A Time Series Analysis of the Temperature in Vancouver: Correlation and Forecast

4/8/2020

Contents

```
knitr::opts_chunk$set(echo = TRUE)
#install.packages('kableExtra')
#install.packages("dplyr")
#install.packages('zoo')
#install.packages("tseries")
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -------------------------------------------------------------------------------- tidyver
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(tseries)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

I. Introduction

Weather conditions are highly influencial to social and personal activities. Effective temperature prediction can help with decision making and resource allocation. This study tries to perform a time series analysis to investigate behaviours of the temperature in Vancouver, using historical weather data of Vancouver, Canada from 2012 to 2019.

1.1 Motivation and the Theme Question

In daily life, we can observe that the temperature today is often correlated with the temperature tomorrow, e.g. if today is a cold today, it is very likely that tomorrow also is, so we start to wonder if there is a connection between the historical temperature data and the future. However, it could also be that the "correlation" is purely due to the seasonal factors in the climate, e.g. since "today" and "tomorrow" are both in winter, they would surely be cold. Hence, we can roughly divide this observed correlation into two parts: seasonal effect, and the correlation between the deseasonal historical and deseasonal future data. In this study, we aim to investigate whether the later effect is significant. Further, we intend to evaluate whether we can use the historical temperature to better forecast future temperature.

1.2 Data Collection and Cleaning

The data was provided by "Climate Canada" and recorded at the station: "VANCOUVER HARBOUR CS". The dataset contains several meteorological measurements including daily maximum, minimum, and mean temperature and their related property flags, daily maximum, minimum, and mean precipitation and their related proeprty flags, as well as many other weather variables. After data cleansing and wrangling, **Date** and **Mean Temperature** between **January 1st, 2012** and **December 31st, 2019** inclusive were kept for further analysis. Moreover, to keep number of days in each year consistent thereby to facilitate periodical analysis, we deleted the two extra days from two leap years, which are 2012-2-29 and 2016-2-29.

The dataset comprises records with missing mean temperature values. The R built-in function "na.fill" in package "zoo" is used to interpolate and fill out the missing values

```
## Initial number of records with NA dates:  0
```

```
## Initial number of records with NA temperatures:  30
```

After interpolation, the dataset has following properties

Table 1: Summary of Data Set

Variable	Type	Unit	Range
Date	discrete Series	date	[2012-01-01, 2019-12-31]
Mean Temperature	numerical	celcius	[-4.1 , 24.8]

Table 1. gives a brief summary of variables we selected.

The dataset is split into training set and test set. The first seven years from 2012 to 2017 inclusive is used for training the time series model. The last one year 2018 is used to test the performance of the model.

II. Analysis

2.1 Preliminary Analysis and Correlation

Below is the plot of the training set

```r
plot(temp.de)
```

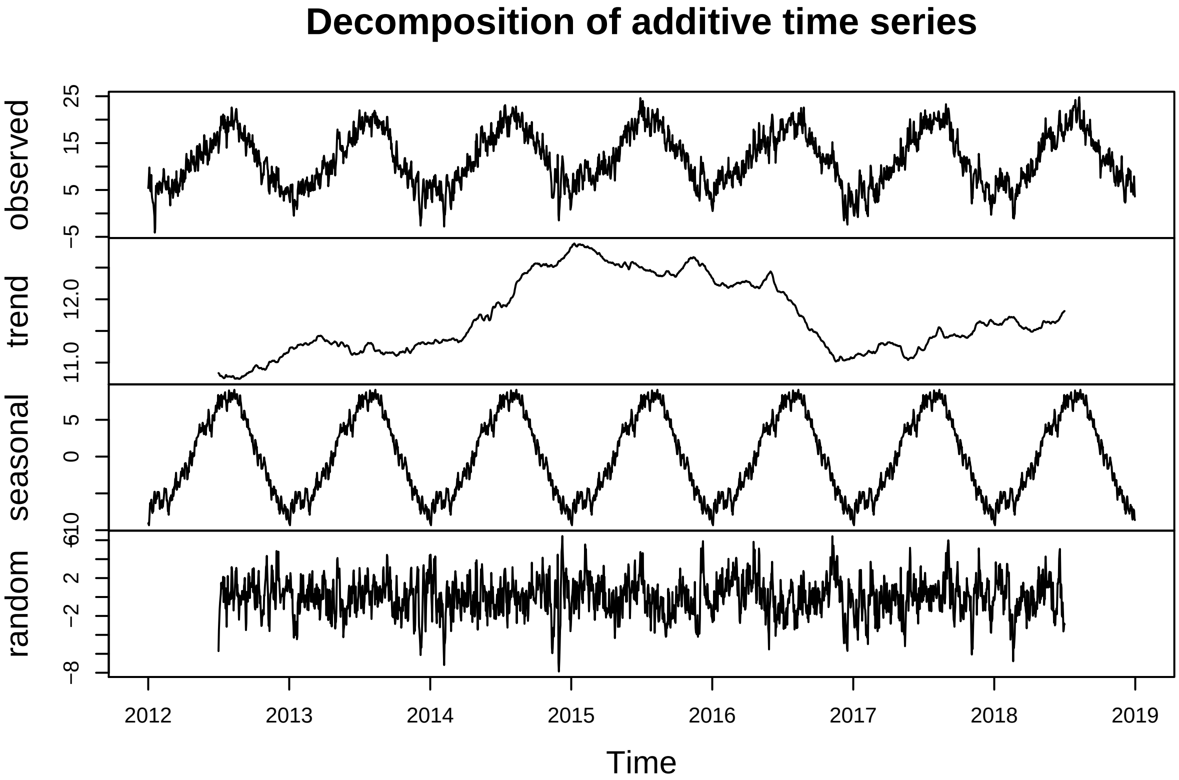
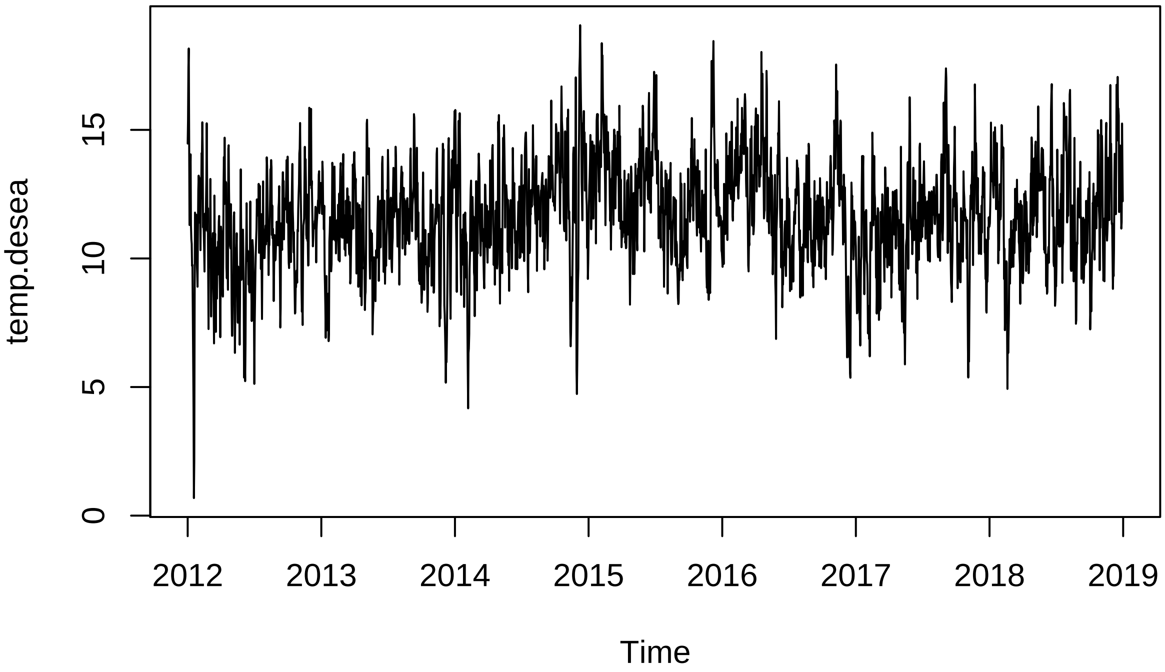


Figure 1. shows there is a strong seasonal pattern within the data and the time series is non-stationary. We deseasonalized the data and used the adf function to test the stationarity of the time series.

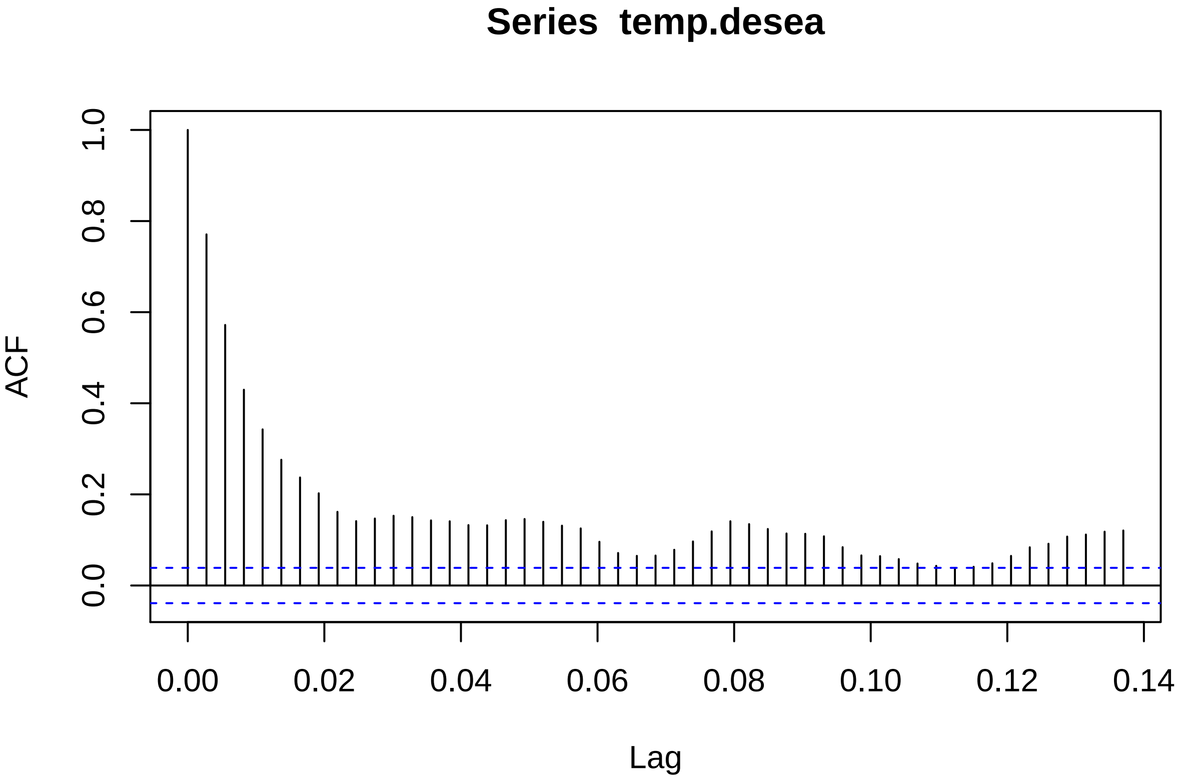
```
## Warning in adf.test(temp.desea): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  temp.desea
## Dickey-Fuller = -9.5134, Lag order = 13, p-value = 0.01
## alternative hypothesis: stationary
```

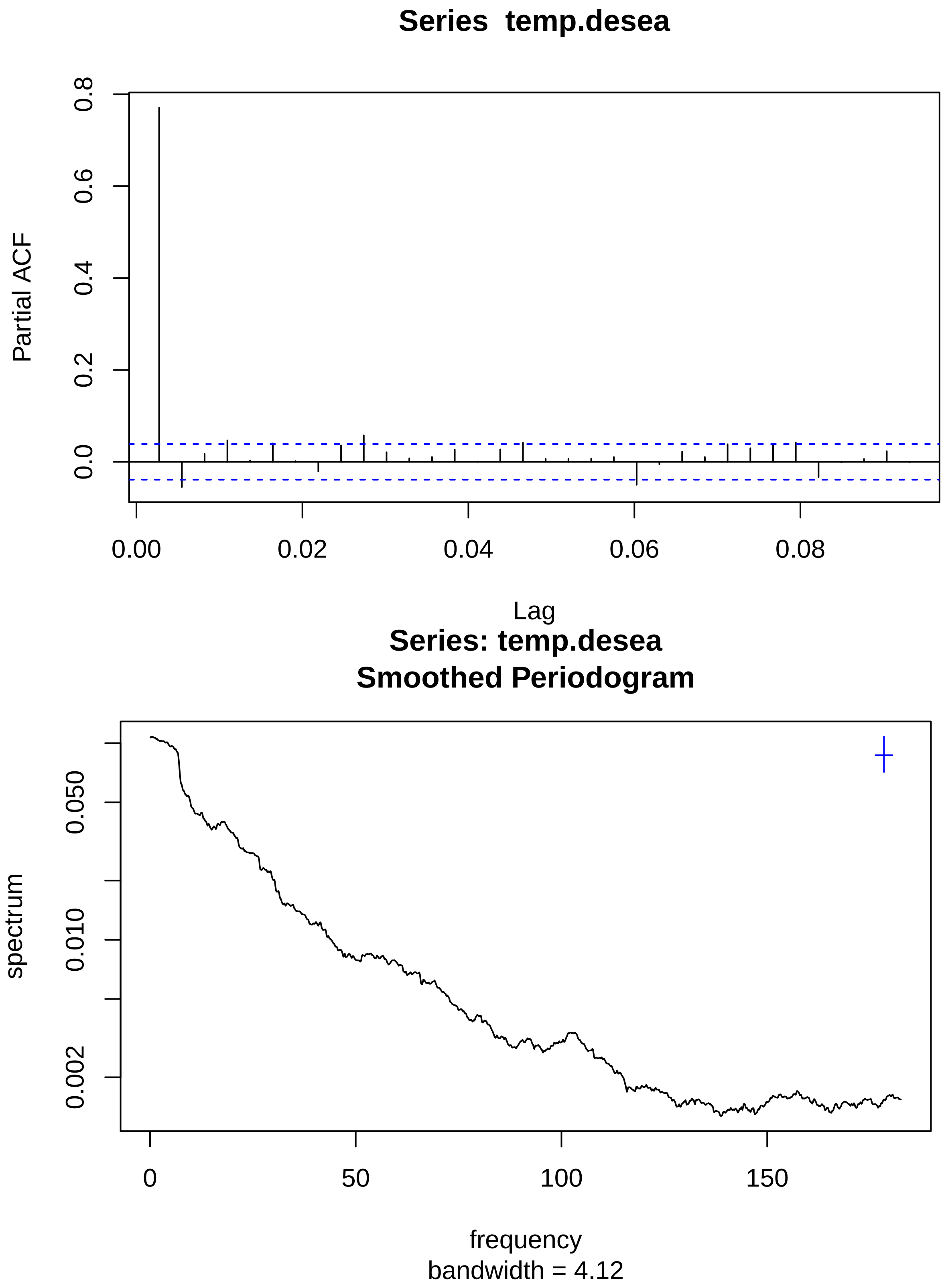
```r
plot(temp.desea)
```



After deseasonalization, Figure 2. shows no trend and seems to become stationary. And the P-value of Augmented Dickey-Fuller test, 0.01 is less than 0.05, thus also indicate that the deseasonalized time series is stationary.

Acf, pacf, and Periodogram of the stationary time series are generated and exhibited blow.





From

Figure 3., It is noticeable that acf shows a sine waved pattern and tails off. Pacf is positively significant at lag=1. Periodogram is dominated by low frequencies and shows no "hidden" period. Now, the first part of our theme question can be answered: indeed, there is still a positive correlation between consecutive deseasonal data aside from the seasonal effect, e.g. the correlation between "today is cold"" and "tomorrow is cold”" is not **only** because of the seasonal effect.

2.2 Forecasting

Knowing that there is indeed a correlation between the consecutive data, we move to investigate whether past data, aside from determine seasonal effect, can be useful in forecasting. To see this, we try to forecast temperatures in two ways: using only the seasonal effect, and use both the seasonal effect and our ARMA model for deseasonal data. If the later method is significantly better, we can conclude that historical data is useful in improving the forecasting results.

2.2.1 Model Construction

We aim to build a forecasting model based on the Box-Jenkins approach. ARMA model is chosen to be apllied according to the features of the data. ML and Grid-search is used to find the best-fit model which has the minimum AIC.

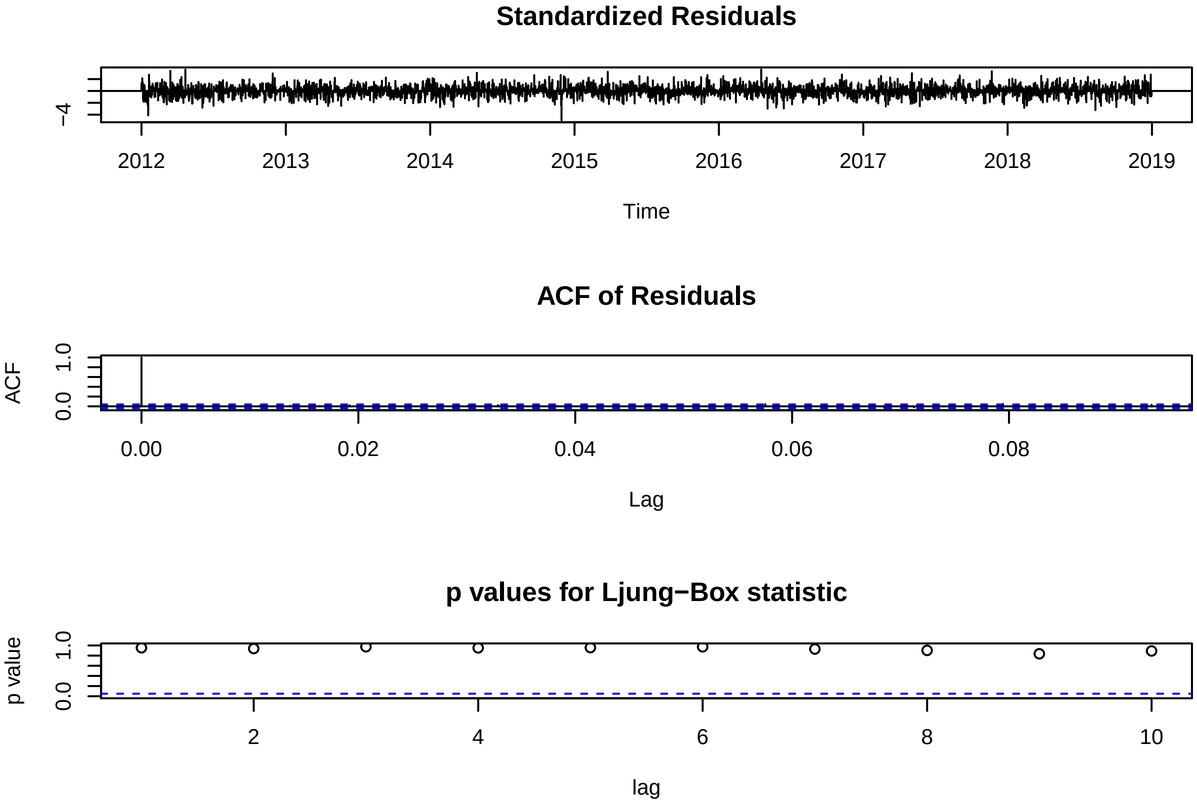
Notice to grader: the code below take minites to run because of the ML convergence problem. Alternatively, you can just read the result from a pre-compuated csv file using the last line.

```
## [1] 8752.107
```

```
##          V1       V2       V3       V4       V5       V6       V7
## 1 11084.125 9621.968 9088.303 8949.568 8853.025 8827.671 8818.908
## 2  8780.167 8774.441 8776.265 8770.994 8772.936 8768.650 8768.278
## 3  8774.706 8776.383 8756.933 8758.338 8758.141 8759.419 8770.425
## 4  8775.808 8756.627 8773.811 8758.145 8759.172 8763.854 8772.511
## 5  8772.008 8773.947 8758.998 8759.343 8778.191 8763.094 8774.365
## 6  8773.992 8758.554 8759.380 8778.531 8776.542 8752.107 8776.474
## 7  8771.998 8773.993 8763.665 8769.926 8753.606 8752.447 8755.171
```

Since p=5 q=5 have minimum AIC, ARMA(5,5) is the best-fit ARMA model

```
## 
##  Box-Pierce test
## 
## data:  arma55$residuals
## X-squared = 0.0033009, df = 1, p-value = 0.9542
```



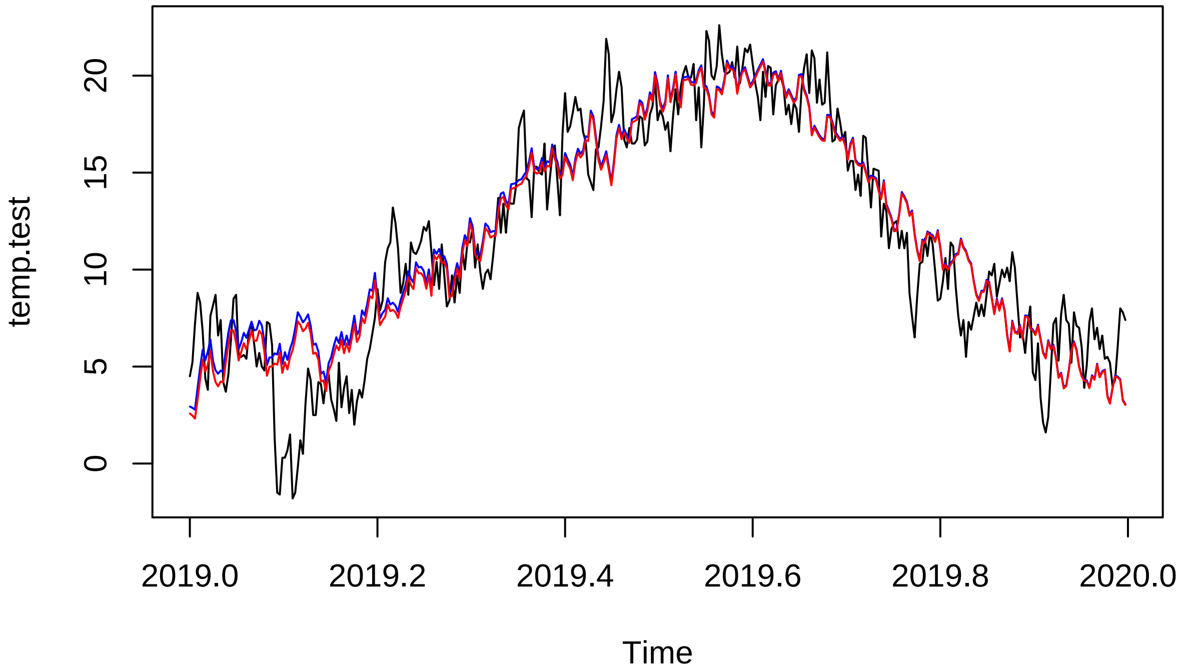
The above model diagnostics suggest that our model is appropriate.

2.2.2 Forecast Temperature in One Year

In this section we aim to compare the forecasting results in the long term, or, in scope of one year. Below we see the difference in prediction using purely the seasonal effect and prediction using both seasonal effect and ARMA model.

Seasonal + ARMA

Pure Seasonal Effect



Comparison

Since the blue and red line looks extremely close. The two forecasts converge very quickly and there is no significant difference in the long term. Therefore, we conclude that at least for long term, historical data won't really improve the forecasting results. This is actually intuitively appealing, since knowing the weather today won't really help me to predict the weather two month from now. In the long term, seasonal effects dominate.

2.2.3 Forecast Temperature in One Day

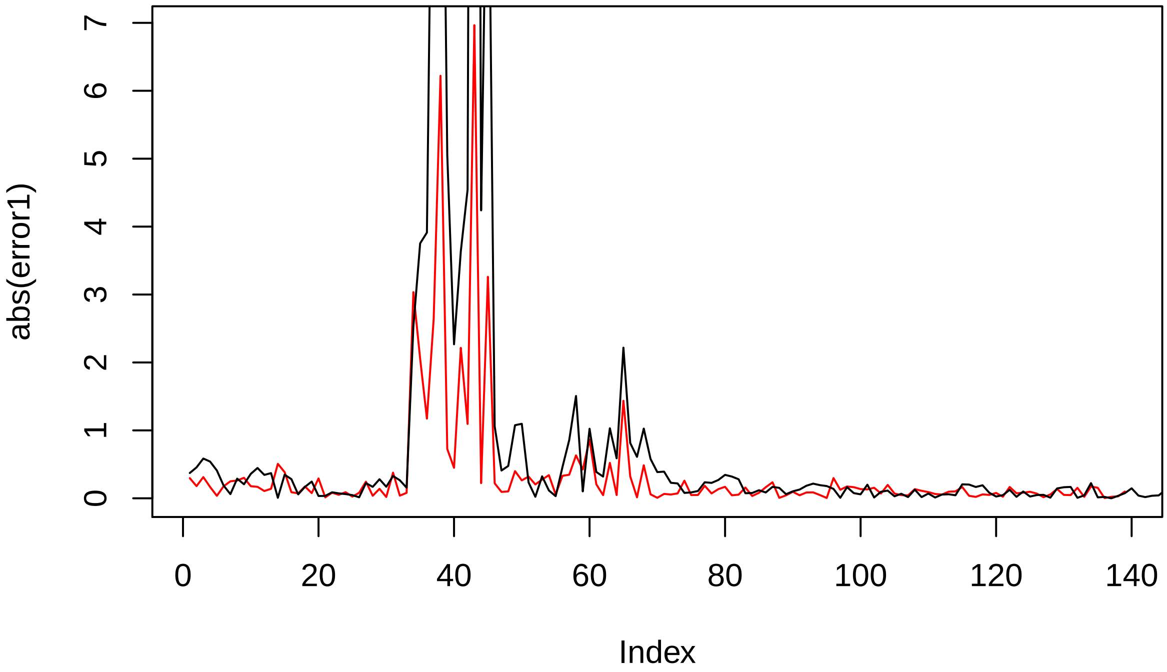
Now we narrow our scope to the forecast to just one day in the future (super short term) and try to see whether the historical data would be useful in short term prediction.

Seasonal + ARMA

The code below does the following: at first, we know the historical data of 7 years and constructing an ARMA model, trying to using it, together with the seasonal effect, to predict the first day's temperature. Then, we "proceed" to the next day and construct a new ARMA model incorporating the new information (the temperature that day), and try to predict the temperature of the day after that. We repeat the process, constructing 365 ARMA models and made 365 one-day predictions. For each predictions, we calculate the relative error of it.

Notice to grader: the code below would potentially take up to half an hour to run due to the computation complexity and ML convergence problem. Alternatively, you can just read the result from a pre-compuated csv file using the last line.

Pure Seasonal Effect



Comparison

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.003389 0.055509 0.110428 0.359702 0.255066 6.965526
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00005  0.04884  0.11552  0.46779  0.25635 32.22005
```

Since the red line is lower than the black one, the method using seasonal effect combined with ARMA model has smaller errors and thus a better method. There is a large difference in Mean errors of two methods (21% vs 47%), while the difference in Median errors is small (8% vs 12%). This suggests that the first model perform better mostly because it can deal with the acyclical, abnormal temperatures better. This result is intuitively appealing, since it essentially shows that when the temperature behaves in accordance with the seasonal effect (pure seasonal effect can give a good result), historical data won't matter that much. However, if the temperature behaves weirdly, e.g. cold whether in the summer, the historical data approach can be more flexible and better at grasping the abnormality thus perform better.

III. Conclusion

This study used a time series analysis approch to study the correlation and forecast methods regarding the temperature in Vancouver based on the meteorological data from 2012 to 2019. The results revealed that there is a positive correlation between consecutive data aside from the seasonal effect. Regarding the prediction power of the deseasonal historical data, we find that it doesn't matter that much when it comes to long term. However, in the short term, due to its flexibility it can greatly enhance the accuracy of the prediction. These findings show that a time-series focus on the deseasonal historical may have the potential in providing more accurate predictions of the weather in the short term. These findings could be helpful to support decisions regarding personal schedules, business management, allocation of medical resources, and other matters.