TIME SERIES FORCASTING USING SARIMA FOR RAINFALL PREDICTION

Final project for EAEE 4000 Machine Learning for Environmental Engineers

Zhenyu Kang

Zk2249

Contents

Introduction

Over the last few decades, the climate of the earth has changed extremely in terms of variation of rainfall and temperature. Changing pattern of precipitation and temperature affect the climatic changes (John and Brema 2018). The changes in rainfall may differ due to different geographical countries' areas. The trending pattern of rainfall has become a significant factor in agricultural countries. According to Yasmeen and Hameed (2018), rainfall projection modelling assumes a combination of probabilistic models, knowledge, and trending patterns of observation. So, the trend in the rainfall behavior has become more important to fit the time series model.

In Machine Learning, a seasonal autoregressive integrated moving average (SARIMA) model is a different step from an ARIMA model based on the concept of seasonal trends. Seasonal variations of the time series can take into account periodic models, allowing more accurate predictions. Seasonal ARIMA (SARIMA) defines both a seasonal and a non-seasonal component of the ARIMA model, allowing periodic characteristics to be captured.

The most difficult investigation is to forecast the rainfall due to their time and space variation (Nyatuame and Agodzo 2018). In the statement of these authors, attempts have been made to predict the behavioural pattern of rainfall and temperature using the stochastic ARIMA model technique in Tordzie watershed of Ghana. Seasonal Autoregressive Moving Average (SARIMA) (3, 0, 3) (3, 1, 3) and SARIMA (3, 1, 3) (3, 1, 3) were identified as appropriate model for forecasting the annual rainfall and maximum temperature for that region.

This report Implements a Time Series Analysis model using the SARIMA algorithm for forecasting of monthly rainfall values. ACF and PACF plots as well as AIC values were used for obtaining suitable parameters for the model.

Data

The monthly average rainfall of Weifang city, China (119.183 , 36.767) data from January, 1980 to November, 2021 (42 years) have been obtained from China meteorological data service center (http://data.cma.cn/). Basic statistics like mean, standard deviation of all data series were calculated for visualizing the data on a monthly basis.

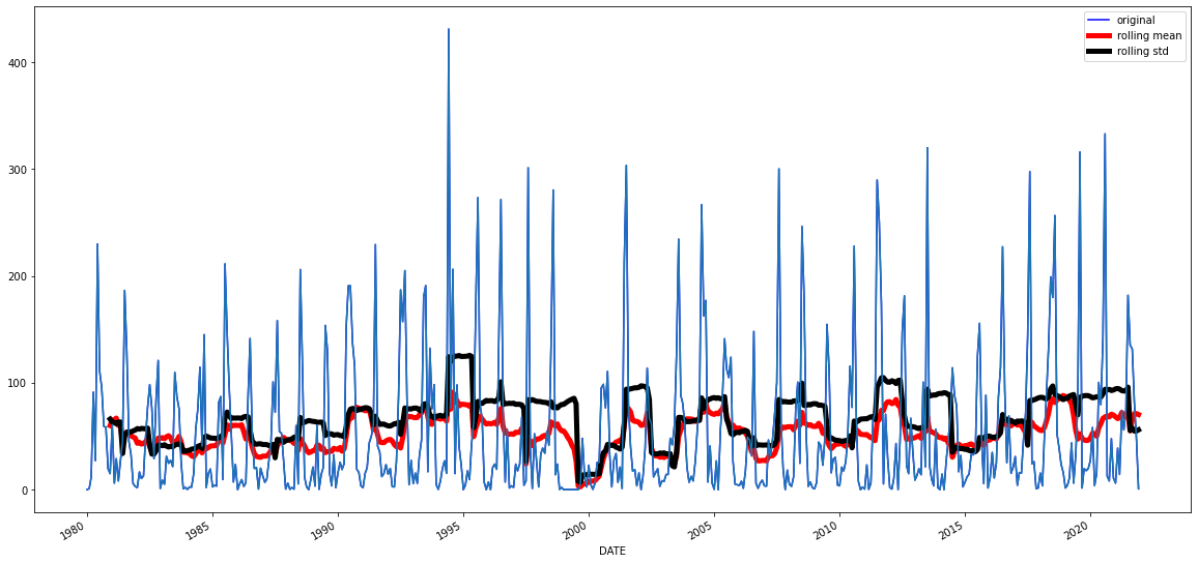

Figure 1. Monthly rainfall data of Weifang from January 1980 to November 2021 (42 years) as well as mean and standard deviation

Methodology

Multiplicative Seasonal ARIMA (SARIMA) Model.

SARIMA model is generally considered as seasonal ARIMA model when the time series follows seasonal effect. The model constructs under consideration with seasonal nature of the series. The general multiplicative Seasonal ARIMA model is expressed as SARIMA (**p**, **d**, **q**) (**P**, **D**, **Q**)$_s$ (Afrifa-Yamoah et al. 2016).

$$\varphi(B)\varphi(B^s)(1-B)^d(1-B^s)^D(Z_t-\mu)=\theta(B)\theta(B^s)\varepsilon_t$$

where $\varphi(B)=1-\varphi_1B-\varphi_2B^2-\cdots-\varphi_pB^p$ (The order p of AR term),

$\theta(B)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$ (The order q of MA term) ,

$\varphi(B^s)=1-\varphi_1B^s-\varphi_2B^{2s}-\cdots-\varphi_PB^{Ps}$(The order P of seasonal AR term) ,

$\theta(B^s)=1-\theta_1B^s-\theta_2B^{2s}-\cdots-\theta_QB^{Qs}$ (Seasonal MA term)

and $\varepsilon_t\sim WN(0,\sigma^2)$, the difference ***d*** non-negative integer. ***s*** is the integer always greater than

one.

Decomposing Time Series.

Time series decomposition reveals the data series is separated into its constituent components, which are generally an irregular component, trend component, seasonal component as well. Figures 2 represents the decomposition plot of rainfall and temperature data series. The figure show the observed time series (top); the estimated irregular or remainder component (second from top); the estimated seasonal component (third from top); the estimated trend component (bottom). The up and down pattern of the observed time series is an indication of the seasonality of both rainfall and temperature data series.

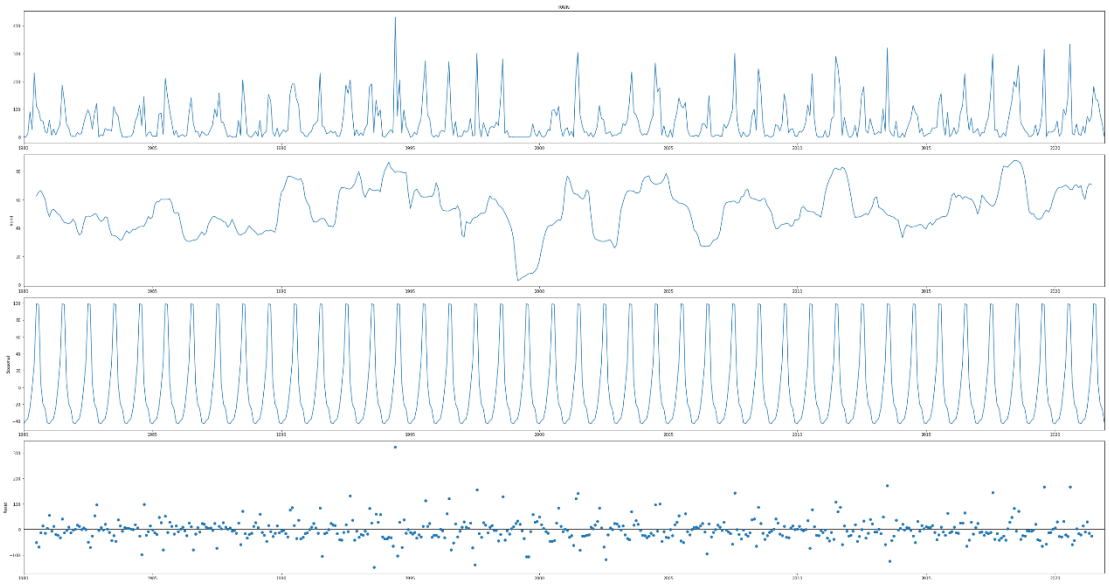

Figure 2. Decomposition of time series.

From Fig. 2, all the rainfall data series has a seasonal effect with the usual upward and downward pattern being experienced yearly over the study period. This indicates that the average monthly rainfall data series of each year was influenced by the seasonal components of all the countries. The rainfall series for Pakistan has larger values of remainder effect (error), hence it is more volatile as compared to the rainfall series in other countries.

Stationarity Test.

For time series modelling, data must be checked of its stationarity. The Augmented Dickey–Fuller (ADF) test and Phillips–Perron (PP) test are applied for testing the stationarity for the data series, From the table, the p value is very less than the significance level of 5% and hence we can make a decision that the data series is stationary.

ADF Statistic	-4.608264
p-value	0.000125

Table 1. Augmented Dickey–Fuller (ADF) test and Phillips–Perron (PP) test result of rainfall dataset

Another method for testing the stationarity of a time series is to determine whether the series is smooth by observing whether the mean and standard deviation of the time series change with time. Observing the results in Figure 1, we can find that the changes of the mean (red line) and standard deviation (black line) of monthly precipitation obviously not shift with time, which indicates that there is no trend effect in the time series, so the series is stationary.

Model Identification and Diagnostic Check.

As the data series is a monthly basis, the significant ACF, PACF plot (Figure 3, 4) confirms at multiple lags of 12. This supports the seasonal differencing with a period of 12 in the data series.

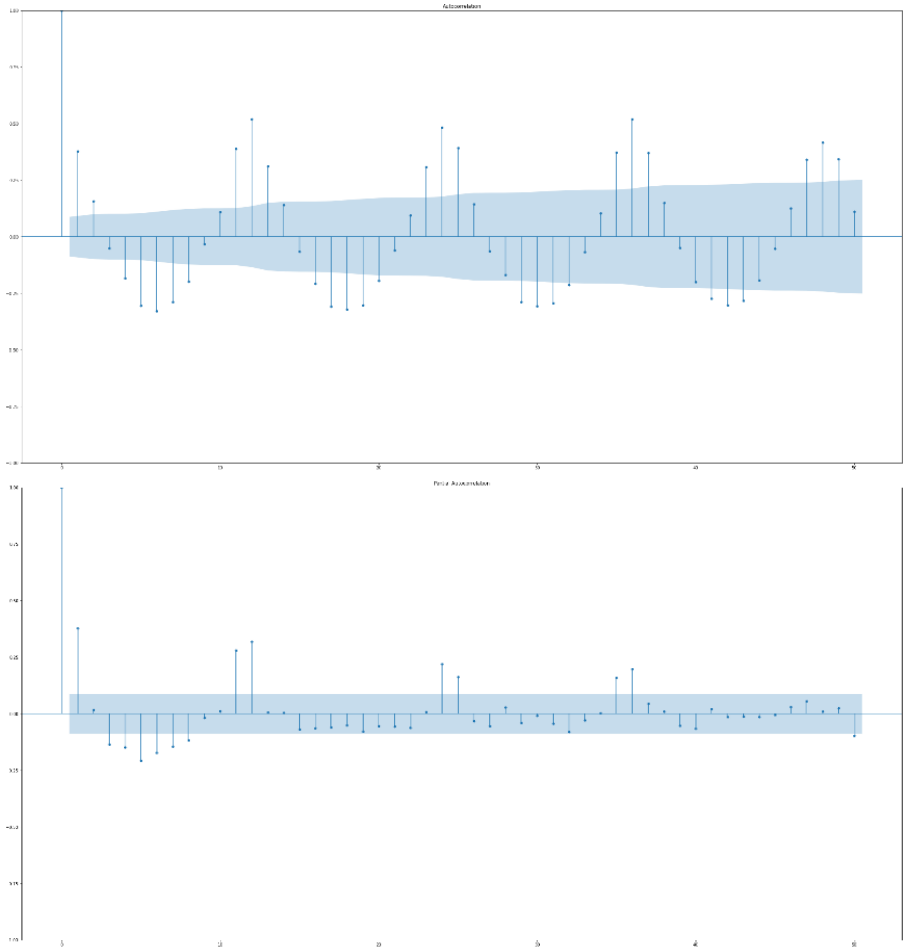

Figure 3. Autocorrelation graph of monthly rainfall dataset

Figure 4. Partial autocorrelation graph of monthly rainfall dataset

The order p and q as well as the seasonal order P and Q of the Seasonal ARIMA (p, d, q) $(P, D, Q)_{12}$ models are identified and estimated through grid search.

The grid search can be traversed to explore different combinations of parameters. For each combination of parameters, a new seasonal ARIMA model can be fitted using the SARIMAX() function of the statsmodels module in Python, and its overall quality can be evaluated. When the grid search traverses the entire parameter environment, we can select the best performing parameters from the parameter set based on the criteria for evaluating time series models.

When evaluating and comparing statistical models with different parameters, each model can be ranked according to the degree of fit to the data or the ability to accurately predict future data points. This report considers the selection of models using the AIC criterion, which is a weighted function of the accuracy of fit and the number of parameters, such that the model with the smallest AIC function is considered the optimal model.

By grid search and based on the AIC criterion, we obtained the lowest AIC value of 3415.0697990988256 for SARIMAX (1, 0, 2) x (4, 2, 5, 12). therefore, we consider this as the optimal choice among all parameter combinations. The results of the optimal parameter combination are shown in the table below.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6302	0.696	0.905	0.365	-0.734	1.995
ma.L1	-0.6160	0.699	-0.881	0.378	-1.987	0.755
ma.L2	0.0473	0.056	0.841	0.400	-0.063	0.158
ar.S.L12	-0.5172	0.132	-3.931	0.000	-0.775	-0.259
ar.S.L24	-0.4857	0.163	-2.987	0.003	-0.804	-0.167
ar.S.L36	-0.4835	0.110	-4.400	0.000	-0.699	-0.268
ar.S.L48	0.0226	0.059	0.384	0.701	-0.093	0.138
ma.S.L12	-1.3436	0.601	-2.235	0.025	-2.522	-0.165
ma.S.L24	0.3206	0.280	1.143	0.253	-0.229	0.870
ma.S.L36	0.3239	0.247	1.313	0.189	-0.160	0.808
ma.S.L48	-0.9012	0.279	-3.229	0.001	-1.448	-0.354
ma.S.L60	0.6094	0.381	1.600	0.110	-0.137	1.356
sigma2	2490.1499	1627.854	1.530	0.126	-700.385	5680.684

Table 2. SARIMAX (1, 0, 2) x (4, 2, 5, 12) Fitting result

Validation.

Model diagnosis is a very important step in fitting a SARIMA model, by which the model is diagnosed to ensure that any assumptions made by the model have not been violated.

Our main concern in model testing is whether the residuals of the model are correlated and are

normally distributed with zero mean. If the SARIMA model residuals are correlated and not normally distributed with zero mean, it indicates that the model can be further improved, and conversely, the model fits well and the model can be considered to have adequately extracted the information from the series. The plot_diagnostic function in Python can quickly generate a model diagnosis and investigate the abnormal behavior of the model. The model diagnosis results are shown in the following figure. From the model diagnostic results, it can be seen that the time series plot of the residuals is basically stable, and the residuals do not fluctuate greatly with time.

From the normal distribution plot in the upper right corner, it can be seen that the residuals of the model are normally distributed. The red KDE line does not deviate from the N(0, 1) line to a large distance. where N (0,1) is a standard normal distribution with mean 0 and standard deviation 1. This is a good indication that the residuals are normally distributed. And the lower left corner of the figure with the Normal Q-Q plot also illustrates that the residuals obey a normal distribution.

The autocorrelation plot of the residuals in the lower right corner of the figure shows that there is no autocorrelation in the residuals, indicating that the residual series is a white noise series.

From this, it can be concluded that the SARIMAX (1, 0, 2) x (4, 2, 5, 12) model is a better fit and can help us understand the original time series data.

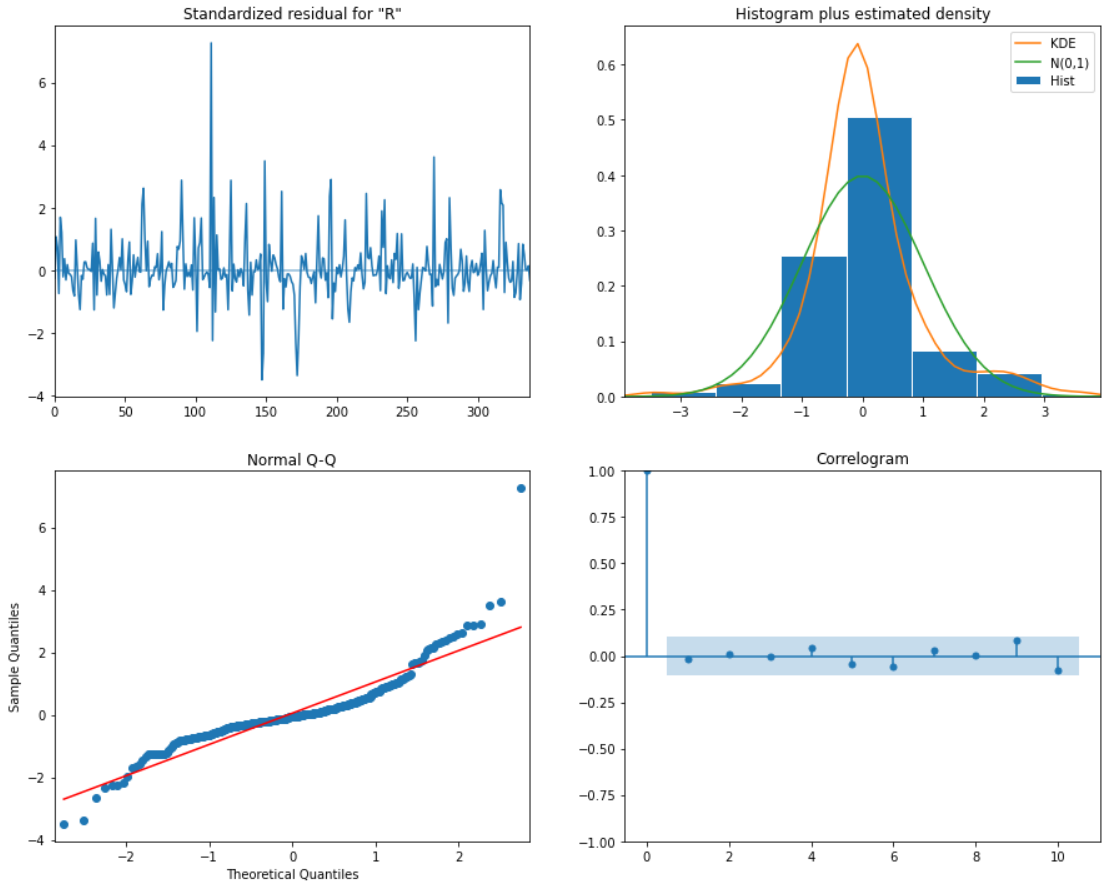

Figure 5. (a) Standardized residual. (b) SARIMAX (1, 0, 2) x (4, 2, 5, 12) model fitting residual distribution. (c) Quantile-quantile graph. (d) Autocorrelation graph of residuals.

Forecasting.

After developing the best fitted time series model, forecasting is carried out for monthly average rainfall of Weifang City. The monthly data from January 2013 to November 2021 are considered for validation of the model can be regarded as in-sample forecast. and the data from January, 2022 to December, 2022 are used as out-sample forecast, satisfied that the residuals of all selected models are found to be approximately stationary and white noise.

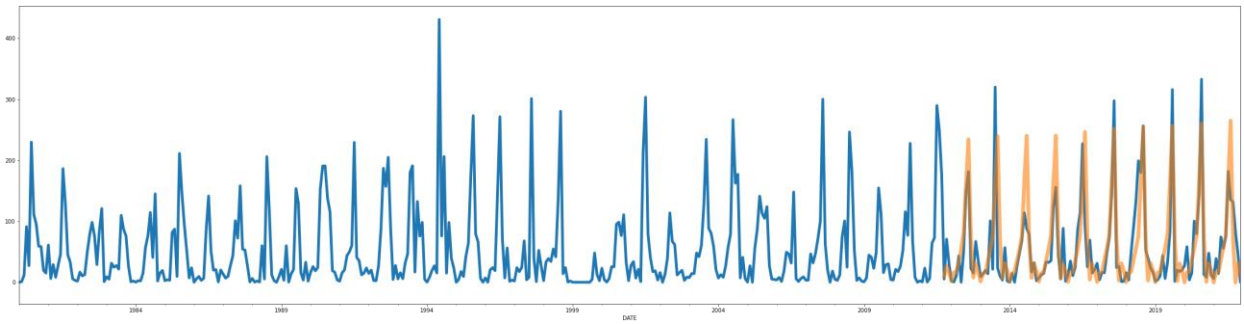

Figure 6. Rainfall forecast from January 2013 to November 2021 compared with history data.

Fig 6 shows the forecasted rainfall from 2013 to 2021 compared with the history rainfall data. Overall, the predicted values of the model match the true values.

2022 January	16.151615	2022 July	234.327912
2022 February	16.317422	2022 August	72.077731
2022 March	32.730454	2022 September	7.894623
2022 April	54.124897	2022 October	29.790282
2022 May	80.698245	2022 November	17.480052
2022 June	160.835977	2022 December	1.457958
Sum	723.8872		

Table 3. predicted monthly rainfall in2022

Table 3 shows the out-sample forecast from January 2022 to December 2022, rainfall amount from the whole year is predicted to be 723.8872 mm.

Conclusion

In this report monthly average rainfall of Weifang city, China have been modelled using Seasonal ARIMA for developing the forecasting model. This study has shown that the data series for rainfall contains three time-series components such as stochastic trend, seasonal and random. The Sen's slope magnitude of rainfall data series has a decreasing trend.

With Pacf, acf plot as well as grid search algorithm, the better parameters combinations of SARIMAX (1, 0, 2) x (4, 2, 5, 12) is selected to forecast rainfall. The in-sample prediction from 2013 to 2021 shows that the forecast basically match the history data. Thus, the out-sample forecast could be reckoned reasonable. The 2022 whole year rainfall is predicted to be 723.8872 mm.

References

1. Ray, Soumik, et al. "Time series SARIMA Modelling and forecasting of monthly rainfall and temperature in the south Asian countries." Earth Systems and Environment (2021): 1-16.
2. Brema, J. "Rainfall trend analysis by Mann-Kendall test for vamanapuram river basin, Kerala." *International Journal of Civil Engineering and Technology* 9.13 (2018): 1549-1556.
3. Hameed, Yasmeen, Bulat Gabidullin, and Darrin Richeson. "Photocatalytic CO2 Reduction with Manganese Complexes Bearing a κ2-PN Ligand: Breaking the α-Diimine Hold on Group 7 Catalysts and Switching Selectivity." *Inorganic chemistry* 57.21 (2018): 13092-13096.
4. Nyatuame, Mexoese, and Sampson K. Agodzo. "Stochastic ARIMA model for annual rainfall and maximum temperature forecasting over Tordzie watershed in Ghana." *Journal of Water and Land Development* (2018).
5. Afrifa-Yamoah, Ebenezer, B. I. Saeed, and Azumah Karim. "Sarima modelling and forecasting of monthly rainfall in the Brong Ahafo Region of Ghana." *World Environment* 6.1 (2016): 1-9.