equation

Xinyi Hu

4/29/2019

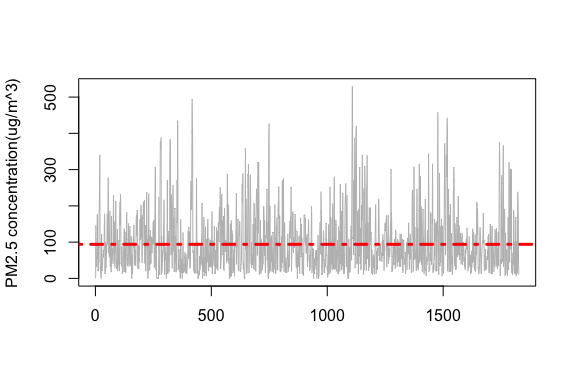
# Time series analysis section

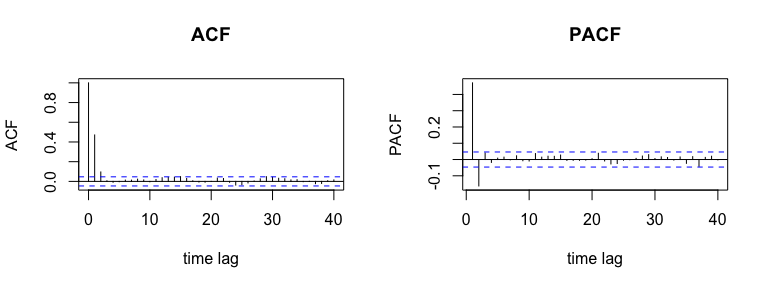
This section is mainly about performing time series analysis on average PM 2.5 daily records and monthly records respectively. Our target in this section is to find out the patterns and features of the PM 2.5 values under different frequencies.

After cleaning the data, we average the PM 2.5 values within 24 hours per day and average the PM 2.5 values in each month respectively. First we analyze the daily PM 2.5 values.

## Part1: Daily PM 2.5 values

Before doing any further analysis, we first plot the daily values of PM 2.5 from 01/01/2010 through 31/12/2014:

 The time series plot shows that changing variablity exists in the daily PM 2.5 data. To remove the changing variability, we perform logarithm on our data and then plot the ACF and PACF of the data:

 Above plots show that both the ACF(autocorrelation function) and PACF(partial autocorrelation function) of daily PM 2.5 records have quick decays so we do not need to do differencing on the data. Also note that in the ACF plot(left), the values of ACF are outside the bounds at lag 1 and lag 2, which suggests that we could fit a MA(2) (moving average 2) model on the records.

Here we use function **auto.arima( )** in R package *forecast* to conduct the model fitting:

fit\_daily = auto.arima(ts(pm\_daily))  
fit\_daily$coef

## ma1 ma2 intercept   
## 0.5553144 0.1135664 4.2399056

To evaluate the quality of this fitted model, we need to perform several hypothesis tests on its fitted residuals. If the fitted model is a good fit, we should expect that the residuals behave like white noise.

**Test of randomness:**

We use Box-Pierce test to test the randomness of the residuals of daily PM 2.5 data.

Box-Pierce test:

1. the data are independently distributed.
2. Under , the test statistic is .
3. Reject if at level .

##   
## Box-Pierce test  
##   
## data: daily\_res  
## X-squared = 8.4383, df = 20, p-value = 0.9885

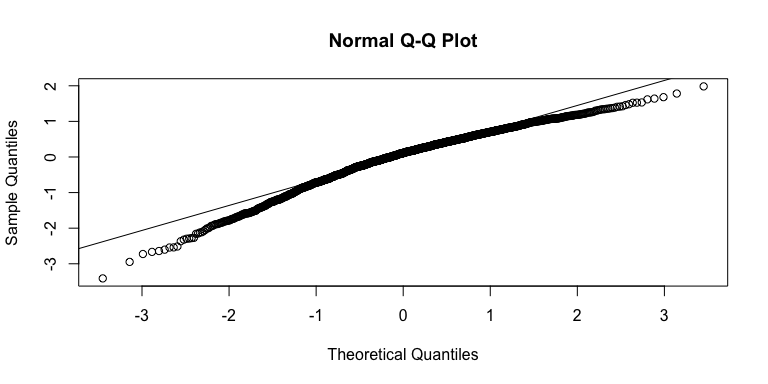
The p-value of Box-Pierce test on the residuals is 0.9885 which is larger than the significance level , so we fail to reject , the residuals are independently distributed.

**Test of normality:**

We first use normal probability plot to evaluate the normality of our data and then use Shapiro\_Wilk test to do the hypothesis testing.

The Shapiro-Wilk test is a test of normality in frequentist statistics:

1. data are normally distributed.
2. Under , the test statistic is ,where is the order statistic and is the sample mean.
3. Reject if at level .



##   
## Shapiro-Wilk normality test  
##   
## data: daily\_res  
## W = 0.96829, p-value < 2.2e-16

The Normal Q-Q plot shows that the points are not approximately lie on the line and heavy tails exist on both sides. Also Shapiro-Wilk test gives us a very small p-value. Hence, we should reject the null hypothesis and conclude that residuals are not normally distributed.

Finally, we compute the sample mean and sample variance of the residuals that are -0.00039(almost zero) and 0.55617 respectively. Therefore,the residuals are not Gaussian white noise, but are independently distributed with sample mean 0 and sample variance 0.55617.

Using the same methods but now we’re doing testing on daily PM 2.5 data.

**Test of randomness:**

##   
## Box-Pierce test  
##   
## data: pm\_daily  
## X-squared = 431.28, df = 20, p-value < 2.2e-16

The p-value of Box-Pierce test on the daily PM 2.5 data is , so we reject , i.e. the daily PM 2.5 data are dependent.

**Test of normality:**

##   
## Shapiro-Wilk normality test  
##   
## data: pm\_daily  
## W = 0.9852, p-value = 1.261e-12

The p-value of Shapiro-Wilk test on the daily PM 2.5 data is , so we reject , i.e. the daily PM 2.5 data are not normally distributed. And this is not surprising because the result of normality test on the fitted residuals shows that the normality is violated.

**Test of stationarity:**

Now we use Augmented Dickey-Fuller test to test whether the daily PM 2.5 data are stationary time series process(whether a unit root is present in a time series sample):

Augmented Dickey-Fuller test:

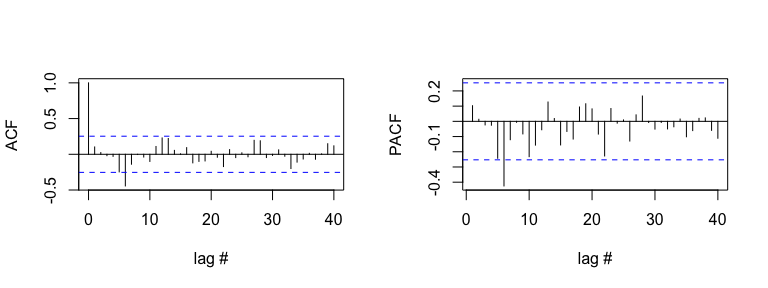
1. The testing procedure is applied to the model:, where is a constant, is the coefficient on a time trend and is the lag order of the autoregressive process.
2. Hypothesis: : (the time series data are **not** stationary) (the time series data are stationary).
3. Under , the test statistic is

##   
## Augmented Dickey-Fuller Test  
##   
## data: pm\_daily  
## Dickey-Fuller = -10.407, Lag order = 12, p-value = 0.01  
## alternative hypothesis: stationary

The p-value of Dickey-Fuller test on the daily PM 2.5 data is 0.01, so we reject , i.e. the daily PM 2.5 data are stationary time series process.

## Part 2: Monthly PM 2.5 values

In this section, we are going to do time series analysis on monthly PM 2.5 data by averaging the PM 2.5 values in each month. The analysis approach is very similar to the last section, first we fit the monthly data a time series model and then perform hypothesis tests on the fitted residuals and the data itself.

 The ACF and PACF shown in the figure above are suggestive of an MA(6) or AR(6) model, as the value of ACF drops dramatically after lag 6 while the value of PACF has a sharp decrease after lag 5. Then we can use function **Arima( )** in R package *forecast* to conduct the model fitting:

fit\_monthly\_ma = Arima(pm\_monthly,order=c(0,0,6))  
fit\_monthly\_ar = Arima(pm\_monthly,order=c(6,0,0))

Here we use **AIC**(Akaike information criterion) to choose which model that we prefer to use for further analysis. The AIC is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.(Wiki link:<https://en.wikipedia.org/wiki/Akaike_information_criterion>)

Model selection for MA(q) model:

Model selection for AR(p) model:

## MA(6) AR(6)  
## AIC 12.83 18.467

We prefer the model with smaller AIC value, so we choose MA(6) model.

**Test of randomness:**

##   
## Box-Pierce test  
##   
## data: monthly\_res  
## X-squared = 8.7963, df = 20, p-value = 0.9851

The p-value of Box-Pierce test on the residuals is 0.9851 which is larger than the significance level , so we fail to reject , i.e. the residuals are independently distributed.

**Test of normality:**

##   
## Shapiro-Wilk normality test  
##   
## data: monthly\_res  
## W = 0.98071, p-value = 0.4591

The p-value is 0.4691, we should fail to reject , residuals are normally distributed.

Hence, the residuals are Gaussian white noise, i.e. independently distributed with sample mean -0.00218 and sample variance 0.0025.

Next, we test the randomness, normality and stationarity of the monthly PM 2.5 records.

##   
## Box-Pierce test  
##   
## data: pm\_monthly  
## X-squared = 27.581, df = 20, p-value = 0.1197

##   
## Shapiro-Wilk normality test  
##   
## data: pm\_monthly  
## W = 0.9882, p-value = 0.8307

##   
## Augmented Dickey-Fuller Test  
##   
## data: pm\_monthly  
## Dickey-Fuller = -3.5647, Lag order = 3, p-value = 0.04378  
## alternative hypothesis: stationary

According to the output, we conclude that monthly PM 2.5 data are stationary time series process, and they are independently and normally distributed.

## Summary

Fitting daily PM 2.5 data by MA(2) model, we obtain non-Gaussian but independently distributed noise and the data itself is stationary, dependent but non normal distributed.

Fitting monthly PM 2.5 data by MA(6) model, the fitted residuals are normally and independently distributed. And the monthly PM 2.5 data are stationary and independent Gaussian process.

We want the fitted residuals to bahave like Gaussian white noise, thusly we consider that the fitted model MA(6) using monthly PM 2.5 data might be a good fit, while the fitted model MA(2) using daily PM 2.5 data does not work well. To build a better time series model for daily PM 2.5 data, one might need to consider more complicated time series models. Compare the AIC and RMSE (Root Mean Square Error: ) of two fitted models, fitted model MA(6) has much smaller AIC value and lower RMSE.

## AIC RMSE  
## Daily PM 2.5 4034.693 0.746  
## Monthly PM 2.5 12.830 0.221

Last but not least, the time series analysis that we have performed in this project is naive due to the lack of knowledge and practical experience. To make more precise time seires models on these two records, one could consider a seasonal ARMA model or doing (parametic or non-parametric) regression with ARIMA errors, which requires more advanced knowledge related to these topics.