ARIMA-Based Time Series Forecast of Monthly Rain

Pranay Trivedi

Received August 02, 2024
Accepted September 25, 2024
Electronic access October 15, 2024

The tropical climate of Singapore leads to significant amounts of rainfall, averaging an excess of 2400 mm annually. This rainfall presents challenges: flash floods, mosquito-borne diseases, and infrastructure damage. This study investigates the feasibility of using an Auto Regressive Integrated Moving Average (ARIMA) model to predict monthly rainfall in Singapore. A time series can be created using the publicly available data from the Meteorological Service of Singapore. The ARIMA(1,0,0)(2,0,0) model with a 12-month seasonality, selected based on AIC values, is employed to forecast the next 12 months of rainfall. This time series-based approach, implemented in Python, predicts future rainfall patterns using historical data. The predicted values closely follow the true values with slight discrepancies at certain peaks and portions.

Keywords: ARIMA, SARIMA, Time Series, Singapore, Rainfall

Introduction

Singapore has always been vulnerable to heavy rainfall and rainstorms throughout the year. It is often considered one of the wettest countries on earth, with a rainfall mean of 2,497 mm annually (World Bank, 2020)[1]. The challenges posed to Singaporeans due to this heavy rainfall are vast. Flash floods are a common experience when the rainfall exceeds the capabilities of the city's drainage systems. With negative economic effects due to insurance claims and the looming threat of climate change worsening flooding, predicting rain patterns is an important tool to reduce the negative effects of flooding (Chow, W.T., 2016)[2]. Rain also poses the threat of increased mosquito-borne diseases, namely dengue. Heavy rain, in certain circumstances, can cause a lagged upsurge in dengue cases by creating additional breeding spots for these insects, effectively increasing the environment's carrying capacity (Cheng, Q.)[3]. For these reasons, an accurate forecast of rainfall could be beneficial to both government agencies and residents of Singapore. Using open-source data from the Meteorological Service of Singapore, the study analyzed average rainfall and temperature since 2009 (Meteorological Service Singapore)[4]. The approach taken in this study uses this past data and applies the forecasting method Auto Regressive Integrated Moving Average (ARIMA) to predict future values. ARIMA models are useful in predicting rainfall patterns due to its ability to effectively capture auto correlations in data. The model's strength lies in modeling trends that have seasonal tendencies, making it an apt model for predicting rainfall in the short term. Furthermore, the ARIMA model has been used and proven effective in analyzing and accurately predicting rainfall patterns in regions such as Tamil Nadu, India (Ashwini et al., 2021)[5]. Initially, an Artificial Neural Network (ANN) approach was considered. A study in the International Journal of Photoenergy, which aimed to predict the solar energy harvested by photovoltaic cells, found that ANN and ARIMA models yield comparable outcomes, but the ARIMA model is more efficient and cost-effective computationally making it a better choice in the paper's particular use case (Fara et al., 2021)[6]. Furthermore, ANN requires more complex data as compared to the ARIMA model which requires only previous data on the meteorological value that needs to be predicted. ANN models often make use of a range of meteorological values to forecast for the targeted meteorological value. Due to computational reasons and simpler data required, an ARIMA model was selected for this particular use case. Python has been used due to its versatility throughout the entire study.

Research Methodology

Data Collection

The data used in this study was collected from the Meteorological Service of Singapore(MSS). The data is downloadable in CSV format and provides daily rainfall statistics from a total of 63 weather stations around Singapore. For the purpose of this paper, data from the weather station located in Admiralty, a region in the north of Singapore, was used. Admiralty was used due to its consistent data: other weather stations had periods where data was not available. Singapore is a small island nation and each of the weather stations records similar data with little variation amongst them. Therefore the statistical benefit of integrating data from multiple weather stations is negligible. Once the information was downloaded in CSV format, using a Python script, the daily values for rainfall were averaged into monthly

rainfall data in mm of rainfall. The monthly data ranges from February 2009 until May of 2016. This made the data 88 months long. Due to a lack of weather stations before 2009, previous data could not be accurately sourced. The data was truncated at 2016 to avoid null data points. Furthermore, there may be uncertainties in the data reported by MSS. Such uncertainties may include errors in instruments causing slight under or over reporting of rain; However, for the purpose of this study we will ignore these slight uncertainties and take the data provided as accurate. Since we are under the presumption that the data provided is wholly accurate, the prediction provided through the time series analysis can also be assumed to be based in certainty as long as it meets the prerequisites to qualify for time series analysis, which is described in detail in the forecasting methods section.

Forecasting methods

In order to use the ARIMA model correctly, the study followed certain steps to get the most accurate results.

1. **Splitting the data set:** Testing and Training: For this study, the data set was split into two sets: a training data set and a testing data set. This is a fundamental step when performing ARIMA forecasting. The training set trains the model to ensure it can predict accurate values. Through the training set, the model can learn the underlying patterns from the historical data in the training set (February 2009 to May 2015) to make stronger predictions for future values. The testing set (May 2015 to May 2016), containing the last 12 points in the data set, is used to evaluate the accuracy of the prediction made by the model. This division is optimal for this sort of testing: the training set is large enough for the model to learn, and the testing is long enough to see patterns in the prediction. The model's accuracy can be deduced by comparing the prediction made by the model against the testing set. This can be analyzed visually, by comparing the graphs of the test against the prediction, or by calculating mathematical error metrics such as mean squared error and root mean squared error.

2. **Augmented Dickey-Fuller (ADF) test for stationary:** Through the ADF test, the time series can be tested to check for stationary. For the time series to be stationary, it would have to have a constant mean and constant covariation; the time series must have a constant variance throughout. The ADF test can be employed to determine if the series is stationary. The ADF test employs the idea of a unit root: if a shock or change in the series at a given point persists indefinitely, there is a unit root present. The null hypothesis for the ADF states that the time series has a unit root, making it non-stationary. The alternative hypothesis is that the series does not have a unit root, making the series stationary. If a P-value below 0.05 is presented when this test is run, the series can be assumed to be stationary (R. Mushtaq, 2011)[7].

3. **Arima Model:** The Auto-Regressive Integrated Moving Average (ARIMA) model is a powerful tool for time series forecasting. It can generate highly accurate short-term predictions based on historical data. The model does require several assumptions, including making sure the data is stationary (A. Hendranata, 2003)[8]. The model consists of three fundamental ideas:

    - Auto-regression (AR): This component leverages past observations within the time series to predict future values. The parameter 'p' signifies the number of past observations incorporated into the model.
    - Integration (I): The integration stage employs differencing to eliminate trends from the data to ensure that the data is fit for the model. The parameter 'd' denotes the number of times the data undergoes differencing.
    - Moving Average (MA): The moving average component factors in past prediction errors to refine the forecast. The parameter 'q' represents the size of the window employed to average these errors.

    Furthermore, seasonality can be included and integrated into the ARIMA model to help account for recurring patterns over a set period. Regular ARIMA models can be extended to incorporate seasonality. They would be expressed as ARIMA(p,q,d)(P,Q,D)s. The definition for 'p', 'q', and 's' remain the same as the traditional model definitions explained above. The second set of parameters, the capital variants, signify a parameter that has been adjusted for the seasonality of the data set. 's' in this scenario represents the period. For every 's' time periods, the data's pattern repeats itself. If one was analyzing monthly data, a s value of 12 would be appropriate given that there are 12 months a year (The Pennsylvania State University)[9]. By incorporating seasonality correctly, ARIMA models can provide more accurate forecasts for time series data with recurring patterns.

4. **Parameter selection:** The success of the ARIMA model is heavily dependent on the parameters selected. In order to select the parameters, one possible method is to use the 'auto_arima' function from the Python library 'pmdarima'. This function goes over possible combinations of parameters and gives each an Akaike information criterion (AIC). The parameter combination with the lowest AIC score is considered the most optimal parameter combination for this model (Bozdogan, 1987)[10]. Using the AIC produced through this function, an accurate ARIMA model can be created.

Residual Analysis

In order to assess and validate the fit of the ARIMA model, a method commonly used is to analyze the residuals. Residuals are the difference between the observed values and the values predicted by the ARIMA model. Ideally, the residuals should be random and normally distributed, indicating that the model has captured the underlying nuances and patterns in the data. In time series analysis residuals are traditionally assumed to be normally distributed to achieve the best fit. To test this, the Jarque-Bera test was employed. The null hypothesis states that the residuals are normally distributed, indicating a good fit. The alternative hypothesis states that the residuals are not normally distributed, indicating a suboptimal fit. If the P-value is greater than the significance level of 0.05, you fail to reject the null hypothesis. This means there is not enough data to conclude that the residuals are not normally distributed, indicating that the model is a good fit for the presented data (Brys, Hubert, & Struyf, 2004)[11].

Results

Stationary Testing

The raw data, training and testing set included, can be seen in figure 1 The Augmented Dickey-Fuller (ADF) test was employed to determine if there is a unit root present, indicating whether the series with the training values is stationary or not. Using the 'adfuller' function available from the 'statsmodels' library on Python, the ADF test can be run. The ADF test yields a P-value of $4.606x10^{-}9$. Since the value is less than 0.05, the null hypothesis can be rejected, indicating the absence of a unit root, hence proving that the training data series is stationary.

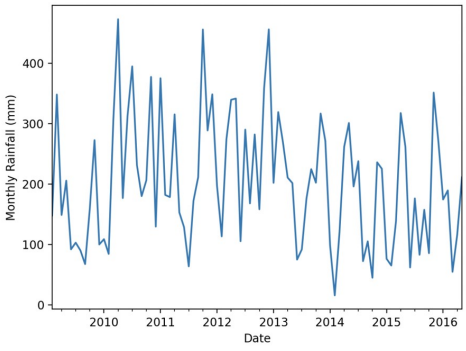

Fig 1: Raw rainfall data obtained from the Meteorological Service of Singapore.

Model Fit

Since the training series data is found to be stationary through the Augmented Dickey-Fuller test, it can be fitted to a model. To fit the model, the Python library 'pmdarima' was used, specifically, the function 'auto_arima'. The function compares possible values and chooses the parameter with the lowest Akaike information criterion (AIC). Given that our data is seasonal, a period of s = 12 is assumed due to the monthly nature of the data. The function's output is shown in table 1.

ARIMA Model	AIC Value
ARIMA(2,0,2)(1,0,1)[12]	1075.076
ARIMA(0,0,0)(0,0,0)[12]	1072.734
ARIMA(1,0,0)(1,0,0)[12]	1069.380
ARIMA(0,0,1)(0,0,1)[12]	1070.689
ARIMA(1,0,0)(0,0,0)[12]	1069.810
ARIMA(1,0,0)(2,0,0)[12]	1068.523
ARIMA(1,0,0)(2,0,1)[12]	1070.743
ARIMA(1,0,0)(1,0,1)[12]	1069.349
ARIMA(0,0,0)(2,0,0)[12]	1070.387
ARIMA(2,0,0)(2,0,0)[12]	1069.300
ARIMA(1,0,1)(2,0,0)[12]	1069.804
ARIMA(0,0,1)(2,0,0)[12]	1069.334
ARIMA(2,0,1)(2,0,0)[12]	1071.260

Table 1 Model fit

With an AIC value of 1068.523, the model ARIMA(1,0,0)(2,0,0)[12] was determined to be the most appropriate model.

Residual Analysis

The residuals can be seen in figure 2. The Jarque-Bera test was implemented to verify whether or not the residuals are normally distributed, and therefore, if the model is a good fit for the data. Using the 'jarque-bera' function from the 'scipy.stats' library, the Jarque-Bera test was conducted. The test yields a P-value of 0.231. Since the P-value is greater than 0.05, there's insufficient evidence to reject the null hypothesis that the residuals are normally distributed. This suggests that the ARIMA model is a suitable fit for the given dataset.

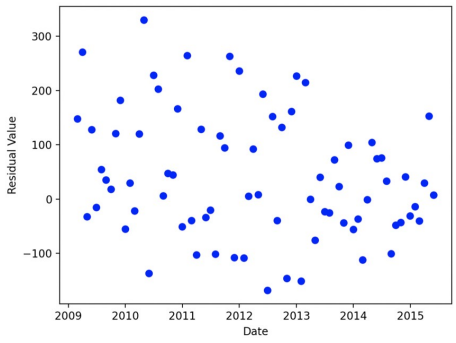

Fig 2: The residual plot of the training set.

Model Forecasting

The SARIMAX model, estimated using the 'statsmodels' library, was fit on the training set. This estimated model was used to generate forecasts for the next 12 months, corresponding to the length of the testing set. The forecast was compared to the actual values in the testing set to assess the model's prediction accuracy as shown in figure 3. The figure shows the raw data with the model's prediction overlaid.

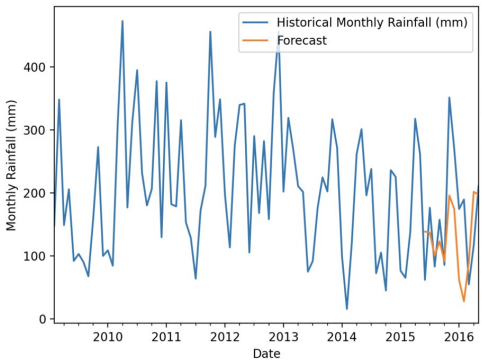

Fig 3: Forecasted Values versus Raw Rainfall Data.

Discussion

Through visual inspection, the forecast generally follows the shape of the test set, with a notable exception in November 2015, where the peaks of the estimates do not reach the actual peaks. This could indicate a downside when applying the model to real-world use cases. The model may underpredict rainfall during peak periods. The model does display some slight limitations with the underestimation of data points and slight lags during certain periods, specifically found during the early months of 2016. Despite this slight lag and peak mismatch, the model successfully captures the overall pattern of the data, including the peaks and troughs. This indicates that, from a graphical perspective, the model provides a good fit for the given use case and data. However, the mean absolute error (MAE) and root mean square (RMSE) error indicate high error. The error values, respectively, are 69.41 and 86.42. These error values are high due to the slight mismatches and lags during the steep months. The closer the RMSE value is to 0, the more accurate it is. Similarly, the closer the MAE value is to 0, the more accurate it is. Since these error metrics revolve around comparing the actual and predicted value at a certain data point, it does not take into account shape and over amplifies the slight differences. Future reiterations could potentially change the parameters in order to reduce the slight mismatches while preserving the forecasting quality for other points of data and reduce the mathematical error values.

Conclusion

The ARIMA-based approach to predicting monthly rainfall for the next 12 months in Singapore has proven to be an effective and accurate model within the bounds of this study. By obtaining public historical weather data from the Meteorological Service Singapore and employing time-series statistics, this study has demonstrated the feasibility of using ARIMA-based models to generate reliable rainfall forecasts even with relatively small data sets. These forecasts are crucial for preparing and responding to heavy rainfall events, which pose significant challenges such as flash floods, economic losses from insurance claims, and increased risks of mosquito-borne diseases like dengue.

The study's methodology included unit root and stationary testing and model fitting to ensure the validity of the ARIMA model generated. The model matched and captured the major trends in the data, despite the slight mismatches between certain data points in the predicted and testing values. The appropriateness of this model indicates the potential positive impact it could have for practical applications in government and residential planning.

Further research on this topic could attempt to refine the model's parameters with larger data sets in order for it to match the peaks and get rid of the slight lags, making the model more useful in practical applications. Furthermore, integrating multiple time series from different meteorological values, such as temperature and humidity, would also aid in getting a more nuanced rainfall prediction. Nonetheless, the current ARIMA(1,0,0)(2,0,0)[12] model serves as a strong and valuable tool for forecasting rainfall in Singapore.

References

1 W. Bank, *Average precipitation in depth*, https://data.worldbank.org/indicator/AG.LND.PRCP.MM?most_recent_value_desc=true&skipRedirection=true&view=map.

2 W. Chow, B. Cheong and B. Ho, *Advances in Meteorology*, 1–11.

3 Q. Cheng, Q. Jing, P. Collender, J. Head, Q. Li, H. Yu, Z. Li, Y. Ju, T. Chen, P. Wang, E. Cleary and S. Lai, *Prior water availability modifies the effect of heavy rainfall on dengue transmission: a time series analysis of passive surveillance data from southern China*, https://doi.org/10.21203/rs.3.rs-3302421/v1, Research Square (Research Square).

4 M. S. Singapore, *Historical Daily Records*, http://www.weather.gov.sg/climate-historical-daily.

5 U. Ashwini, K. Kalaivani, K. Ulagapriya and A. Saritha, 6th International Conference on Inventive Computation Technologies (ICICT.

6 L. Fara, A. Diaconu, D. Craciunescu and S. Fara, *International Journal of Photoenergy*, 1–19.

7 R. Mushtaq, *Augmented dickey fuller test*.

8 A. Hendranata, *ARIMA*, Autoregressive Integrated Moving Average),"2003.

9 T. P. S. University, *4.1 seasonal arima models: Stat 510*, Statistics Online Courses, PennState.

10 H. Bozdogan, *Psychometrika*, **52**, 345–370.

11 G. Brys, M. Hubert and A. Struyf, *A robustification of the Jarque-Bera test of normality*.