Mathematical Statistician and Engineering Applications

ISSN: 2094-0343

2326-9865

Study to Forecast Drought Using SARIMA Model

[1]Dhruvendra Kumar Chourishi

Department of Computer Sc. and Engg. Govt. Women's poly College Bhopal, Bhopal M.P. India

dhruve23@yahoo.com

[2]Dr. Anil Rajput

OSD, Higher Education Department, Govt. of Madhya Pradesh, Bhopal M.P India

. dranilrajput@hotmail.com

[3]Dr. Sanjeev Gour

Department of Computer Sc, Career College Bhopal, Bhopal M.P India

sunj129@gmail.com

[4]Kshmasheel Mishra

Reader, Institute of Computer Science, Vikram University , Ujjain M.P, India

Article Info
Page Number:11157 - 11167
Publication Issue:
Vol 71 No. 4 (2022)

Article History
Article Received: 15 September 2022
Revised: 25 October 2022
Accepted: 14 November 2022
Publication: 21 December 2022

Abstract:

Drought is an unpredictable disaster with a cumulative nature. Its occurrence, persistence, end, and severity cannot be characterized or formulated through any single parameter. Drought is a complex phenomenon, and nobody can predict its onset, persistence, severity, or end accurately.

In our research area of Madhya Pradesh, 90% of the rainfall occurs due to the southwest Monsoon, but irregular, low rain, and early withdrawal of monsoon can cause drought-like conditions in the area. This drought problem is more severe in arid and semi-arid zones like the Bundelkhand area.[5] The frequent occurrence of dry conditions in a particular region has an adverse impact on many sectors, especially agriculture.

Several models and indices have been developed to understand drought characteristics, but no single model is entirely effective at predicting them accurately. SARIMA (Seasonal ARIMA) is one of the most powerful and popular statistical models used to predict monthly time series precipitation data.

KEYWORDS: Precipitation, Time Series Analysis, SARIMA Model, Forecasting.

---

1. Introduction:

Droughts are one of the major natural disasters, caused by abnormally dry conditions that persist for a long time in a particular area of land. Many countries around the world, including China, Africa, and the US, are affected by this problem, but those with economies that rely on agriculture are particularly vulnerable. The severity of droughts is increasing day by day, leading to adverse effects on the environment, economy, and society.

However, predicting droughts is complex and difficult. No real-time monitoring system has been developed to forecast them accurately. Weather conditions cannot be monitored by simply analyzing observed data. Accurate and timely measurements are necessary for predicting and forecasting weather. While many parameters must be considered to monitor drought, we will focus on rainfall, the primary source of water in Madhya Pradesh, which occurs due to the southwest and northeast monsoons. For this study, we will use monthly precipitation data from the last 40 years, from 1981 to 2021.

1.1 Study area: In this research work, we focused on the geographical area of Madhya Pradesh, which is located between latitude 22.9734° N and longitude 78.6569° E. Specifically, we selected some districts in the Bundelkhand area, including Chhatarpur, Sagar, Panna, Datia, Damoh, and Tikamgarh. Despite being situated in the heart of the country, Madhya Pradesh has no coastline, and many areas within the state suffer from water scarcity, which is the lack of sufficient available water resources to meet the requirements of the region. The primary sources of water in Madhya Pradesh are rivers, lakes, and wells, which depend on natural rainwater, as there are no other major sources of water. There are three main seasons in Madhya Pradesh: summer, winter, and monsoon.

Although the history of drought in Madhya Pradesh is not well documented, we will discuss some of the major historical drought events that have occurred in various districts of the state. The primary monsoon

season for rainfall in Madhya Pradesh is from June to September. Yearly consolidated data from 2011 to 2014 shows that there is great variation in rainfall for almost every year, and this uncertainty or untimely monsoon changes many parameters of drought monitoring. An analysis of the rainfall data for various years reveals that there have been significant variations in rainfall in Madhya Pradesh.

The average temperature on Earth is about 61^{o}F (16^{o}C), but this temperature varies depending on the season, climatic conditions, vegetation, forestation, and location (latitude and longitude), among other factors. In Madhya Pradesh, there has been very little variation in the maximum and minimum temperature from 1984 to 2013. However, the limited resources of water in the state cannot keep pace with the growing population and industries, which has led to a decline in groundwater levels.

The objective of this study is to describe and oversee the current drought prediction system, enhance the Autoregressive Integrated Moving Average (ARIMA) models, and produce forecasts utilizing seasonality. Specifically, the study aims to utilize the SARIMA (Seasonal ARIMA) model to improve drought forecasting accuracy [4]. A problem with the ARIMA model is that it supports non seasonal data so it is to be adjusted through seasonal differencing.

The ARIMA model is a commonly used statistical model for forecasting linear time series data. [1] Box and Jenkins are widely credited with popularizing and developing the ARIMA (Autoregressive Integrated Moving Average) model. In fact, the ARIMA method is often referred to as the Box-Jenkins model. Box and Tiao discussed the general transfer function model that is utilized by the ARIMA procedure (1975) [1]. When the ARIMA model is augmented with additional time series as input variables, it is commonly referred to as the ARIMAX (Autoregressive Integrated Moving Average with Explanatory Variables) model. [6] In Pankratz's (1991) work, the ARIMAX model is referred to as dynamic regression. This terminology reflects the fact that the model combines the dynamics of both auto regression and moving average models with the explanatory power of regression models. [1] The ARIMA model procedure is highly flexible in terms of identifying univariate time series models, estimating model parameters, and producing forecasts. This flexibility enables the ARIMA method to be customized to fit a diverse range of time series data, making it a versatile and powerful tool for time series analysis and prediction.

2. Methodology:

2.1 Data: In order to implement the SARIMA Model in our research area we have supplied the historical time series data set of monthly precipitation in our model. In this, we have taken the data from 1981 to 2021 of all our six research areas of Bundelkhand in Madhya Pradesh. Nowadays many finance, marketing, and research-based forecasting are being performed through time series models. The ARIMA model is a highly popular time series model due to its ability to effectively model and capture trends and patterns in time series data. [2].

In order to apply the SARIMA model to time series data, the initial step is to transform the non-stationary time series into a stationary one, prior to analysis.[8,9]. To confirm the stationarity of the data using the Dickey-Fuller test, we perform a hypothesis test where the null hypothesis assumes the data is non-stationary.[7]

Station	**ADF test statics**	**p- value**	**legs used**	**Number of Observation**
Chhatarpur	-9.053054935	4.78E-15	13	418
Damoh	-5.295232717	5.60E-06	14	429
Datia	-8.173004282	8.52E-13	14	429
Panna	-5.001927316	2.20E-05	13	418
Sagar	-5.001602689	2.20E-05	14	429
Tikamgarh	-4.780932551	5.91E-05	13	430

Table 1 : Parameters to determine the stationarity of a time series

The ideal p-value in a SARIMA model should be less than 0.05, which indicates that the null hypothesis can be rejected at a 95% confidence level, and the relationship between the dependent variable and independent variables is statistically significant.

For the ADF test, if the calculated test statistic is less than the critical value at a given significance level, typically 0.05 or 0.01, then the null hypothesis of non-stationarity is rejected, and the series is considered stationary. Thus, the ideal ADF test result is a significant rejection of the null hypothesis, indicating that the series is stationary.

2.2 Seasonal ARIMA Model

Auto regression Model

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \emptyset_3 y_{t-3} \ldots\ldots\ldots\ldots\ldots + \emptyset_p y_{t-p} + \varepsilon_t$$

y_t *depends only on its past values* $y_{t-1}, y_{t-2}, y_{t-3}$ *etc.*

Moving average model

$$y_t = c + \emptyset_1 \varepsilon_{t-1} + \emptyset_1 \varepsilon_{t-2} + \emptyset_1 \varepsilon_{t-3} \ldots\ldots\ldots\ldots + \emptyset_1 \varepsilon_{t-q} + \varepsilon_t$$

$\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}$. error term assume to be the noise process with mean zero

ARIMA Model

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \emptyset_3 y_{t-3} \ldots\ldots\ldots\ldots + \emptyset_p y_{t-p} + \emptyset_1 \varepsilon_{t-1} + \emptyset_1 \varepsilon_{t-2} + \emptyset_1 \varepsilon_{t-3} \ldots\ldots\ldots\ldots + \emptyset_1 \varepsilon_{t-q} + \varepsilon_t$$

The ARIMA model is represented by the notation (p,d,q), where p corresponds to the number of autoregressive terms in the non-seasonal component of the dataset, d denotes the number of differences required to achieve stationarity in the non-seasonal component of the dataset, and q represents the number of moving average terms in the non-seasonal component of the dataset.

[3] In the ARIMA model, the parameters p, d, and q are non-negative integers

SARIMA model is simply an ARIMA model with the added feature of seasonality,

By incorporating seasonality components, the ARIMA model is able to effectively capture and forecast complex time series data

While in SARIMA model (P, D, Q)S, where P, D, and Q are the seasonal components of the model, and S is the time span of the repeating seasonal pattern. These seasonal components are included in the

ARIMA model to account for periodic fluctuations that occur over fixed intervals, such as daily, weekly, or monthly cycles. By incorporating these seasonal components, the ARIMA model is better equipped to capture and forecast time series data with recurring patterns.

2.3 Seasonal Components of Model

Here's a some parameter of SARIMA model are arL1,arL2, maL1, arS.L12, ar.S.L24 these parameters decides goodness of model ar.L1 refers to the effect of the previous observation, and ar.L2 refers to the effect of the observation two time periods ago. The larger these parameters, the greater the influence of past observations on the current one. The coefficient of ar.L1 and for all the six stations suggests a positive relationship between the current observation and the previous observation." and has a statistically significant effect of different level at current observation .

ma.L1: This is a moving average (MA) parameter representing the effect of past errors on the current observation. The larger this parameter, the more weight is given to past errors, and the smoother the time series is likely to be. ma.L1 is not statistically significant at the 5% level, meaning it may not have a significant effect on the current observation

ar.S.L12 and ar.S.L24: These are seasonal autoregressive parameters, similar to ar.L1 and ar.L2, but for seasonal patterns. ar.S.L12 represents the effect of the observation 12 time periods ago (i.e., the same month in the previous year), and ar.S.L24 represents the effect of the observation 24 time periods ago (i.e., two years ago in the same month). ma.S.L12 and ma.S.L24 is highly statistically significant at the 0.1% level, meaning it has a very significant effect on the current observation.

2.4 Model Identification

Akaike's Information Criterion (AIC) is a useful tool for both selecting an appropriate statistical model and determining the order of an ARIMA model. It measures the goodness of fit of a model while taking into account the complexity of the model. When comparing multiple models, the model with the lowest AIC value is typically chosen as the best-fitting model for the given data.

$$AIC = -2\log(L)+2(p+q+k+1)$$

where L is the likelihood of data k=1, if $c \neq 0$, k=0 if $c=0$

and BIC is

$$BIC = AIC + [\log(T) - 2](p+q+k+1)$$

The Hannan-Quinn information criterion (HQIC) is a commonly used method for model selection in the ARIMA modeling framework

HQIC is used to select the optimal ARIMA model by taking into account the trade-off between model fit and model complexity

It is given as

$$HQC = -2\,Lmax + 2k\ln(\ln(n))$$

Where Lmax is the log-likelihood, k is the number of parameters and n is the number of observation

The optimal model is determined by selecting the model with the minimum values of AIC, BIC, and HQIC. The values for each of the criteria are obtained by fitting different ARIMA models to the time series data and comparing their performance. The model with the lowest combined value of AIC, BIC, and HQIC is considered the best fit for the data.

Station	**Models**	**AIC**	**BIC**	**HQIC**	**Number of observations**
Tikamgarh	SARIMAX(1,1,1)x(1,1,1,12)	4793.459	4813.79	4801.487	444
	SARIMAX(1,1,2)X(1,1,2,12)	4803.533	4831.996	4814.771	444
	SARIMAX(2,1,1)X(2,1,1,12)	**4595.693**	4624.156	4606.931	444
					0
Sagar	SARIMAX(1,1,1)x(1,1,1,12)	4966.758	4987.089	4974.785	444
	SARIMAX(1,1,2)X(1,1,2,12)	**4957.758**	4985.89	4968.665	444
	SARIMAX(2,1,1)X(2,1,1,12)	4979.767	5008.23	4991.005	444
					0
Datia	SARIMAX(1,1,1)x(1,1,1,12)	4592.163	4612.494	4600.19	444
	SARIMAX(1,1,2)X(1,1,2,12)	4593.535	4621.998	4604.773	444

	SARIMAX(2,1,1)X(2,1,1,12)	**4595.693**	4624.156	4606.931	444
					0
Chhatarpur	SARIMAX(1,1,1)x(1,1,1,12)	4791.1	4811.431	4794.128	444
	SARIMAX(1,1,2)X(1,1,2,12)	**4784.846**	4813.309	4796.084	444
	SARIMAX(2,1,1)X(2,1,1,12)	4791.569	4820.032	4802.807	444
					0
Damoh	SARIMAX(1,1,1)x(1,1,1,12)	4934.225	4954.556	4942.253	444
	SARIMAX(1,1,2)X(1,1,2,12)	**4919.551**	4947.103	4930.341	444
	SARIMAX(2,1,1)X(2,1,1,12)	4945.002	4973.465	4956.241	444
					0
Panna	SARIMAX(1,1,1)x(1,1,1,12)	4848.014	4868.345	4856.042	444
	SARIMAX(1,1,2)X(1,1,2,12)	**4844.364**	4872.827	4855.602	444
	SARIMAX(2,1,1)X(2,1,1,12)	4845.748	4874.211	4856.986	444

Table 2 : AIC, BIC, and HQIC are measures used to select the best model for research area

2.5 Fitting and Prediction graph of research area

SERIMA Model (1,1,1)x(1,1,1,12) Forecasting Result

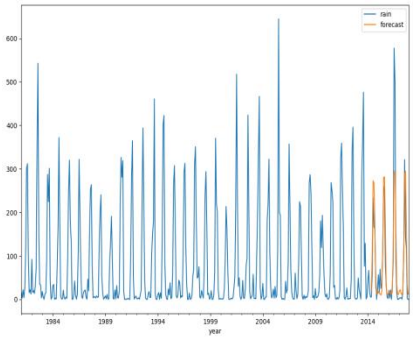

Fig -1 Chhatarpur

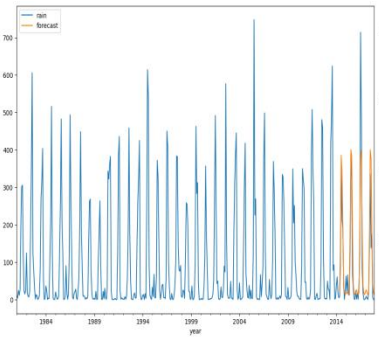

Fig-2 Damoh

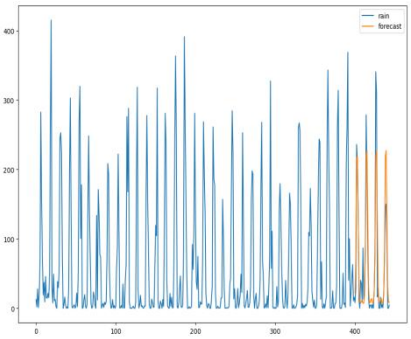

Fig-3 Datia

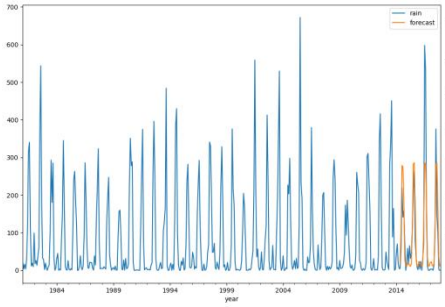

Fig-4 Panna

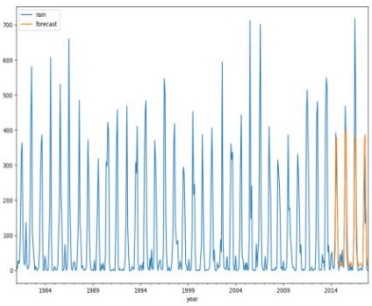

Fig-5 Sagar

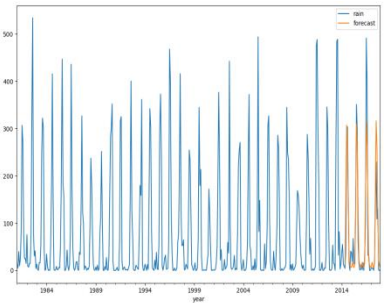

Fig-6 Tikamgarh

Approximately 20% data for testing and 80 % data as sample

3. Results and Discussion:

A seasonal ARIMA model with orders (1,1,1)x(0,1,1,12) was developed. The p-value of the AR(1) coefficient was less than 0.05, indicating its statistical significance, and the same was true for the MA(1) coefficient. The p-value of the seasonal AR(12) coefficient is greater than 0.05, indicating its non-significance. The p-value of the seasonal MA(12) coefficient is less than 0.05, indicating its statistical significance.

Several estimation models were developed, and they were almost similar to the previous model. However, this model had a better fit, which can be attributed to the numerical optimization used. This model provided excellent results using all three criteria (AIC, BIC, and HQIC) for all six research stations.

4. Conclusion:

In this study of seasonal ARIMA models for predicting drought occurrence in different areas of the Bundelkhand region, time series data from the past 40 years was utilized. To assess the accuracy and

quality of the models, three major criteria, namely AIC, BIC, and HQIC, were tested for all stations. The model with the minimum value of these criteria was deemed the best fit for the data.

Several SARIMA models were evaluated to determine the best fit, and it was concluded that the SARIMA (1, 1, 1)x(1, 1, 1,12) model was the most appropriate for the data, as it accurately captured the underlying patterns and fluctuations in the time series.

The predicted models indicate that drought occurrence is expected to decrease in Chhatarpur and Panna, while it is likely to increase in Damoh, Datia, Sagar, and Tikamgarh. Table-1 shows that the value of AIC, BIC, and HQIC for Tikamgarh,

Datia, Chhatarpur, Damoh, and Panna are minimum, indicating that these are good models, except for Sagar. Despite the inconsistent frequency and distribution of rainfall, the region faces the challenge of sustaining crop yield due to a decreasing and erratic trend in annual rainfall. This pattern is expected to continue in the coming years.

Further work is required to evaluate and apply other forecasting models and methods to obtain better accuracy in forecasting results.

Reference:

1. Box, George EP, and George C. Tiao. "Intervention analysis with to economic and environmental problems. " Journal of the American Statistical Association 70.349 (1975): 70-79.
2. Box G.E.P., Jenkins G.M.: Time Series Analysis Forecasting and Control, 525 pp. HoldenDay, San Francisco(1976)
3. Cryer, J. D. and K.S. Chan,. Time Series Analysis with Application in R. 2nd Edn., Springer, New York, ISBN-10: 0387759581(2008), pp: 454.
4. Guo, Z.W.. The adjustment method and research progress are based on the ARIMA model. Chinese J. Hosp. Stat., 161(2009): 65-69.
5. Han, P., P.X. Wang and Y.J. Wang,. Drought forecasting based on the standardized precipitation index at different temporal scales using ARIMA models. Agric. Res. Arid Areas, 26(2008): 212-218.
6. Pankratz, A. Forecasting with Dynamic Regression models, Wiley Interscience, (1991)

7. Stoffer, D.S. and R.H. Dhumway,. Time Series Analysis and its Application. 3rd Edn., Springer, New York, ISBN-10: 1441978658(2010), pp: 596.
8. Wang, J., Y.H. Du and X.T. Zhang,. Theory and Application with Seasonal Time Series. 1st Edn., Nankai University Press, Chinese. (2008)
9. Wang, Y.,. Applied Time Series Analysis. 1st Edn., China Renmin University Press, Beijing(2008)