Time Series Analysis and SARIMA Modelling of Precipitation in Singapore

Xuhui Wang, Jasmine Gao, Chao Gao

December 2021

Introduction

For every region in the world, rainfall, as a ubiquitously measured meteorological variable, is very important for agricultural production, natural resource management and industry development. On the growing importance of rainfall studies in many fields including climate change research and high performance computing, different users, starting from a scientist to a policy maker, need rainfall data analysis and prediction in advance for their application. Since the economy of Singapore is mainly driven by exports, logistics and tourism, the analysis and prediction of rainfall has considerable value, in both theory and application, to Singapore manufacturing and service industries. Our study firstly provides an analysis of the rainfall data during a long period of time in Singapore, then, we would derive a SARIMA model and make the diagnostics and assessment, and finally, we would make reasonable prediction for the future time dependence as a reference that could be used in the industries which may need statistics about rainfall. For seasonal ARIMA models, it takes at least 50 but preferably more than 100 observations to achieve a good analysis and forecast performance (Box, Jenkins, Reinsel, & Ljung, 2015). Another important thing is to choose the sample size which might effect the time series trend correctly (Box & Tiao, 1975). Because the variation of rainfall in a region is easily subject to many variables including atmospheric temperature, wind speed and wind direction, each of which does not vary over time regularly, it is not that helpful to simply increase the sample size for the model determination. Given the above reasons, we choose to track the monthly rainfall data in Singapore for 20 years from January 2001 to July 2020 cut from the original data with the sample size of 38 years (NEA-Singapore, 2020) (Approved by the instructor Dr. Zhou). As an example, Shamsnia *Et al.* used statistics of relative humidity and monthly average temperature and precipitation of Abadeh Station for 20 years as well deriving an ARIMA model and achieved good performance (Shamsnia, Shahidi, Liaghat, Sarraf, & Vahdat, 2011).

Preliminary Analysis

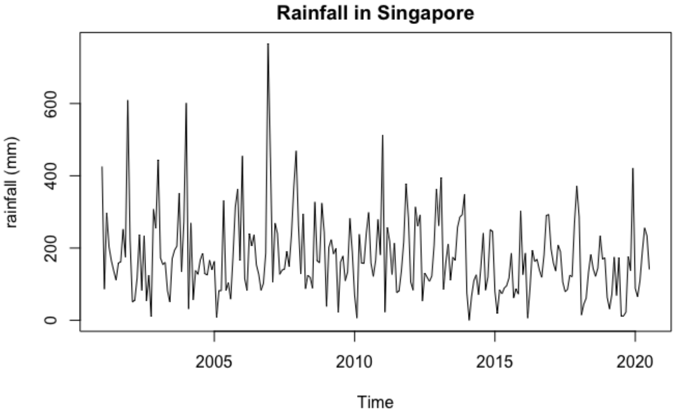


Figure 1: Raw data of rainfall in Singapore

From the plot of raw rainfall data in Figure 1, it is clear that there are some significant outliers within the time series data considered on monthly basis, however, the series still looks relatively stationary by a brief observation. By performing the Augmented Dickey–Fuller test on the raw data as a reference, the series shows stationarity since p-value = 0.01 (the detailed test results included in appendix due to the limited number of pages), which preliminarily verifies our initial observation. The regular patterns of ups and downs in Figure 1 could be considered as an indication of seasonality. To further analyze the characteristics of the raw data such as trend or seasonality, we decompose it into various components for the next step.

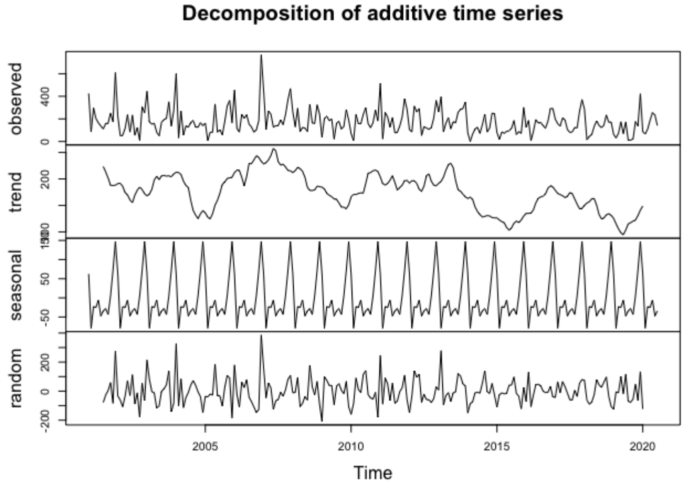
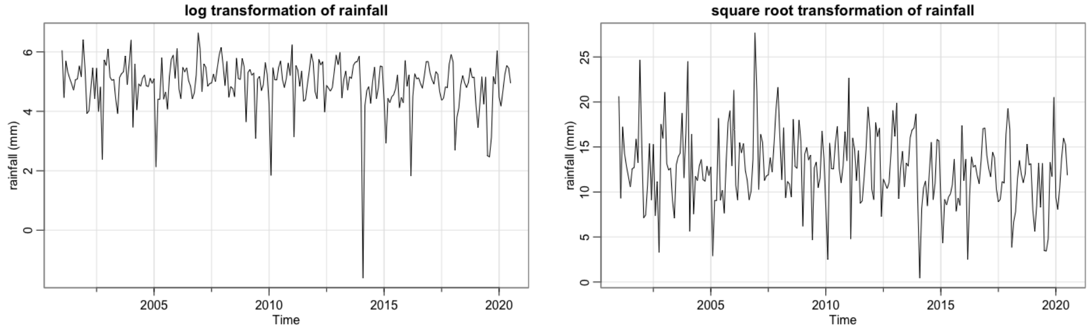


Figure 2: Decomposition of raw data of rainfall in Singapore

As could be seen from the plot of decomposition in Figure 2, the trend seems quite irregularly stable over time, despite a few ups and downs over the whole period. Furthermore, the random effect could also be considered as a stable and stationary-looking series over the whole period. However, by looking at the seasonal part, we would notice that there are seasonal effects with regular rising and falling patterns appearing yearly over the whole period, which implies that regular monthly rainfall recorded each year was influenced by these two patterns of the seasonality component. Given the above observations, in order to derive the appropriate model, we may make a seasonal difference and choose the seasonal lag parameter $s = 12$ later on, since the time series is monthly data.

Before the model derivation, we need to ensure the stationarity of the series visually and stabilize the mean and variance. Thus, in spite of the verification by the previous ADF test, it is still meaningful to make transformation on the time series to standardize the data.

We perform two transformations, logarithm and square root, and investigate their performance statistically to choose the better one. Apparently, comparing with the logarithm transformation method as shown in Figure 3(a), the square root transformation method provides more stable mean and variance without significant outliers, as shown in Figure 3(b).



(a) Logarithm

(b) Square root

Figure 3: Transformation of raw rainfall data

Since it might be necessary to perform a seasonal differencing to make a stationary time series as mentioned above, we discuss the characteristics of the data after square root transformation and seasonal differencing. As shown in Figure 4, it is clear that the data has constant mean and variance. By performing another Augmented Dickey–Fuller test on the data after square root transformation and seasonal difference as a reference again, the series also shows stationarity since p-value = 0.01 (the detailed test results included in appendix due to the limited number of pages). So far, we have examined the stability of the transformed and differenced time series, and we now move on further to the analysis of ACF and PACF about the raw data to verify our ideas.

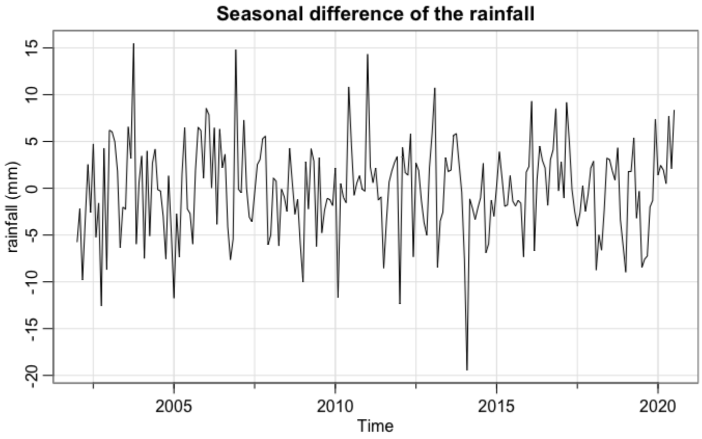
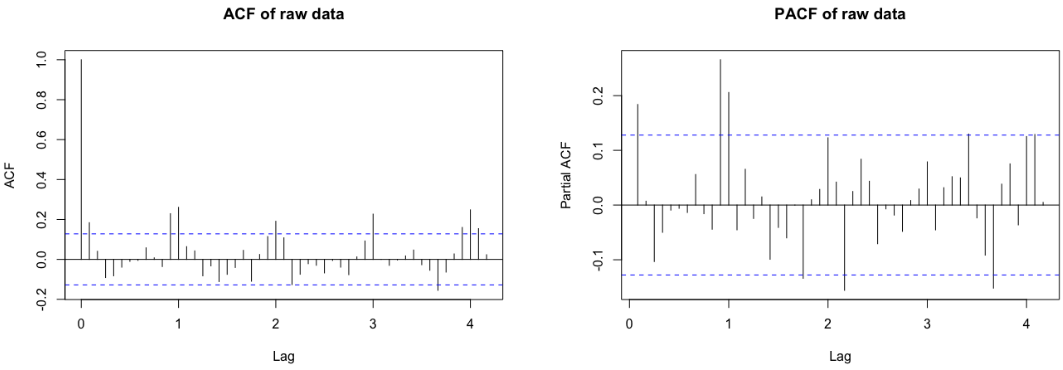


Figure 4: Rainfall data after square root transformation and seasonal differencing

Model Identification

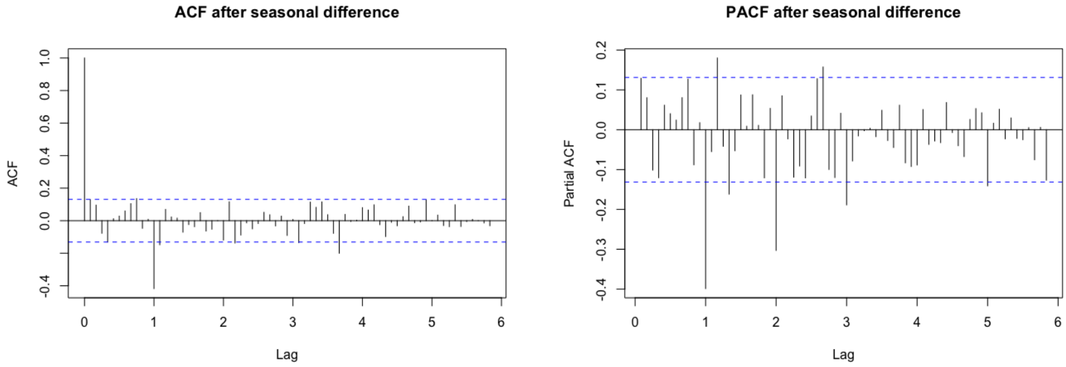


(a) ACF of raw data

(b) PACF of raw data

Figure 5: ACF and PACF of raw rainfall data

The sample auto-correlation function (ACF) and the partial auto-correlation (PACF) about original rainfall data are shown in Figure 5. There is no apparent pattern in ACF as spikes appear every seasonal lags, which implies that it is necessary to further investigate the characteristics of the time series by transformation and seasonal differencing as mentioned in the last section.



(a) ACF of transformed data

(b) PACF of transformed data

Figure 6: ACF and PACF of rainfall data after seasonal difference and square root transformation

After we have examined the mean and variance of the time series in the last section, we will look at ACF and PACF of the series after transformation and seasonal differencing. From Figure 6(a), by inspecting the early lags, we can see that ACF goes to zero very fast, it is safe to say the time series is now stationary. Next, let us investigate the ACF and PACF together to determine possible models.

From Figure 6, a spike at lag 12 in ACF is significant while no other is significant at lags; The PACF represents an exponential decay at the seasonal lags, which is 12, 24, 36 etc. Hence, the seasonal part of the SARIMA model has a moving average term of order 1. For the non-seasonal part, both ACF and PACF tail off. Therefore, we assume that the non-seasonal part has a moving average term of order 1 and an autoregressive term of order 1. Based on the features protrayed by the two plots, we obtain an initial model, which is SARIMA $(1, 0, 1) \times (0, 1, 1)_{12}$. In order to propose the best model with the minimum AIC, we investigate the neighboring models using nested loop and the result is shown in Figure 7, where the best model to be chosen for the rainfall data is SARIMA $(0,0,2) \times (0,1,1)_{12}$ with the smallest AIC score of 5.5755.

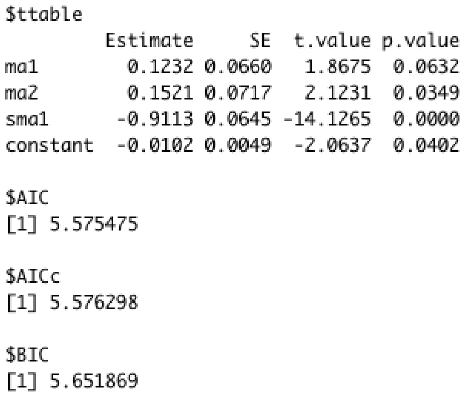
Table 1: SARIMA(p,0,q) ×(0,1,1)12 model

	MA0	MA1	MA2	MA3
AR0	5.5891	5.5861	5.5755	5.5827
AR1	5.5833	5.5883	5.5836	5.5897
AR2	5.5816	5.5868	5.5887	5.5954
AR3	5.5824	5.5898	5.5946	5.5950

Figure 7: Table of AIC values of the model cluster

Assessment & Forecasting

The model parameters of the proposed model are all significant as shown in Figure 8. We now illustrate the diagnostics plot in terms of the standardized residuals plot, ACF of residuals plot, the residual QQ plot and the Ljung box plot.



```
$ttable
         Estimate     SE  t.value p.value
ma1        0.1232 0.0660   1.8675  0.0632
ma2        0.1521 0.0717   2.1231  0.0349
sma1      -0.9113 0.0645 -14.1265  0.0000
constant  -0.0102 0.0049  -2.0637  0.0402

$AIC
[1] 5.575475

$AICc
[1] 5.576298

$BIC
[1] 5.651869
```

Figure 8: ttable of the final model

The standardized residuals plot from Figure 9 illustrates that the residuals have constant mean and variance. Furthermore, the residual Q-Q plot claims that the distribution of the residuals of the final model is Gaussian. The ACF of residuals also looks like an ACF of the white noise. Additionally, in Ljung box plot, all p values are above the blue line, which implies that the residuals are all uncorrelated. Therefore, our proposed final model is justified.

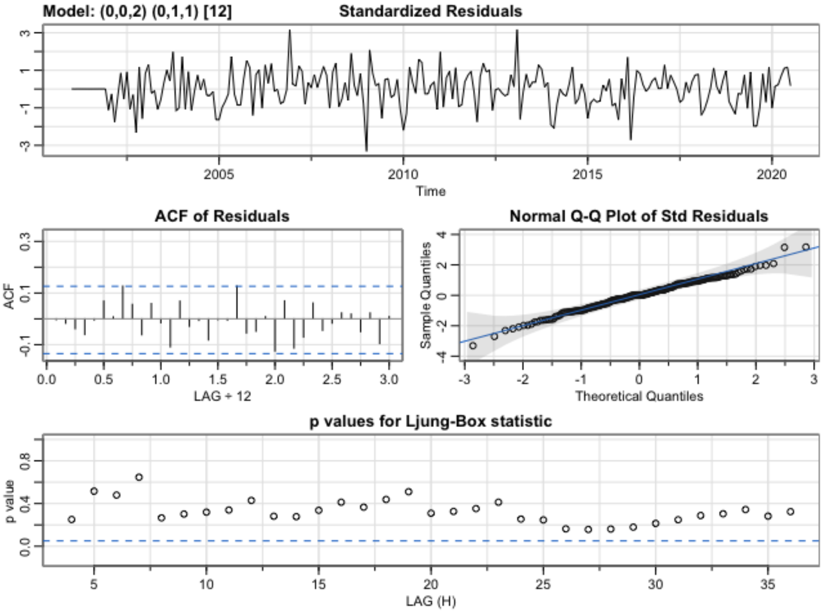


Figure 9: Diagnostics of final model

The proposed model is demonstrated as the equation 1, where x_t represents the square root of the rainfall data, w_t represents the white noise and B denotes the backshift operator.

$$(1-B^{12})x_t = (1+0.1232_{(0.0660)}B+0.1521_{(0.0717)}B^2)(1-0.9113_{(0.0645)}B^{12})w_t+0.0102_{(0.0049)} \tag{1}$$

So far, we have done the analysis and model derivation. As the final step, here we perform forecasting for the rainfall data given the time dependence. We make the prediction for the next ten-month observations based on our proposed model. From Figure 10, where the red points represent the predicted rainfall data and the blue curves represent the values restored by (the predicted value based on the square root data$\pm 2\times se)^2$, we could observe that the predicted data agree well with the previous observations, and it is able to replicate much of the seasonal variation in the original rainfall data. In addition, all the points lie within the two blue error bands. Therefore, we conclude that the predictive power of our proposed SARIMA model $(0,0,2) \times (0,1,1)_{12}$ is very appreciable.

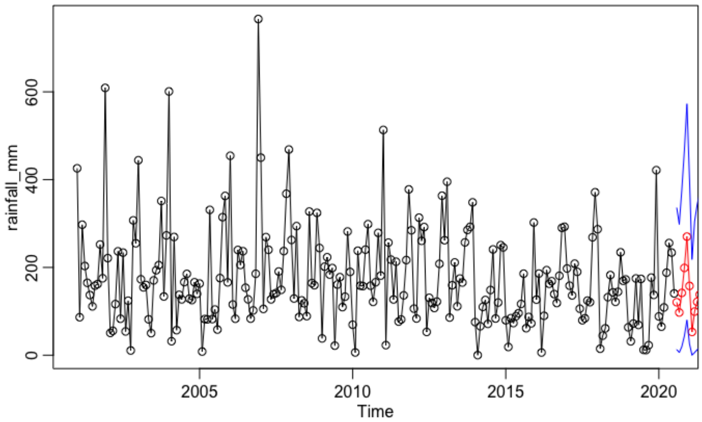


Figure 10: Prediction for the next 10 monthly observations

Conclusion

In this project, firstly, we introduced the importance of research on rainfall analysis and prediction. Then we analyzed the time series of rainfall data in Singapore over the recent 20 years, in which, by performing the square root transformation and seasonal difference of the raw data, we stabilized the mean and variance. Then we achieved the stationary time series by observing the ACF. Proceeding to conduct model identification and diagnosis, we proposed a SARIMA model and made the justification. Finally, we moved on to predict the next ten-month data using our own proposed model and the predictions agree well with the previous data. The evaluation to the proposed model and the predicted values show that SARIMA is a reasonable and appropriate model to fit rainfall data. One of the concerns generated during our work is that we believe that a time series model has different facets, as in the case where the rainfall in one region could depend on more than one variable, so the modelling and prediction of rainfall might not be accurate enough with the time dependence only. Therefore, we look forward to working on a multivariate time series analysis again in the future.

References

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Box, G. E., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, *70*(349), 70–79.

NEA-Singapore. (2020). Monthly rainfall in singapore. *https://www.kaggle.com/kelvinchow1979/monthly-rainfall-in-singapore*.

Shamsnia, S. A., Shahidi, N., Liaghat, A., Sarraf, A., & Vahdat, S. F. (2011). Modeling of weather parameters using stochastic methods (arima model)(case study: Abadeh region, iran). In *International conference on environment and industrial innovation* (Vol. 12, pp. 282–285).

Appendix

Work Distribution

Throughout the whole project, Eric is mainly responsible for the introduction and preliminary analysis part. Jasmine is mainly responsible for the model identification part. Chao is mainly responsible for the assessment and prediction part. Despite everyone's work distributed, we collaborated and communicated in a reciprocal way of studying. We are also enormously grateful to have had this precious opportunity to learn from Dr. Zhou and work on such a meaningful project.