# A COMPARISON OF ARIMA AND RANDOM FOREST TIME SERIES MODELS FOR URBAN DROUGHT PREDICTION

Ninoslava Tihi[1]*,
[0009-0004-8009-8120]
Srđan Popov[2]
[0000-0003-1215-3111]

[1]The Higher Education Technical School of Professional Studies in Novi Sad, Novi Sad, Serbia

[2]Faculty of Technical Sciences, Novi Sad, Serbia

Abstract:
One of the most devastating natural hazards that can have huge impact on ecosystems, agriculture, water supply and society as a whole is drought. Unlike hydrological and agricultural droughts, urban drought mainly affects populated cities and towns and represents a huge challenge for authorities in terms of managing public health and water supply. Therefore, forecasting urban drought is of great importance. This papers aims to present methodology of data collecting, preprocessing and forecasting meteorological parameters using ARIMA and Random Forest time series models and comparing them based on certain metrics in order to find the best prediction model. Our findings based on RMSE metrics used for evaluation of model accuracy, suggest that ARIMA model outperforms Random Forest model and therefore it is selected as the best model for urban drough prediction.

Keywords:
SARIMA, Random Forest, R, Univariate Time Series, Drought Prediction.

## INTRODUCTION

Urban drought can be classified as one of the very detrimental catastrophic events and its conseqences can be devastating. Drought impacts all the areas of society from health, agriculture to economy, energy and the environment. Around 55 million people are directly affected by droughts every year worldwide and almost 700 million are at risk of being displaced due to drought consequences by 2030 [1]. Therefore, proper water plans and careful monitoring is needed in order to adequately manage and conserve precious water resources. Forecasting urban drought can be beneficial for municipalities and water management agencies in order to make adequate decisions to prevent, prepare and respond in timely manner with appropriate measures.

Drought prediction can be very challenging and it depends on various important factors such as availability and quality of the data, accuracy level and usage of drought prediction model [2]. There are many machine learning models that are being used for time series predictions and each one of them has its own advantages and disadvantages.

Correspondence:
Ninoslava Tihi

e-mail:
tihi@vtsns.edu.rs



DOI: 10.15308/Sinteza-2024-51-56

Selection of the best model depends on various factors like climate changes, regional characteristics of the drought, nature of the time series, machine learning algorithm that is being used and many others. Therefore, when choosing a best model for drought prediction, we need to carefully consider several models before choosing the best one according to our needs. In this paper, we used three of the most well-known and commonly used models for predicting time series, namely ARIMA and Random Forest models. They have not been chosen only for their popularity but also for their accuracy and reliability.

## 2. TIME SERIES FORECASTING

Time series forecasting has a numerous applications in various industries. It is one of the most used technique in fields like business, finance, healthcare, retail, social studies forecasting, pattern recognition, weather forecasting and many others. Time series forecasting can be defined as a process of analyzing time series by using various statistics methods and then modeling data by using certain models. In other words, forecasting involves predicting future values over a certain period of time based on analyzing the trends of historical data. Time series can be seen as a set of data points collected over certain period of time.

Various drought prediction models can be divided into four categories [2]:

1. Stohastic;
2. Physical;
3. Machine learning; and
4. Deep learning models.

All these models have certain advantages over others so selecting the most suitable model from the vast range can be difficult and challening. Furthermore, it is also possible to overlook the ideal model if you are not familiar with some less known and used models. So in order to find the best one, you need to consider the nature of the problem, unique requirements, the type of your data and the model you want to use. It is best to try different suitable models, compare them using certain metrics and then choose the best one or use hybrid model which combines the strengths of both models.

### 2.1. STOHASTIC MODELS

One of the stohastic models we used in this research is very popular linear model called AutoRegressive Integrated Moving Average (ARIMA) and its seasonal version called Seasonal AutoRegressive Integrated Moving Average model. Both models are quite simmilar and very effective for predictiong drought. ARIMA model consists of three main components. First component is called autoregressive (AR) and it uses linear regression to model the relationship between past and current values. Second one is called integrated component (I) which is used for differencing. Differencing is a process of transforming data from non-stationary to stationary. Stationarity describes certain statistical properties of the time series. If the time series is stationary, it means that its mean and variance do not change over time. Standard notation for non-seasonal ARIMA model is ARIMA (p,d,q) where each of these mentioned parameters can have integer values of 0 or 1. Value 1 indicates that the certain component is used in the modeland value 0 means that the model component is not used. SARIMA model is used for seasonal time series and besides three ordinary components it has also three seasonal components. Its notation is SARIMA (p,d,q)(P,D,Q)s where the parameter s represents the length of the seasonal cycle [3].

### 2.2. MACHINE LEARNING MODELS

Another model we used is Random Forest which is a very popular machine learning model. These models are very good for identifying complex patterns in order to predict future drought. Random forest belongs to a group of supervised machine learning models that uses regression for predicting numerical values. This model is a type of ensemble learning method that combines set of decision trees to make accurate prediction. The final output of the model is done by aggregating predictions from individual decision trees. More decision trees prevents overfiting which is a common problem in building a forecasting model and it also leads to higher accuracy of the model.

## 3. METHODOLOGY

### 3.1. SOFTWARE

For the research purposes, we used R programming language since it is very popular statistical programming language used for all kinds of statistical computation, data preparation, transformation and visualization. R consists of many packages which can be downloaded from R archive network called CRAN (http://cran.r-project.org). As an IDE for R, we used RStudio (version 2021.09.0 Build 351) [4].

### 3.2. DATA SOURCE

Data was collected from various sensors which belong to the type of network called "Proactive Network". This sensor station is located in the municipality of Novi Sad. Data was acquired from January 2014 to December 2018. All the sensors are divided into several categories according to the parameters they are measuring. There are six sensor categories which are measuring the weather parameters and one extra category for measuring the state of the battery power [5]:

1. Soil moisture sensors (SM1, SM2, SM3, SM4, SM5, SM6);
2. Humidity sensor (AH1);
3. Wind speed sensor (WS1);
4. Wind direction sensor (WD1);
5. Air temperature sensor (AT1);
6. Precipitation sensor (PP1) and
7. Battery power state sensor.

Recordings are made hourly so in total there are 288 rcordings per day. All the recorded data is being sent to the web portal where it can be downloaded in CSV format. Figure 1 shows the web portal where the data was downloaded from.

Data set consists of four atrributes. The first represents the time and date of the measured valueand the second one is identification number of the device where the parameter is measured. The third one is a sensor identifier and the last represents the actual value of the parameter. The structure of the data set is shown in Table 1 [6].

### 3.3. DATA PREPARATION

Raw data is downloaded from the web portal in CSV format, and loaded in RStudio. Then, it is converted and stored in a data frame which is a common data structure in R programming that organizes data into tabular form.

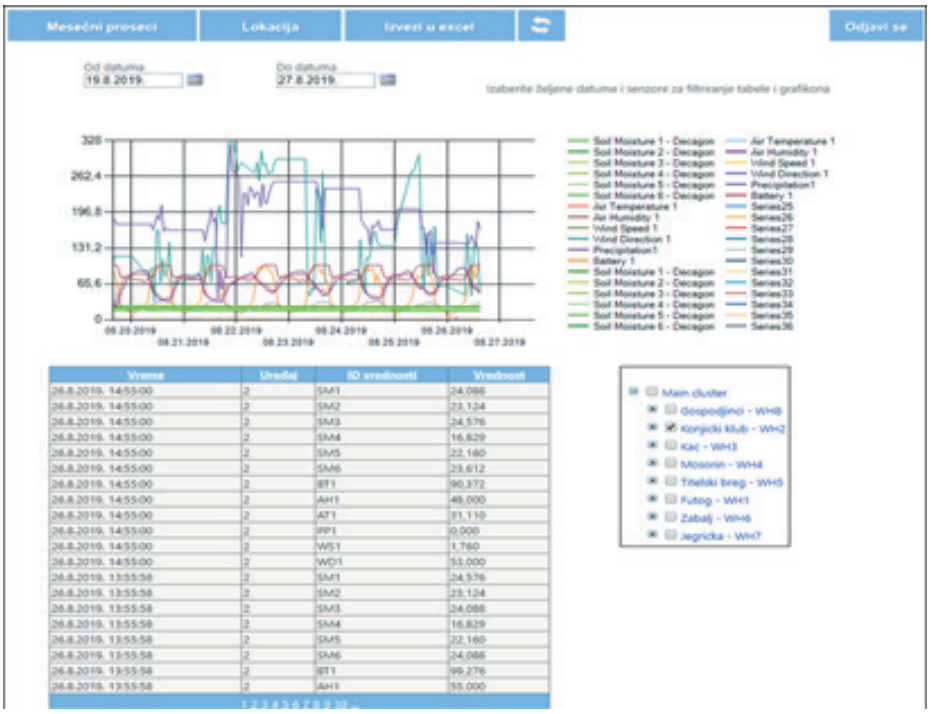


**Figure 2.** MacLeamy BIM Curve.

**Table 1.** Structure of the data set.

| Attribute name | Description |
|---|---|
| Time and date | Time and data of the measured parameter |
| Device | Identification device number |
| ID | Sensor identifier value |
| Value | Measured parameter value |

This data structure consists of rows and columns and it is mostly used in data analytics. Next step is to filter the data so that only relevant parameters are being used. There are several parameters which can indicate the severity and duration of a drought, but we chose only three relevant ones for this research such as air temperature, soil moisture and precipitation. The new tables are created, one for each time series. Tables are called "Precipitation", "Soil Moisture" and "Temperature". Since the granularity of the data is very low, we had to use aggregation functions to calculate mean values for each of the time series. So, air temperature and soil moisture time series are aggregated using the "mean " function in order to calculate mean values. We calculate the mean values for air temperature and soil moisture because the sensors measure these values 24 times a day and using the sum function would show unrealistically high values. On the other hand, the nature of precipitation accumulation process requires using "sum" function in order to calculate the total amount in a day. After being aggregated, data are being sorted by months and years and transformed into time series object. The final step of data preparation process is to divide the data set into two subsets, one for building a model and the other for model validation. Usually, 80% of the original data set is taken for building a model and the rest is used for model verification [7].

### 3.4 DATA MODELING

Data modeling involves careful prediction technique selection based on the nature of the time series and the chahracterstics of the chosen algorithm. The most popular techniques used for time series with a seasonal component are SARIMA and RandomForest methods. They will be used only on Temperature time series as an example of model building. Function called "arima" is used for building SARIMA model and function "randomForest" is used for building DeepForest model. Both functions are part of forecast package which needs to be installed and loaded first.

### 3.5. PERFORMANCE METRICS

Performance metrics is important part of selecting the best model. It is used to evaluate the accuracy and performance of both models. There are dozens of metrics that can be used and each one of them provide certain information about the model performance. Regression models use metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and others. For this research purposes, RMSE metrics was used. It can be mathematically represented as:

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\hat{y}_i\right)\right)^2}$$

**Equation 1.** Root Mean Squared Error equation

Root Mean Squared Error is a square root of the mean squared error and it measures the square root f the average of the squared difference between the target value and the predicted value. It addresses some of the disadvantages in MSE.

## 4. RESULTS AND DISCUSSION

The graphical representation of average values of Temperature time series is shown in Figure 2. Average temperatures are shown for the period from January 2014 to December 2018 on Y-axis. Figure shows that there are variations and certain oscillations in temperature based on season and the lowest temperature was in January 2017.

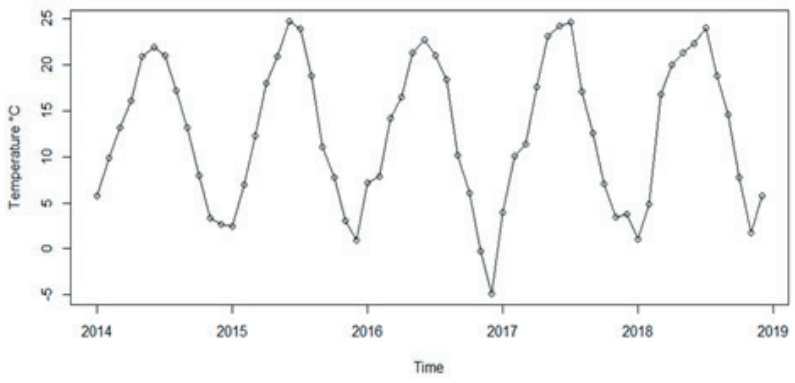


**Figure 2.** Graphical representation of Temperature time series.

### 4.1. SARIMA METHOD RESULTS

Choosing the ideal parameters for SARIMA model requires certain experience and expertise. In order to find the optimal combinations of ordinary and seasonal parameters, funcion "auto.arima" was used. This functions is a part of forecast package and it returns the best SARIMA model according to certain information criteria. It automates the model parameters selection process by searching over range of possible models based on certain constraints provided. After applying auto.arima function, it is found that the best selected model for predicting the air temperature is ARIMA $(1,0,1)(1,1,0)_{[12]}$. The coefficients and values of AIC (**A**kaike **I**nformation **C**riteria), AICc (**A**kaike **I**nformation Criteria corrected) and BIC (**B**ayesian **I**nformation **C**riteria) information criteria are shown in Table 2 [8].

Forecast data obtained from the chosen SARIMA model, along with original data and validation data is shown in Figure 3 [8]. The green line represents historical data from original data set and blue line represents validation data. Red line represents prediction values for 12 months ahead.

### 4.2. RANDOMFOREST METHOD RESULTS

For building a RandomForest model, function "randomForest" was used. This functions contains two parameters ntree and mtry. First parameter sets number of trees that are being used. For our research, ntree parameter was set to 100. The forecast data obtained using this model is shown in Figure 4. The blue line represents the historical data and red line represents the forecast data for 12 months ahead.

### 4.3. MODEL EVALUATION

Model evaulation is performed by using RMSE metrics obtained for both models which is shown in Table 3.

Based on values of RMSE, accuracy of ARIMA model is higher as its RMSE is lower than one of RandomForest model. A higher RMSE in RandomForest model indicates that there are certain deviations from the residuals to the actual values compared to ARIMA model. Therefore, the conclusion is that ARIMA model is the best model to fit the data.

**Table 2.** Coefficients and criteria values of best SARIMA model.

| Model | Coefficients | | | Criteria | | |
|---|---|---|---|---|---|---|
| | AR1 | MA1 | SAR1 | AIC | AICc | BIC |
| ARIMA (1,0,1)(1,1,0)[12] | -0.6036 | 0.9209 | -0.4079 | 228.59 | 229.52 | 236.08 |

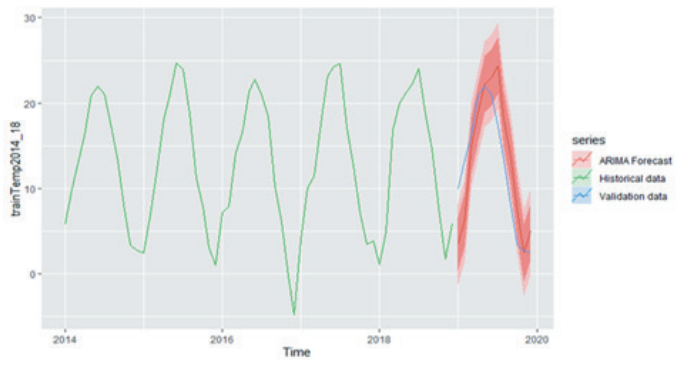


**Figure 3.** ARIMA $(1,0,1)(1,1,0)_{[12]}$ forecast.

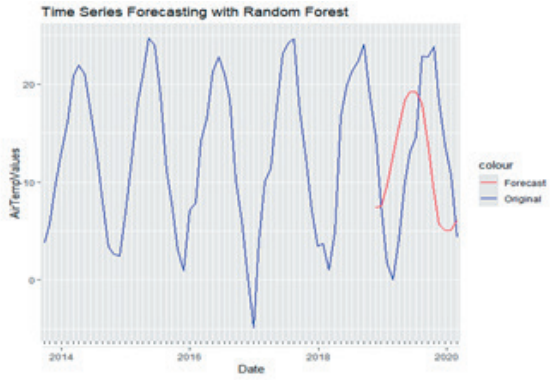


**Figure 4.** RandomForest model forecast.

**Table 3.** RMSE metrics for both models.

| Measure | ARIMA $(1,0,1)(1,1,0)_{[12]}$ model | RandomForest model |
|---|---|---|
| RMSE | 2.083858 | 8.693726 |

## 5. CONCLUSION

ARIMA and RandomForest methods are one of the most common methods for forecasting time series. They both have certain advantages and disadvantages. This research presented a methodology of selecting the best prediction model based on metrics such as RMSE. Both models were applied on data set of average temperature values from period of January 2014 to December 2018. Best ARIMA model was automatically selected by using function auto.arima. RandomForest model was obtained by setting the parameter for number of trees to be 100. The evaluation of both model is performed comparing RMSE metrics and the conclusion is that ARIMA model has lower value of RMSE and therefore better prediction accuracy. The future work would include other relevant metrics used for evaluation of models or tuning the model parameters in order to improve the accuracy.

## 6. REFERENCES

[1] W. H. Organization, 2024. [Online]. Available: https://www.who.int/health-topic/drought#tab=tab_1.

[2] N. Nandgude, T. P. Singh, S. Nandgude and M. Tiwari, "Drought prediction: A comprehensive review of different drought prediction models and adopted technologies," *Sustainability*, vol. 15, no. 15, p. 11684, July 2023.

[3] T. Mills, Applied Time Series Analysis - A Practical Guide to Modeling and Forecasting, Loughborough, The UK: Elsevier Inc., 2019.

[4] "What is R?," [Online]. Available: https://www.r-project.org/about.html.

[5] N. Tihi, S. Popov, J. Bondžić and M. Dujović, "Visualization of Big Data as Urban Drought Monitoring Support in Smart Cities," *Fresenius Environmental Bulletin*, vol. 30, no. 01, pp. 716-722, 2021.

[6] N. Tihi, S. Popov and M. Kavalić, "A comparison of Seasonal ARIMA and Holt-Winters Exponential Smoothing Models for Urban Drought Prediction," in *International Multidisciplinary Conference "Challenges of Contemporary Higher Education-CCHE 2024"*, Kopaonik, Srbija, 2024.

[7] N. Tihi and S. Popov, "Application of Holt-Winters Exponential Smoothing Method in Urban Drought Prediction," in *Proceedings of Engineering Conference*, Bečići, Montenegro, 2023.

[8] M. Todorov, N. Tihi, S. Popov and B. Stamatović, "Time series models for weather forecasting in smart cities," in *Proceedings of AlfaTech 2024*, Belgrade, Serbia, 2024.