.

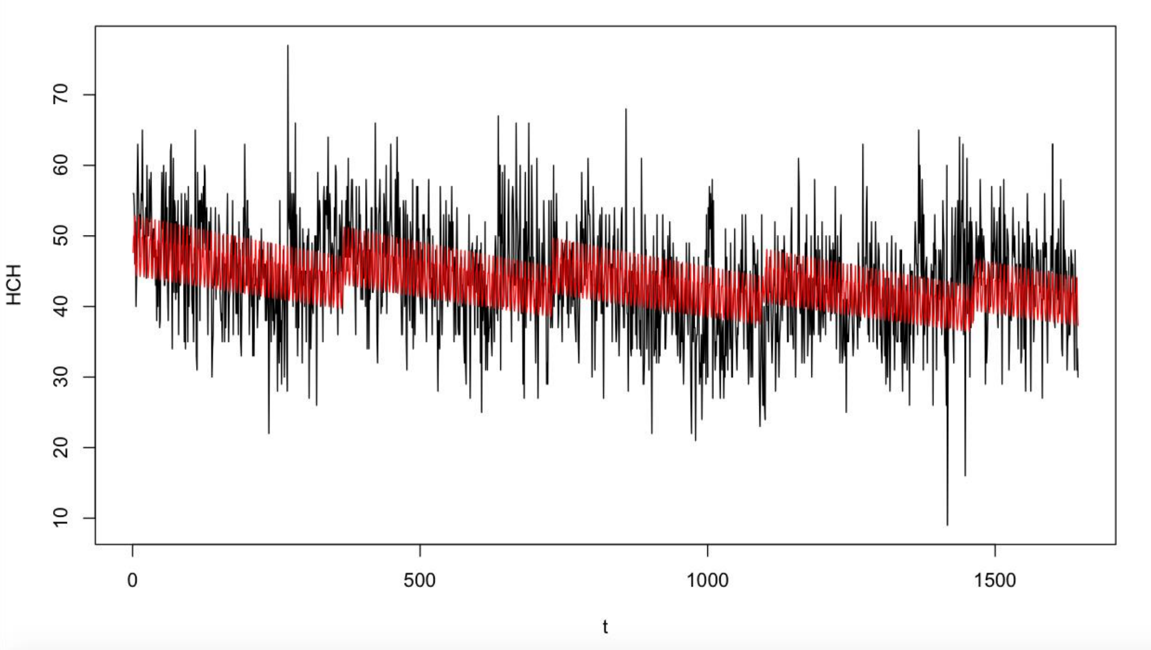


Figure 9. Plot of observed data (black line) versus fitted data (red line).

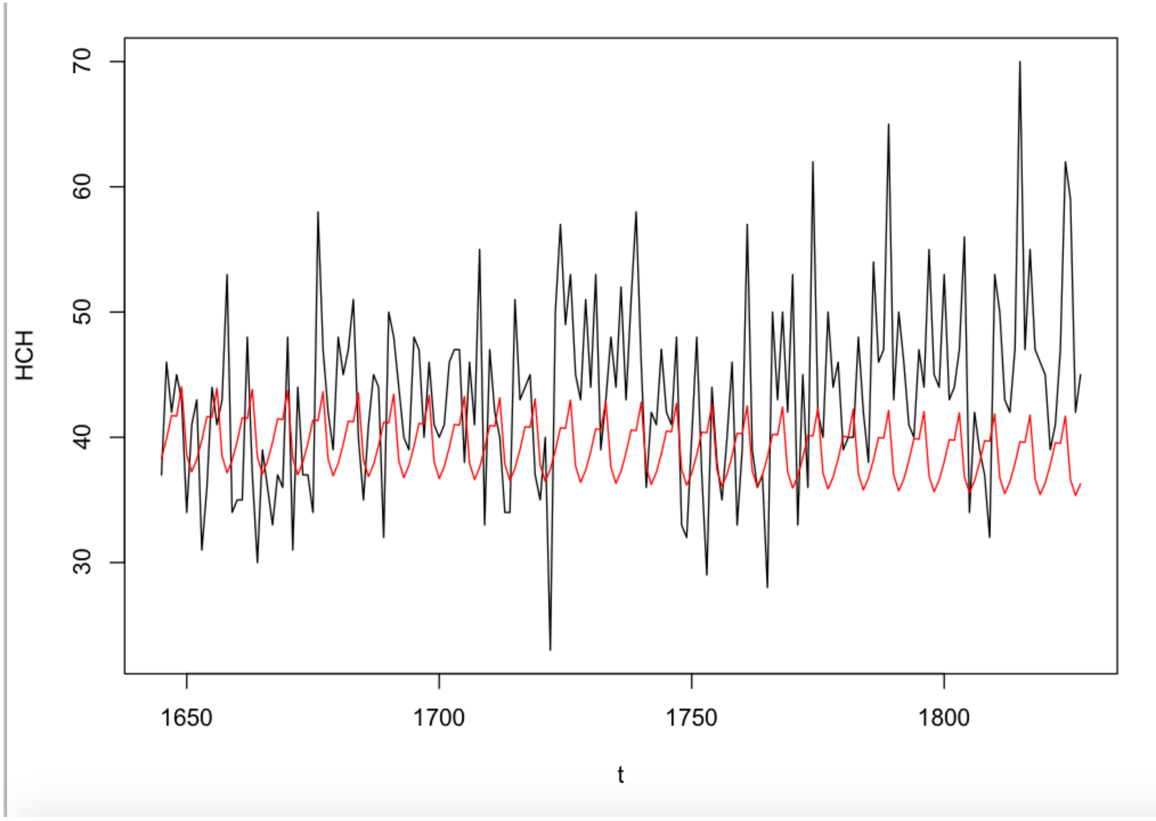
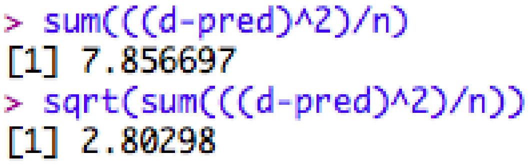


Figure 10. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 7.856697
> sqrt(sum(((d-pred)^2)/n))
[1] 2.80298
```

From Figure 10, the MSPE is 7.856697, and the SSPE is 2.80298.

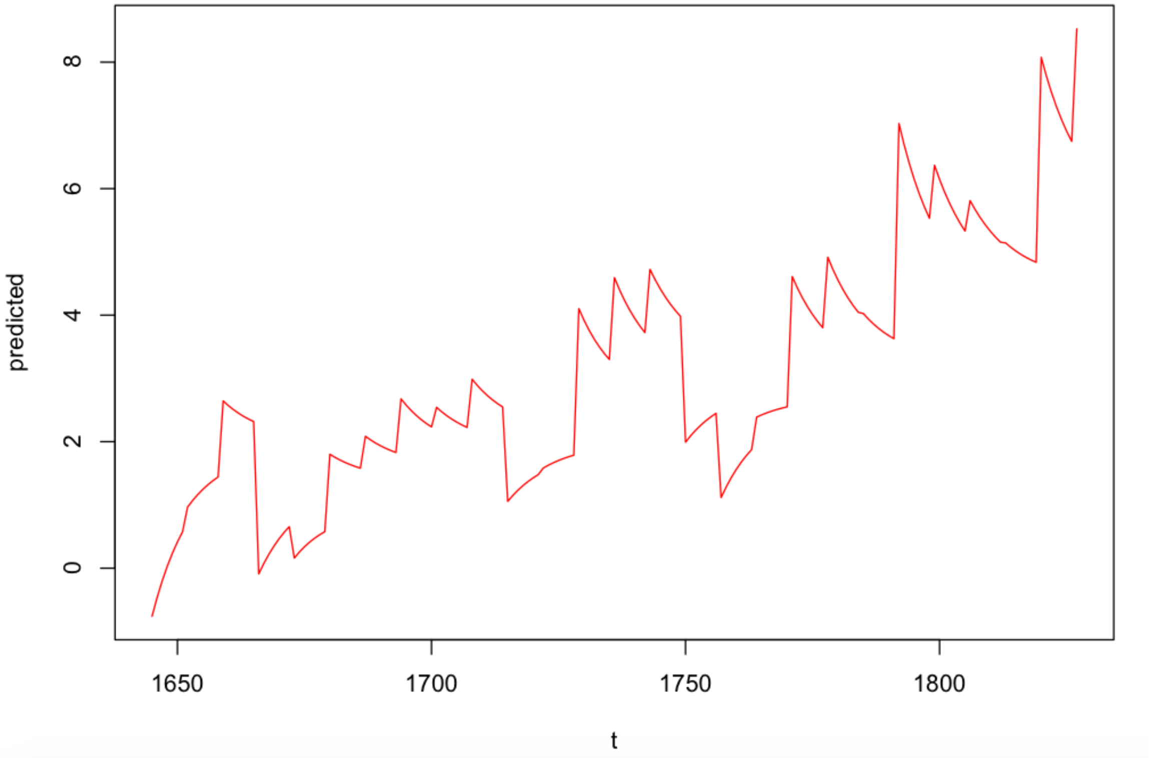


Figure 11. Plot of predicted residuals.

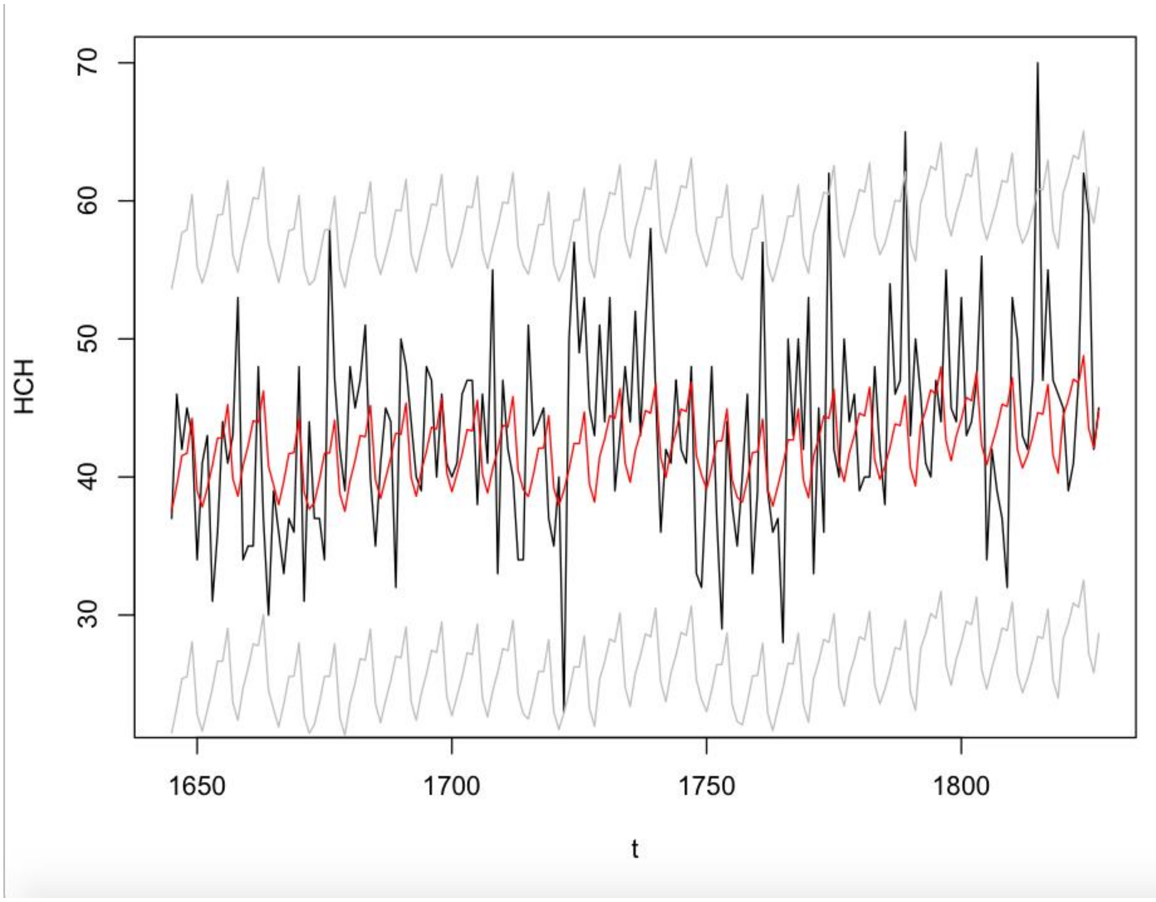
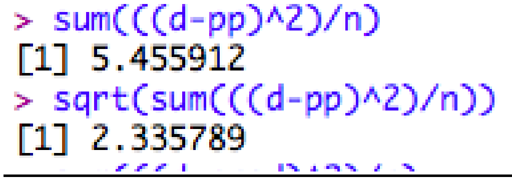


Figure 12. Plot of observed values(black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure 12, the prediction results look roughly same as the Figure 8. We also calculate the calculate the mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 5.455912
> sqrt(sum(((d-pp)^2)/n))
[1] 2.335789
```

In this case, MSPE is 5.455912 and SSPE is 2.335789. We found these are more than prediction by GAM and ARMA effect, which means that improvement cannot be achieved.

For QEII site:

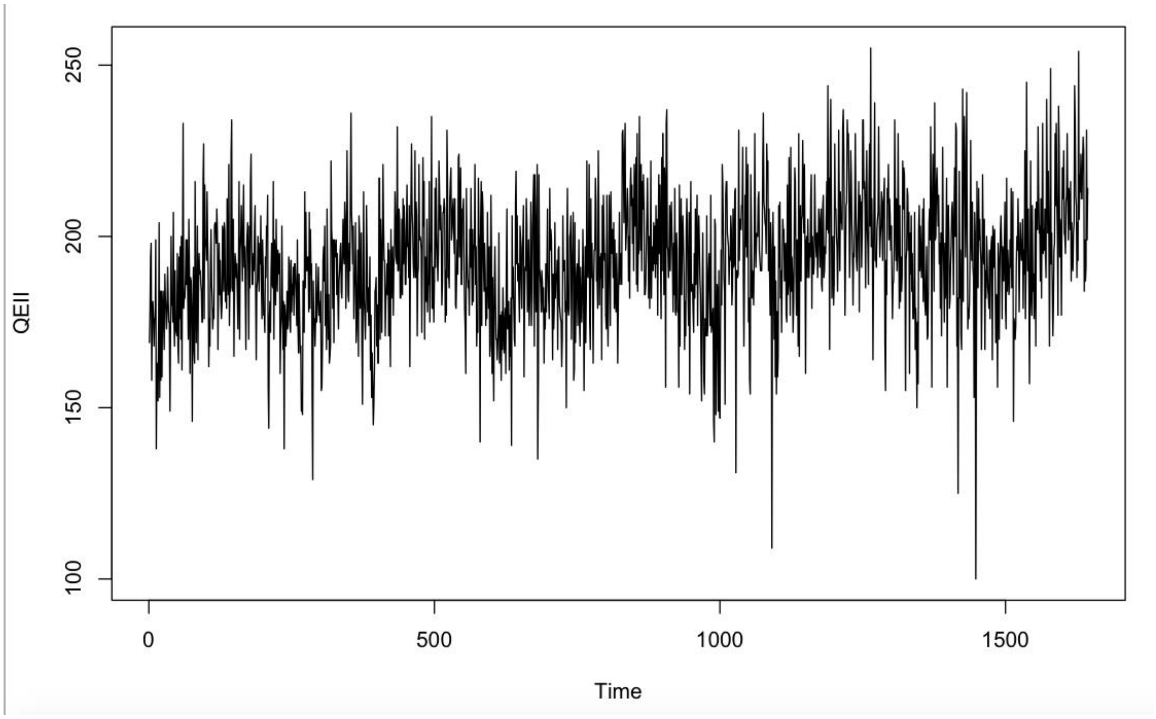


Figure 1. Time series plot of observed data

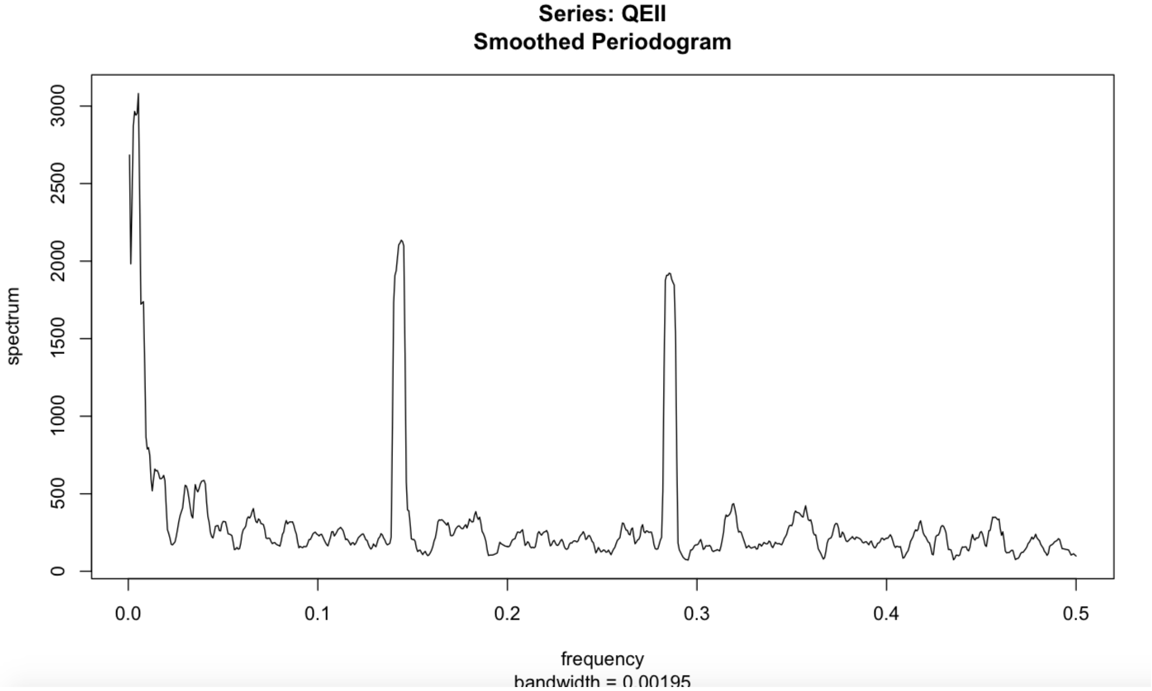


Figure 2. spectra diagram of observed data

From Figure 1, we can see there is periodic signal, and a slightly increasing trend in every year. From Figure 2, the peak is at frequency 0.14, we can infer there is a weekly period. Another peak is at frequency 0.28, which means there is 3.5 days' period (half of a week). In this case, we only consider the weekly period due to avoid overfitting. Therefore, we define week as a categories variable to smooth data. And for days in every year, we cannot ensure whether there is linear relation, so we define day is smooth function. For the whole data, we use GAM to smooth and remove periodicity. Then we fit GAM for observed data. The function is $\log(\lambda) \sim \alpha + \beta * t + \gamma * \text{week} + s(\text{day})$.

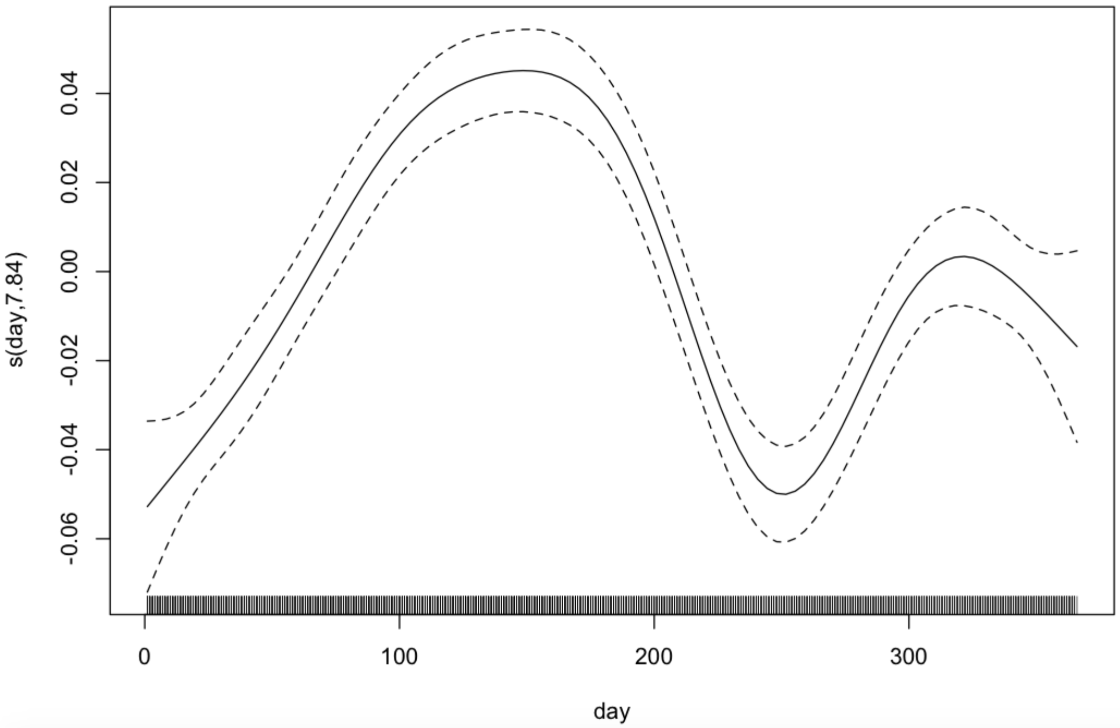


Figure 3. Plot of the smooth components(day) of the fitted GAM.

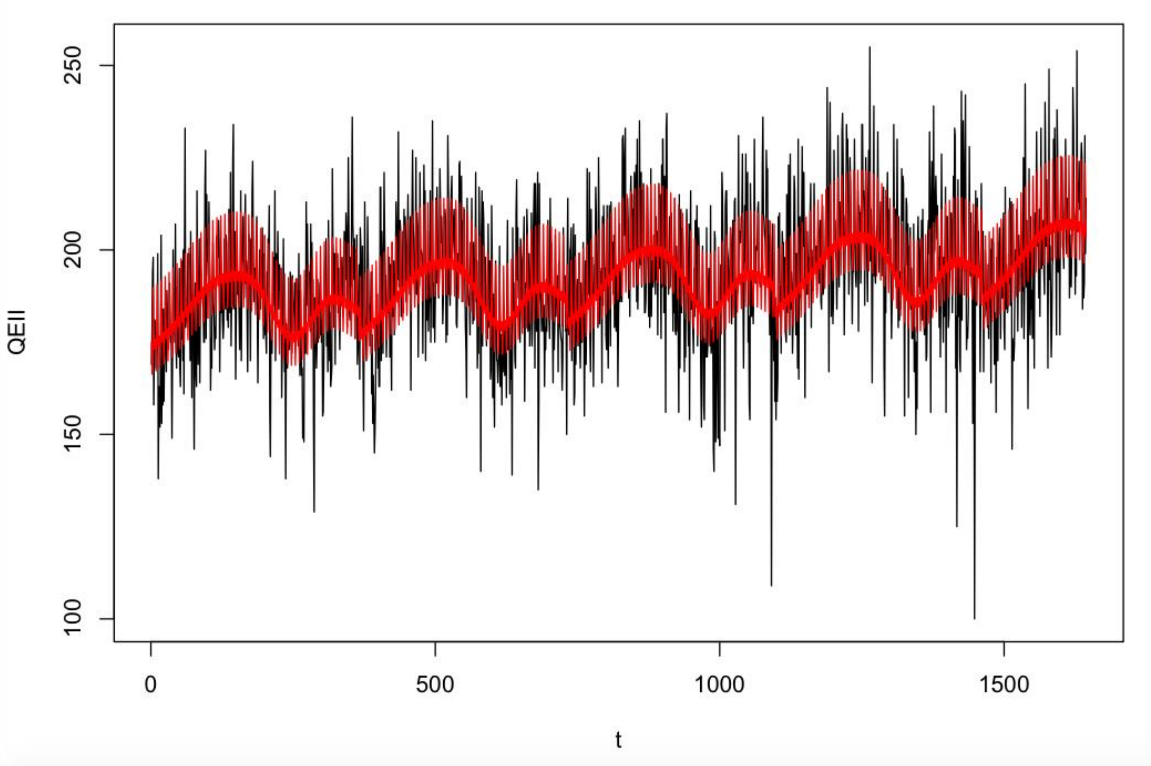
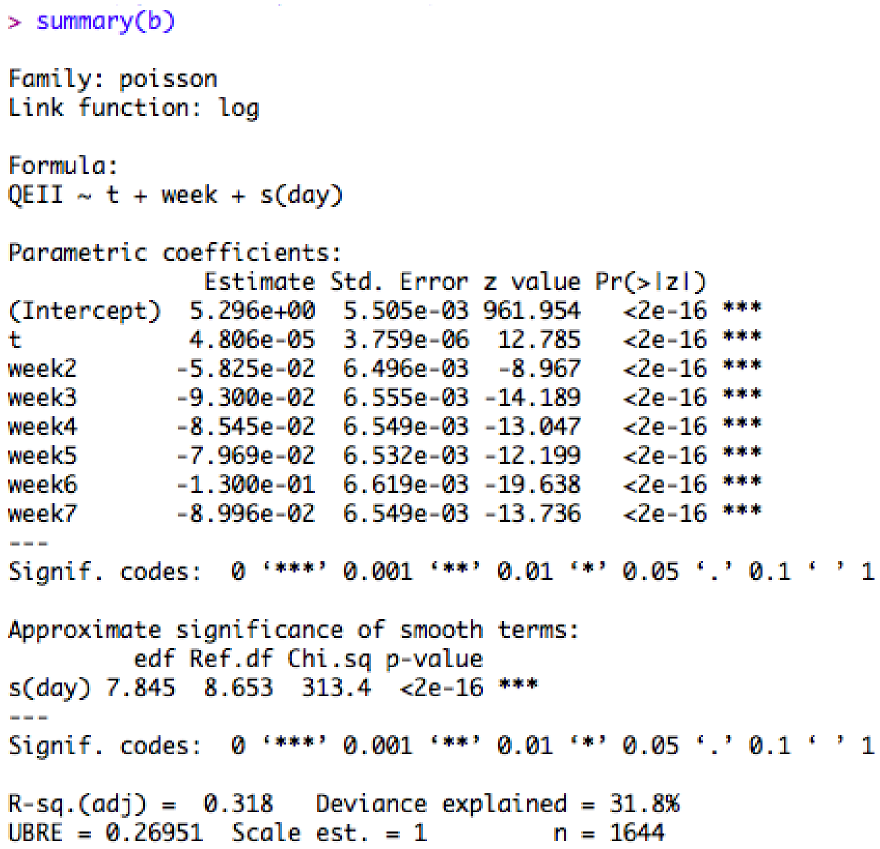
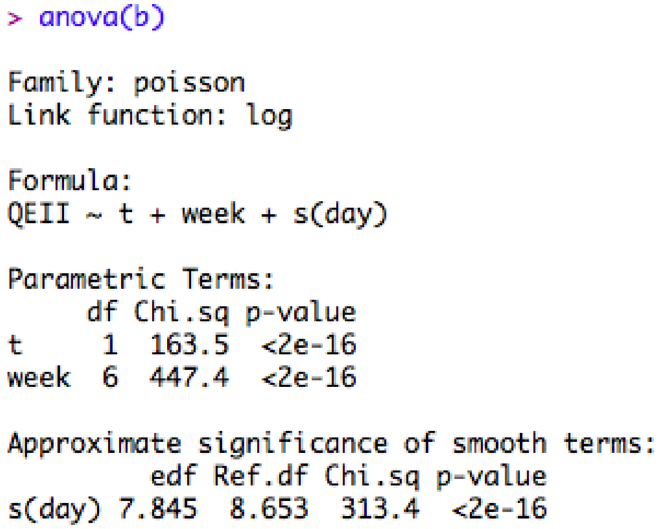


Figure 4. Plot of observed data (black line) versus predicted data (red line).

From Figure 3 and 4, we can see the fitted data versus the original data and check the smooth components behaviors. In Figure 3, it shows the pattern in the year.



```
> summary(b)

Family: poisson
Link function: log

Formula:
QEII ~ t + week + s(day)

Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.296e+00  5.505e-03 961.954   <2e-16 ***
t            4.806e-05  3.759e-06  12.785   <2e-16 ***
week2       -5.825e-02  6.496e-03  -8.967   <2e-16 ***
week3       -9.300e-02  6.555e-03 -14.189   <2e-16 ***
week4       -8.545e-02  6.549e-03 -13.047   <2e-16 ***
week5       -7.969e-02  6.532e-03 -12.199   <2e-16 ***
week6       -1.300e-01  6.619e-03 -19.638   <2e-16 ***
week7       -8.996e-02  6.549e-03 -13.736   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(day) 7.845  8.653  313.4  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.318   Deviance explained = 31.8%
UBRE = 0.26951  Scale est. = 1         n = 1644
```



```
> anova(b)

Family: poisson
Link function: log

Formula:
QEII ~ t + week + s(day)

Parametric Terms:
     df Chi.sq p-value
t     1  163.5  <2e-16
week  6  447.4  <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(day) 7.845  8.653  313.4  <2e-16
```

From ANOVA, there are all significant for variables.

Then we predict the GAM for the next year. Firstly, we consider the fitted values prediction for GAM.

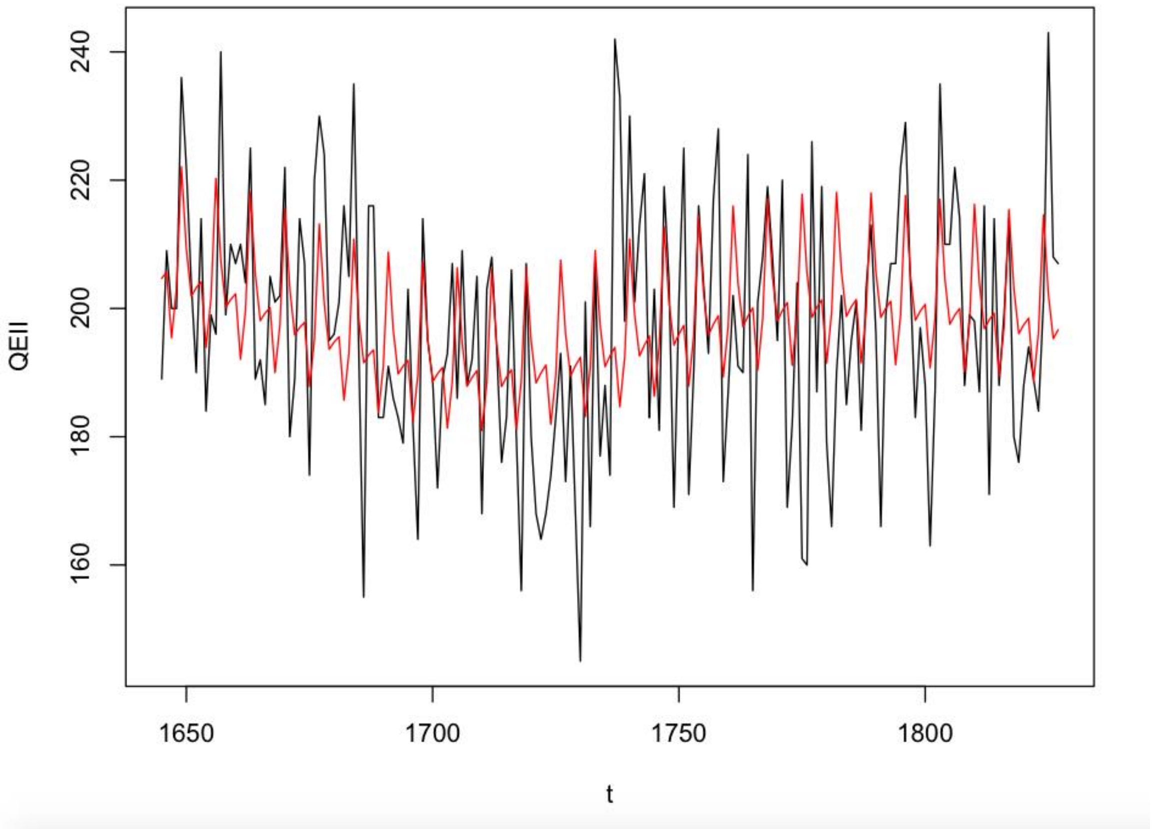
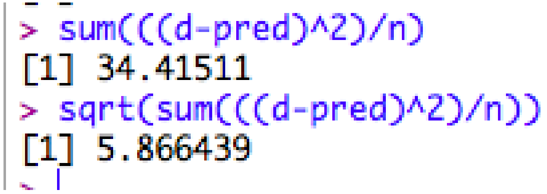


Figure 5. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 34.41511
> sqrt(sum(((d-pred)^2)/n))
[1] 5.866439
```

In the case, the MSPE is 34.41511, and the SSPE is 5.866439.

Secondly, we consider the residuals prediction. From Figure 6, we can see the residuals is approximately normal distribution.

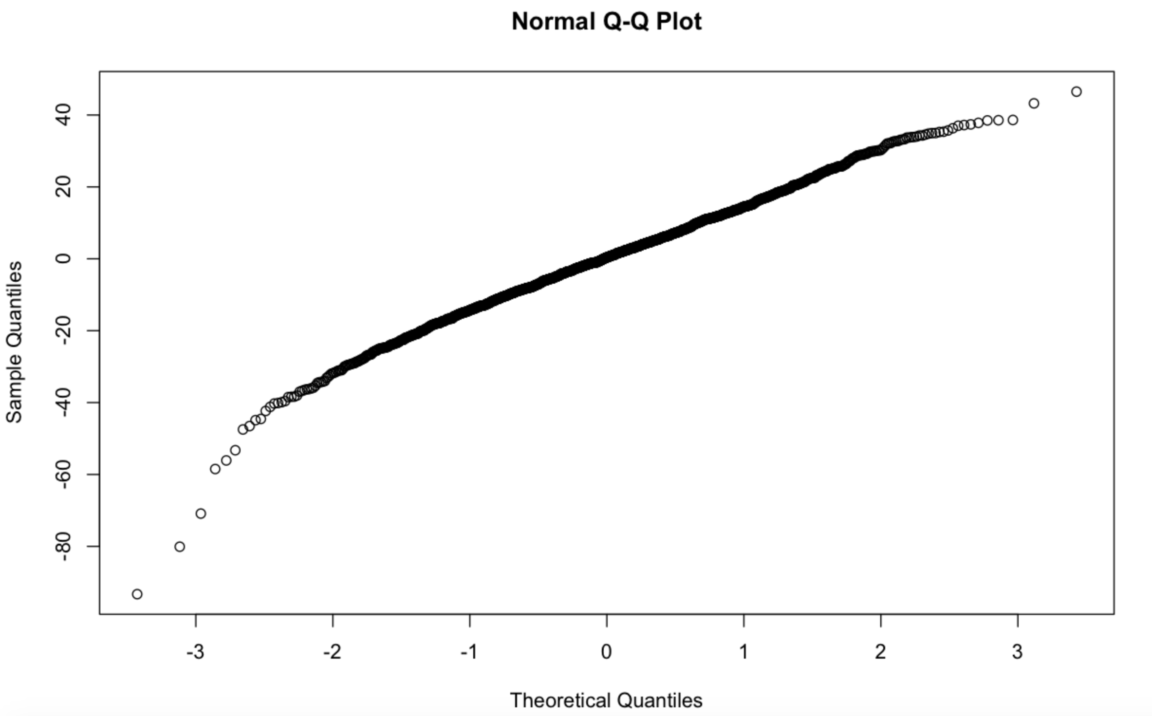
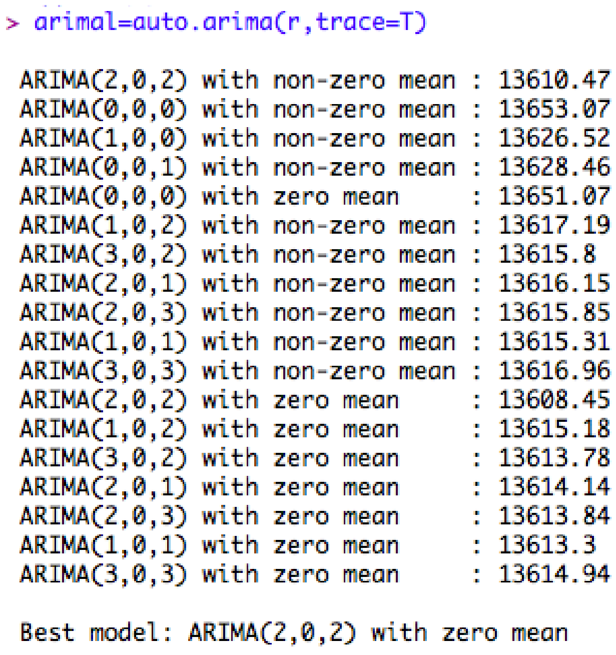


Figure 6. Residuals QQ plot.

Then we use ARMA to fit the residuals. In this case, we choose ARIMA(2,0,2).
The ARIMA(2,0,2) is $X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} = \varepsilon_t + \theta_1 X_{t-1} + \theta_2 X_{t-2}$



```
> arima1=auto.arima(r,trace=T)

 ARIMA(2,0,2) with non-zero mean : 13610.47
 ARIMA(0,0,0) with non-zero mean : 13653.07
 ARIMA(1,0,0) with non-zero mean : 13626.52
 ARIMA(0,0,1) with non-zero mean : 13628.46
 ARIMA(0,0,0) with zero mean     : 13651.07
 ARIMA(1,0,2) with non-zero mean : 13617.19
 ARIMA(3,0,2) with non-zero mean : 13615.8
 ARIMA(2,0,1) with non-zero mean : 13616.15
 ARIMA(2,0,3) with non-zero mean : 13615.85
 ARIMA(1,0,1) with non-zero mean : 13615.31
 ARIMA(3,0,3) with non-zero mean : 13616.96
 ARIMA(2,0,2) with zero mean     : 13608.45
 ARIMA(1,0,2) with zero mean     : 13615.18
 ARIMA(3,0,2) with zero mean     : 13613.78
 ARIMA(2,0,1) with zero mean     : 13614.14
 ARIMA(2,0,3) with zero mean     : 13613.84
 ARIMA(1,0,1) with zero mean     : 13613.3
 ARIMA(3,0,3) with zero mean     : 13614.94

 Best model: ARIMA(2,0,2) with zero mean
```

Then we predict everyday' residual in next year by weekly circulation.

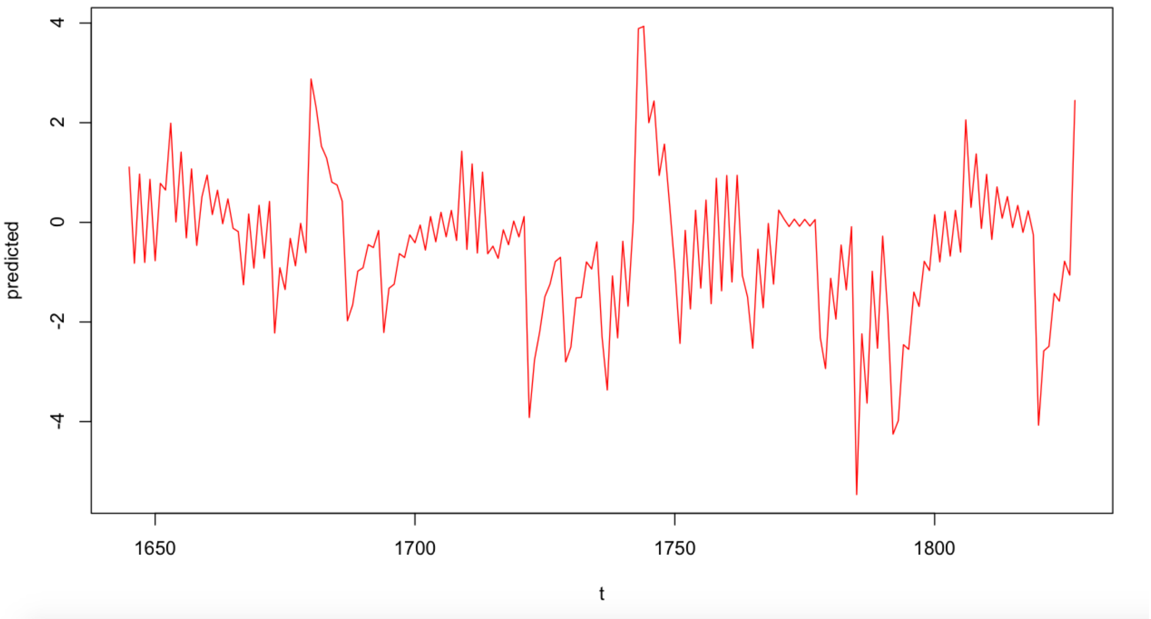


Figure 7. Plot of predicted residuals.

Then we add these two prediction (GAM prediction and residuals prediction), and choose 5% significant level.

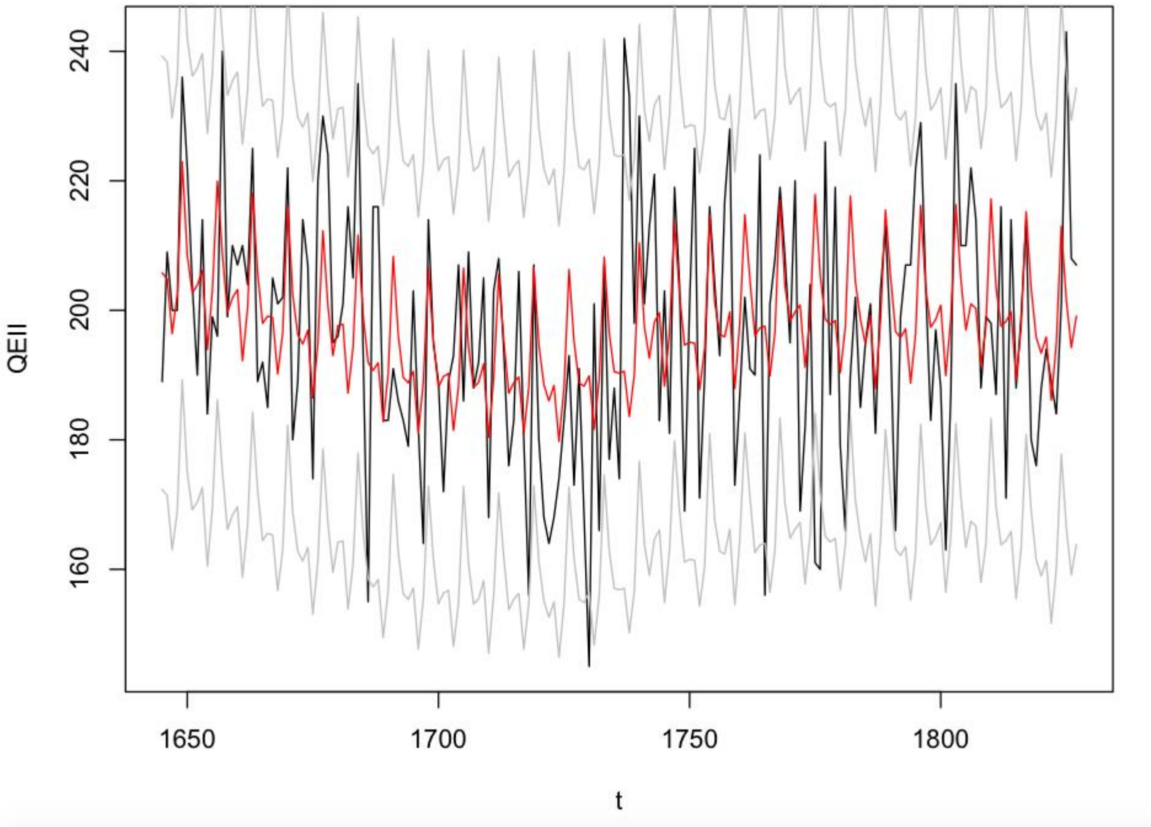
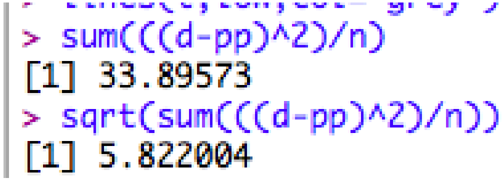


Figure 8. Plot of observed values (black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure8, we can see all the original data are in the confidence interval, and for predicted data, the trend is almost the same as the original data. Then we calculate the mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 33.89573
> sqrt(sum(((d-pp)^2)/n))
[1] 5.822004
```

In the case, the MSPE 33.89573, and the SSPE is 5.822004.

Next, we try to use log of data to refit this model. The process is as same as above. Hence we get these results.

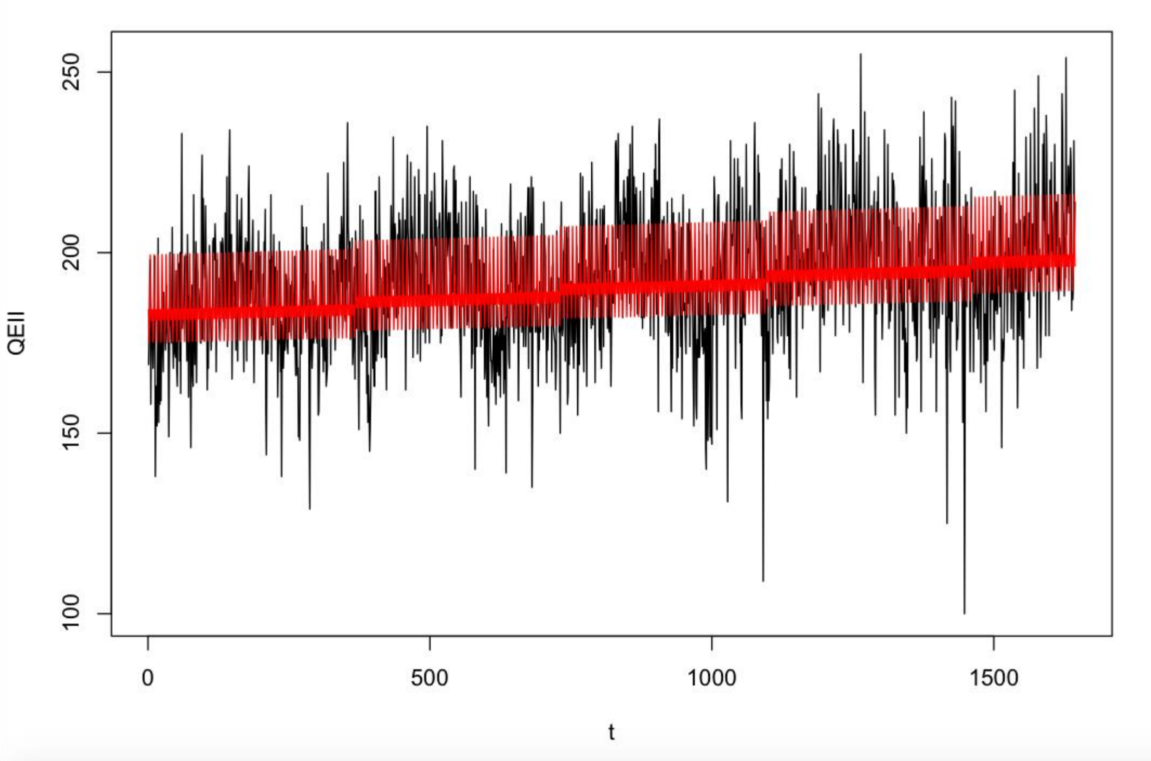


Figure 9. Plot of observed data (black line) versus fitted data (red line).

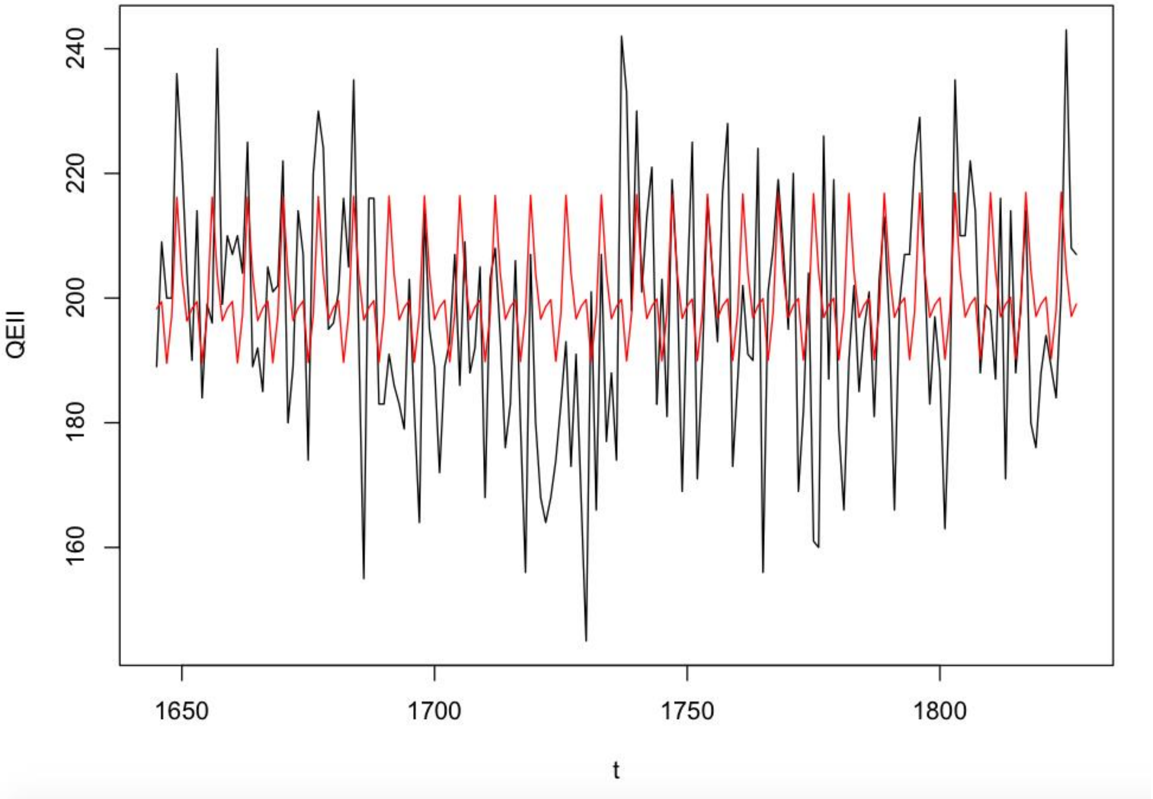
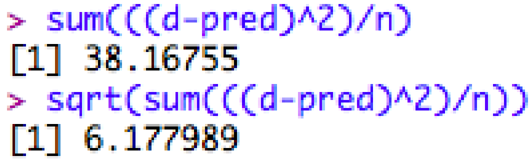


Figure 10. Prediction plot of observed data (black line) versus predicted data (red line).

If we only consider GAM effect prediction, we can calculate mean squared prediction error and standard squared prediction error.



```
> sum(((d-pred)^2)/n)
[1] 38.16755
> sqrt(sum(((d-pred)^2)/n))
[1] 6.177989
```

From Figure 10, the MSPE is 38.16755, and the SSPE is 6.177989.

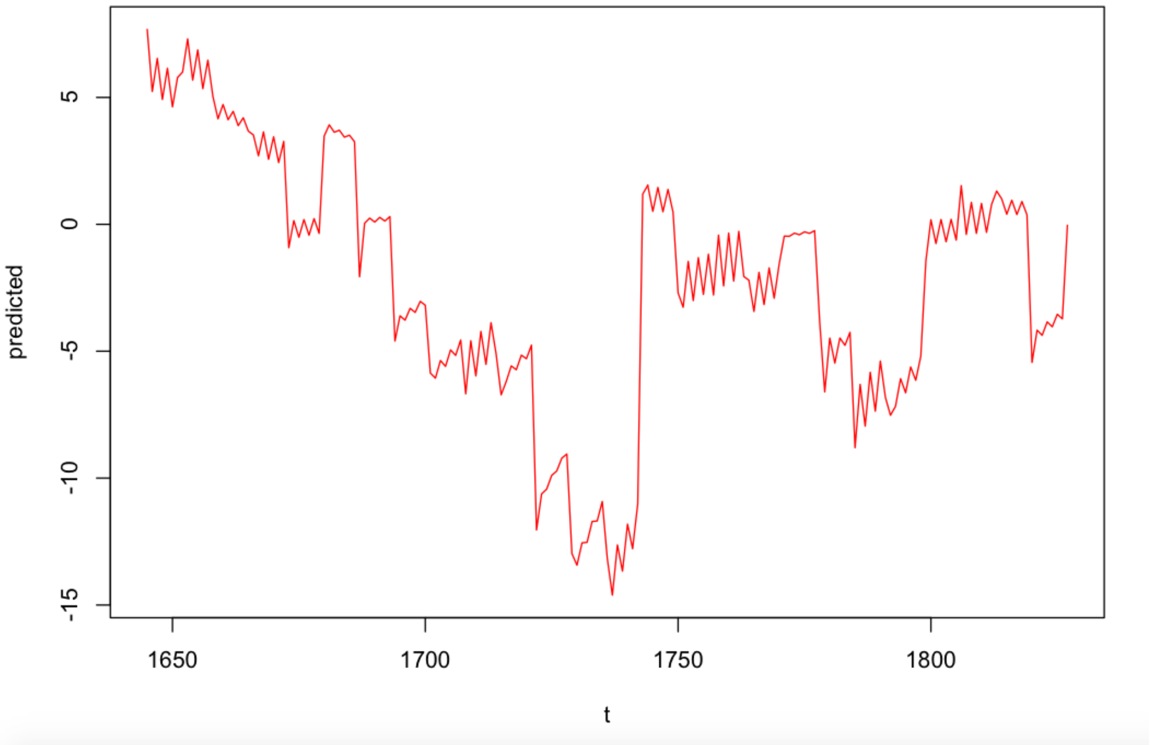


Figure 11. Plot of predicted residuals.

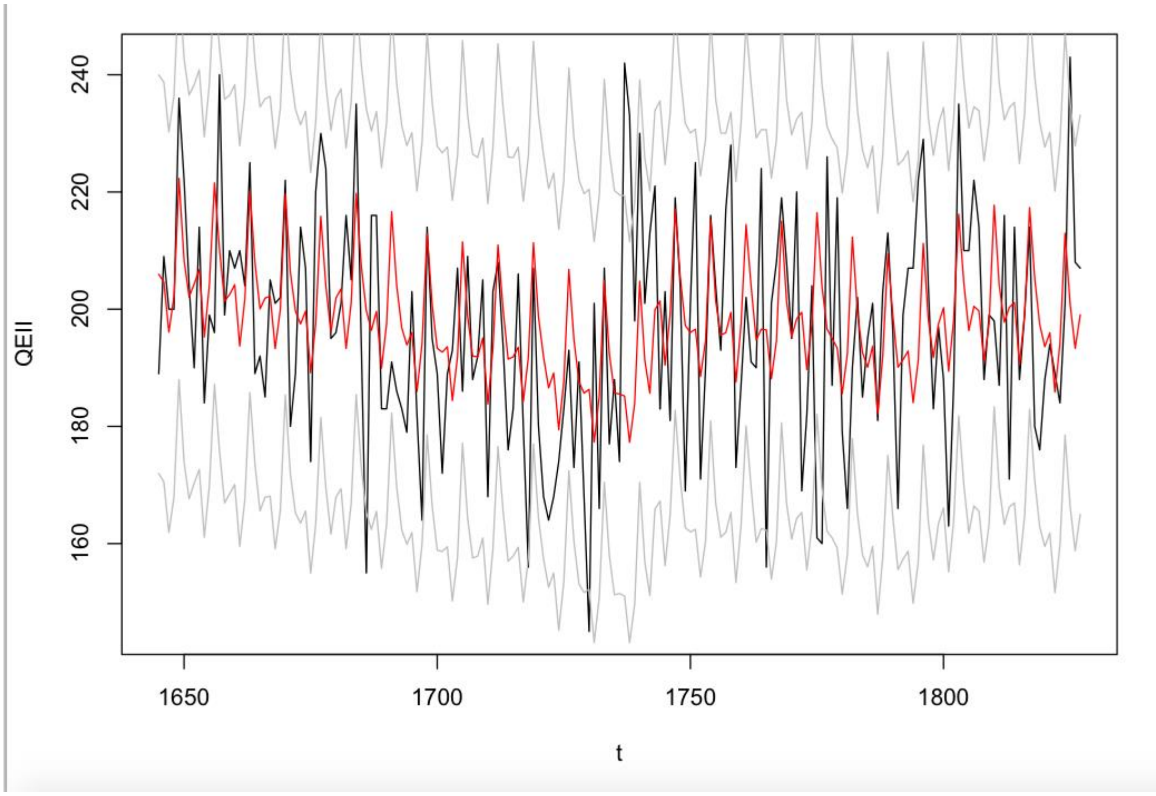
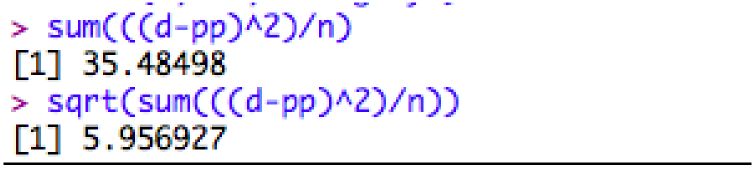


Figure 12. Plot of observed values(black line) and predicted values(red line), and 95% confidence interval(grey line).

From Figure 12, the prediction results look roughly same as the Figure 8. We also calculate the calculate the mean squared prediction errors and standard squared prediction errors.



```
> sum(((d-pp)^2)/n)
[1] 35.48498
> sqrt(sum(((d-pp)^2)/n))
[1] 5.956927
```

In this case, MSPE 35.48498, and SSPE is 5.956927. We found these are more than prediction by GAM and ARMA effect, which means that improvement cannot be achieved.

Conclusion

From the results, we can get this table.

Prediction error	CCHC	DGH	HCH	QEII
SSPE only by GAM	3.630821	3.654963	2.672192	5.866439
SSPE by GAM and ARMA	3.608406	3.62875	2.332943	5.822004
SSPE by log of data	3.66111	3.612282	2.335789	5.956927

From the table, we can find that using GAM and ARMA to predict data gets less prediction error. Only in the DGH site, we use log of data to predict data, improvement can be achieved.

In this project, prediction error of QEII is slightly large, we can find more information about QEII to fit other proper model to make prediction error smaller in the future work.

Appendix

```
hosp=read.csv("/Users/zhangxinyue/Desktop/hospital.csv",header=T)
library(nlme)
library(mgcv)
library(forecast)
cchc=hosp$CCHC[1:1644]
check=spec.pgram(cchc,spans=11,log="no")
plot.ts(cchc,type="l")
n=length(cchc)
t=1:n
week=as.factor(c((5:7),rep(c(1:7),234),(1:3)))
day=c((1:366),rep(c(1:365),3),(1:183))
a=data.frame(cchc,t,week,day)
b=gam(cchc~t+week+s(day),family=poisson(link="log"),method="GCV.C
 p",data=a)
plot(b,pages=1,seWithMean=TRUE)
plot(t, a$cchc,ylab="CCHC",type="l")
lines(t, b$fitted, col="red")
summary(b)
ANOVA(b)
t=1645:1827
week=as.factor(c((4:7),rep(c(1:7),25),(1:4)))
day=c((184:366))
new=data.frame(t,week,day)
p=predict.gam(b,new,type="response",se=TRUE)
pred=p$fit
se1=p$se
d=hosp$CCHC[1645:1827]
plot(t, d,ylab="CCHC",type="l")
lines(t, pred, col="red")
pred1=matrix(pred,nrow=183,ncol=1)
se1=matrix(se1,nrow=183,ncol=1)
r=residuals(b,type="response")
qqnorm(r)
arima1=auto.arima(r,trace=T)
fit=arima(r,order=c(1,0,2))
predicted=matrix(nrow=189,ncol=1)
i=1
while(i<=183){
 pred2=predict(fit,n.ahead=7)
 predicted[i]=pred2$pred[1]
 predicted[i+1]=pred2$pred[2]
 predicted[i+2]=pred2$pred[3]
 predicted[i+3]=pred2$pred[4]
 predicted[i+4]=pred2$pred[5]
 predicted[i+5]=pred2$pred[6]
 predicted[i+6]=pred2$pred[7]
 r=c(r,hosp$CCHC[i+1644]-pred[i],hosp$CCHC[i+1645]-
 pred[i+1],hosp$CCHC[i+1646]-pred[i+2],hosp$CCHC[i+1647]-
 pred[i+3],hosp$CCHC[i+1648]-pred[i+4],hosp$CCHC[i+1649]-
 pred[i+5],hosp$CCHC[i+1650]-pred[i+6])
 fit=arima(r,order=c(1,0,2))
 i=i+7
}
predicted=predicted[1:183]
plot(t,predicted,type="l",col="red")
pp=pred1+predicted
r2=residuals(b,type="response")
fit2=arima(r2,order=c(1,0,2))
se2=matrix(nrow=189,ncol=1)
i=1
while(i<=183){
 pred2=predict(fit,n.ahead=7)
 se2[i]=pred2$se[1]
 se2[i+1]=pred2$se[2]
 se2[i+2]=pred2$se[3]
 se2[i+3]=pred2$se[4]
 se2[i+4]=pred2$se[5]
 se2[i+5]=pred2$se[6]
 se2[i+6]=pred2$se[7]
 r2=c(r,hosp$CCHC[i+1644]-pred[i],hosp$CCHC[i+1645]-
pred[i+1],hosp$CCHC[i+1646]-pred[i+2],hosp$CCHC[i+1647]-
pred[i+3],hosp$CCHC[i+1648]-pred[i+4],hosp$CCHC[i+1649]-
pred[i+5],hosp$CCHC[i+1650]-pred[i+6])
 fit2=arima(r,order=c(1,0,2))
 i=i+7
}
se2=se2[1:183]
se=se1+se2
up=pp+1.96*se
low=pp-1.96*se
plot(t, d,type="l",ylab="CCHC")
lines(t,pp,col="red")
lines(t,up,col="grey")
lines(t,low,col="grey")
sum(((d-pp)^2)/n)
sqrt(sum(((d-pp)^2)/n))
sum(((d-pred)^2)/n)
sqrt(sum(((d-pred)^2)/n))


#use log of data #
hosp=read.csv("/Users/zhangxinyue/Desktop/hospital.csv",header=T)
library(nlme)
library(mgcv)
library(forecast)
cchc=hosp$CCHC[1:1644]
n=length(cchc)
t=1:n
week=as.factor(c((5:7),rep(c(1:7),234),(1:3)))
day=c((1:366),rep(c(1:365),3),(1:183))
a=data.frame(cchc,t,week,day)
b=gam(log(cchc)~t+week+s(day),family=poisson(link="log"),method="
GCV.Cp",data=a)
plot(b,pages=1,seWithMean=TRUE)
plot(t, a$cchc,ylab="CCHC",type="l")
lines(t, exp(b$fitted), col="red")
summary(b)
anova(b)
t=1645:1827
week=as.factor(c((4:7),rep(c(1:7),25),(1:4)))
day=c((184:366))
new=data.frame(t,week,day)
p=predict.gam(b,new,type="response",se=TRUE)
pred=exp(p$fit)
se1=exp(p$se)
d=hosp$CCHC[1645:1827]
plot(t, d,ylab="CCHC",type="l")
lines(t, pred, col="red")
pred1=matrix(pred,nrow=183,ncol=1)
se1=matrix(se1,nrow=183,ncol=1)
r=cchc-exp(b$fitted)
qqnorm(r)
arimal=auto.arima(r,trace=T)
fit=arima(r,order=c(1,0,2))
predicted=matrix(nrow=189,ncol=1)
i=1
while(i<=183){
 pred2=predict(fit,n.ahead=7)
 predicted[i]=pred2$pred[1]
 predicted[i+1]=pred2$pred[2]
 predicted[i+2]=pred2$pred[3]
 predicted[i+3]=pred2$pred[4]
 predicted[i+4]=pred2$pred[5]
 predicted[i+5]=pred2$pred[6]
 predicted[i+6]=pred2$pred[7]
 r=c(r,hosp$CCHC[i+1644]-pred[i],hosp$CCHC[i+1645]-
pred[i+1],hosp$CCHC[i+1646]-pred[i+2],hosp$CCHC[i+1647]-
pred[i+3],hosp$CCHC[i+1648]-pred[i+4],hosp$CCHC[i+1649]-
pred[i+5],hosp$CCHC[i+1650]-pred[i+6])
 fit=arima(r,order=c(1,0,2))
 i=i+7
}
predicted=predicted[1:183]
plot(t,predicted,type="l",col="red")
pp=pred1+predicted
r2=cchc-exp(b$fitted)
fit2=arima(r2,order=c(1,0,2))
se2=matrix(nrow=189,ncol=1)
i=1
while(i<=183){
 pred2=predict(fit,n.ahead=7)
 se2[i]=pred2$se[1]
 se2[i+1]=pred2$se[2]
 se2[i+2]=pred2$se[3]
 se2[i+3]=pred2$se[4]
 se2[i+4]=pred2$se[5]
 se2[i+5]=pred2$se[6]
 se2[i+6]=pred2$se[7]
 r2=c(r,hosp$CCHC[i+1644]-pred[i],hosp$CCHC[i+1645]-
pred[i+1],hosp$CCHC[i+1646]-pred[i+2],hosp$CCHC[i+1647]-
pred[i+3],hosp$CCHC[i+1648]-pred[i+4],hosp$CCHC[i+1649]-
pred[i+5],hosp$CCHC[i+1650]-pred[i+6])
 fit2=arima(r,order=c(1,0,2))
 i=i+7
}
se2=se2[1:183]
se=se1+se2
up=pp+1.96*se
low=pp-1.96*se
plot(t, d,type="l",ylab="CCHC")
lines(t,pp,col="red")
lines(t,up,col="grey")
lines(t,low,col="grey")
sum(((d-pp)^2)/n)
sqrt(sum(((d-pp)^2)/n))
sum(((d-pred)^2)/n)
sqrt(sum(((d-pred)^2)/n))
```