A Real-Time Weather Forecasting and Analysis

Sushmitha Kothapalli,
Dept of Computer Science and Engineering,
Hubballi, India
sushmithachowdary261993@gmail.com

S. G. Totad
Dept of Computer Science and Engineering
Hubballi, India
totad@bvb.edu

***Abstract*-- Weather forecasting is the attempt by meteorologists to predict the weather conditions at some future time and the weather conditions that may be expected. The climatic condition parameters are based on the temperature, wind, humidity, rainfall and size of data set. Here, the parameters temperature and Humidity only are considered for experimental analysis. The data is collected from the temperature and humidity sensor called DHT11 sensor, which helps in detecting the temperature and humidity values of a particular region or location. The raspberry pi is used for storing the collected data to the cloud, with the help of Ethernet shield for uploading the data online. The data stored in cloud is generated in the form of CSV, JSON, XML files which is used for further analysis. The correlation analysis of the parameters helps in predicting the future values. The ARIMA model that gives better results for time-series data is used for predicting the values for forthcoming.**

Keywords: Analysis, Arduino Board, ARIMA model, Cloud, Correlation, Raspberry Pi, Time-series data.

I. INTRODUCTION

The data analysis plays an important role in discovering useful information, making predictions and decision making. The data analysis is used in many rapidly emerging fields like Healthcare, Weather Conditions, Media, Agriculture, Education, and E-commerce etc. for the business development and to reach the ever increasing customer satisfaction. Analyzing the data involves cleaning, transforming and building data model for the available dataset. So time-series data i.e. the continuous weather data of a particular region to predict the future weather conditions for the data analysis to predict the further weather conditions.

The DHT11 sensor reads temperature and humidity values of a particular region (the real-time data is been collected for this project), using Raspberry pi and Arduino board. The Arduino board connects to raspberry pi and sensors; the data read by sensors is stored in raspberry pi, which is connected with monitor and works as a CPU for the monitor. The raspberry pi stores the data in public cloud and the analysis is done on collected data using R (R studio) which is very effective for data analysis.

R comes with inbuilt functions, which helps in effective analysis. The time series comes with many models Auto Regressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), Multiple Linear Regression (MLR) models which provides good results for time-series data compared to other models [1]. Here the ARIMA model for the time series data yields better results than ANN and MLR [2]. The survey on these models is mentioned in next section. R comes with inbuilt ARIMA and Predict functions; with these functions the future values of weather are predicted. In time-series data Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) should be found before the prediction starts, The ACF and PACF values should be known to predict the values.

The organization of this paper is as follows. Section II provides information about the survey done on selected issue i.e, the methodologies that are used for processing and analyzing the real-time weather forecasting data. Section III gives the detail description about the techniques used for collecting and storing the real-time weather conditions. Section IV describes the analysis of collected data and also the predicts the future weather data values. Section V concludes which methodology is accurate compared to other. VI) References: The literature review is done by reading many papers and journals, in this section the detail about the papers and journals used for this study is mentioned.

III. RELATED WORK

The time-series data has several models which predicts the future weather data, Here only three models are considered Multiple Linear Regression, Artificial Neural Network and Auto-Regressive Integrated Moving Average model to know the best model for performing the data analysis, and gives better results than other techniques on time-series data.

a. Multiple Linear Regression

The Multiple Linear Regression (MLR) gives the information about the nature of relationship [3], through the MLR equation, however the mean value varies as the value of other variable in an equation changes. The MLR equation is as follows:

$$y = \alpha + \beta x + e$$

From the above equation, y is a dependent variable, x is an independent variable where value varies from 1 to n, e is an error value, α and β are the co-efficients; where α and β value

978-1-5386-0814-2/17/$31.00 ©2017 IEEE

depends on person who builds the model. Multiple Linear Regression is used for building prediction model which generates the potential predictors[4], and predicts rainfall for upcoming years[5]. An Artificial Neural Network and Decision Tree can be used to build efficient model.

b. Artificial Neural Network model

Artificial Neural Network model (ANN) is based on neural network, where the hidden layers of network affects the results of this model, whenever the layers of neural network increases the outcome of the model changes. The predictive model is built by analyzing the hidden layer nodes [6]; those three nodes found that the predictive model is best; the average rainfall over India is predicted. As the layers of the network increases the error factor will decrease [7]. The information about the monsoon data of Himalayas and predicting the weather of Himalayas can be done using the hybrid models of Artificial Neural Network [8]. This is the survey on MLR and ANN, the work on ARIMA is as follows.

c. Auto Regressive Integrated Moving Average

Auto Regressive Integrated Moving Average (ARIMA) is fit for the analysis of time-series data, which provides better results for model building, understanding the data and predicting the future values of data. The general form of ARIMA model is given below:

$$Y_t = a_0 + \sum a_i \, . \, y_{t-i} + \sum b_j \, . \, e_{t-j}$$

i = 1,2,3,……p and j = 1,2,3,…….q

where Y_t indicates the stationary stochastic process with non-zero mean, a_0 is the constant co-efficient, e_i represents noise distribution term, a_i autoregressive co-efficient, b_j is the moving average co-efficient. This model helps in building tools to predict the lead time in environment policy [9]. The hybrid fuzzy model can be built using ARIMA, which gathers time-series data [10], helps in handling and finding the uncertainties to provide better results.

III. PROPOSED TECHNIQUE

The proposed work has different techniques to collect the data and storing it. Appropriate sensors are used to collect the real-time data, based on the need of the parameters. The DHT11 collects the temperature and humidity values; Then if it is for pressure one more sensor is to be used. The architecture diagram of the proposed work is shown in the figure 1 .

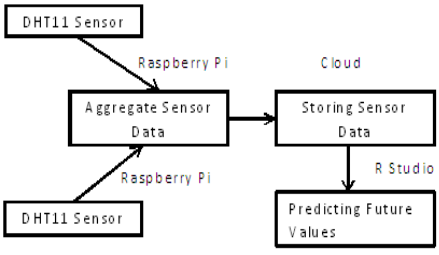

Figure 1: Architecture of the Project

The hardware connections are made based on parameter requirements. Here temperature and humidity are considered for experimental analysis, The DHT11 sensor is connected to Arduino board and the Arduino board is connected to Raspberry Pi through a serial USB. The architecture of the hardware connections is given as follows;

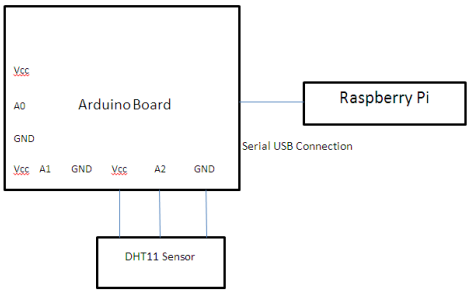

Figure 2: Circuit Diagram for Collecting Data

The Arduino board and DHT11 sensors are connected through VCC, A2 and GND pins. The VCC pin supplies power to sensor, A2 pin collects data from the sensor and GND helps in grounding to handle the circuit blasts. The Arduino UNO-R3 doesn't come with a Wi-Fi shield to connect with cloud online. For this purpose Raspberry Pi is used to store the data in cloud. The Arduino board is connected with Raspberry Pi using serial USB connection and the collected data is stored in public cloud through Thingspeak. The data collected for one year is stored in form of CSV file and the file is downloaded from cloud to perform analysis.

R studio is used to perform data analysis, which provides required results for the user. Therefore the correlation between the parameters are found to check how the parameters affect the presence of other parameter. Then the correlation functions are applied on the data to get the knowledge about the data. Using the observations done on the correlation functions performed on the data, the prediction model is used. Further a small survey is done on the model of time-series data. The ARIMA model is used for the further analysis of the data, to predict the values for upcoming year.

IV. ANALYSIS AND RESULTS

The parameter values collected from DHT11 sensor are stored on internet using cloud, which is used for further analysis using R studio. The data collected through Raspberry Pi can be viewed on webpage using a separate IP address. The figure shows the result from web server.

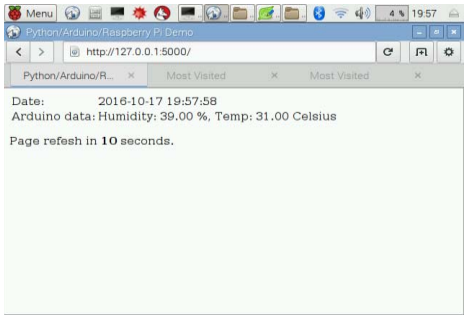

Figure 3: Data collection

The analysis is done by finding the correlation between the parameters, through correlation functions on the parameters and then continues to predict the future values of weather. The correlation analysis is carried out on three parameters, they are; Temperature, Humidity and Pressure. The correlation between the parameters is given as follows;

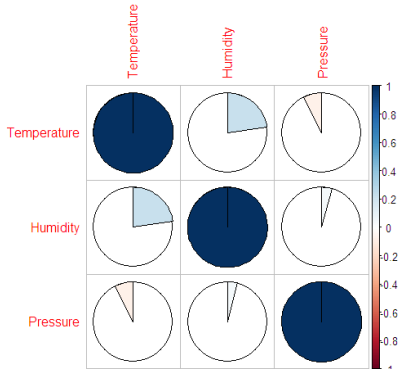

Figure 4: Correlation between parameters

From the above graph, the light blue to dark navy blue indicates that the correlation is high as the color gets darker. If the color is light, then it indicates that it is very weakly correlated. The correlation between temperature and humidity is around 0.4-0.6. Hence the correlation between those two parameters is good compared to temperature and pressure whose value is around -0.2, which shows that the existence of pressure parameter doesn't affect the value of temperature. The humidity and pressure are also correlated around 0.2.Hence further only temperature and humidity parameters are considered.

The average temperature can be found using the following histogram,

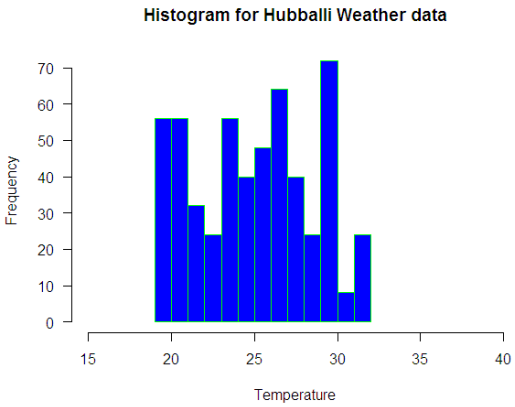

Figure 5: Histogram for Average Temperature

The above graph shows that the average temperature lies between 25-28 degree celsius. The bar with high spike indicates average temperature value. The graph shows that the temperature doesn't go below 17 and above 34 degree Celsius. The Auto Correlation function on the parameters is given as follows;

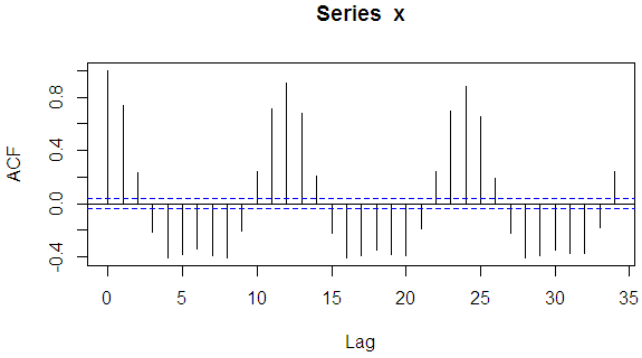

Figure 6: Auto Correlation Function

The above Auto Correlation Function (ACF) graph shows how the variables are related to each other. The correlation between the present value and the next value is considered i.e. it performs only between the successive variables. The Lag indicates the number of rows in an interval of time. Some variables have high correlation value 0.8, and some with low correlation -0.4. The Partial Auto Correlation Function (PACF) is given by;

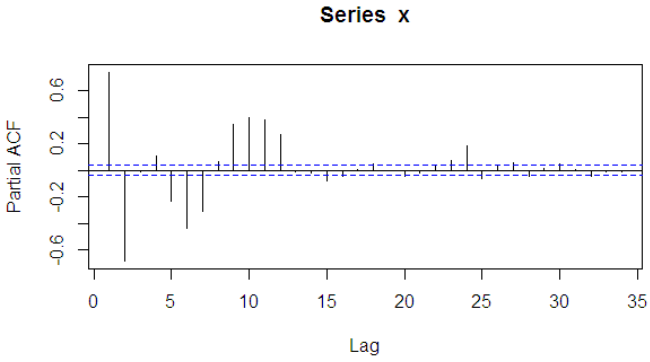

Figure 7: Partial Auto Correlation Function

The Partial Auto Correlation Function (PACF) considers continuous variables with respect to present variable. It finds the PACF until it finds the correct variable that can be used. The prediction of future values is carried out using the ARIMA model in R. before prediction starts ACF and PACF which we already performed. The predicted future year's values are represented in the form of graphs, because I am using R. The predicted values graph is given below;

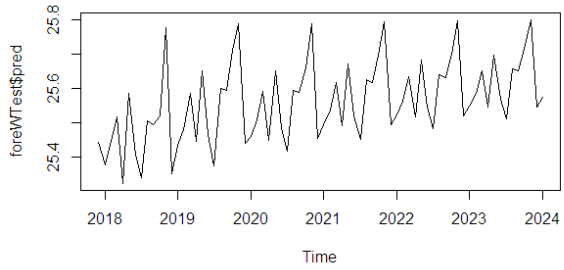

Figure 8: Predicted future values graph

The graph's x-axis represents the year's for which the weather data to be predicted, y-axis indicates the average temperature values where the predicted data lies. The dataset used for this work starts from the month of June. So, the first value of every year indicates the June month values. The high spike near to an end of the year clearly indicates April and May months, because the month starts from the June month, remaining months have low values due to the rainy and winter season month. The following graph indicates the standard error graph, which helps in gaining information like which year can be accurately predictable, the graph is as follows;

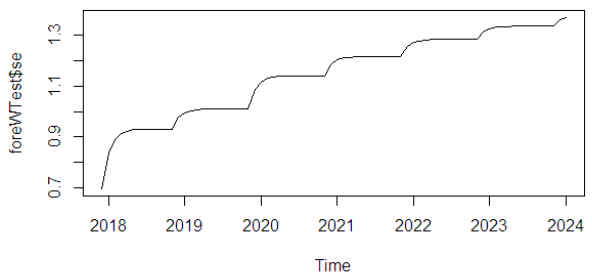

Figure 9: Standard Error Graph

The standard error indicates that the predicted value nearly plus or minus of error value. The above graph shows that the values for year's 2018 and 2019 nearly accurate, but for 2020 and above the values are not accurate. The standard error value for 2018 and 2019 is less i.e. less than 1, but for 2020 and above crosses the value 1 which cannot be an accurate for the prediction. From these analysis and results this project is concluded in next section.

V. CONCLUSION

The real-time data i.e. time-series data is gathered and analysis is performed on this dataset using R. Data can be collected using other devices rather than Arduino and Raspberry Pi. Here the values are predicted by implementing the ARIMA model using R studio, the prediction can be done for every month and every year. The proposed system takes dataset whose values starts from June, But it can start from any Month. Only first two successive years give accurate results in prediction of future weather values, because of the standard error values as the year's increases. During the prediction two or more models can be used for same dataset, to find the accuracy of each model and also find which model is appropriate for predicting the weather parameters. Addition to this the number of parameters can also be considered.

REFERENCES

[1] Edward N. Lorenz "Dynamical And Empirical Methods Of Weather Forecasting" Massachusetts Institute Of Technology.

[2] Mathur, S., and A. Paras. "Simple weather forecasting model using mathematical regression." Indian Res J Exten Educ: Special 1 (2012).

[3] Monika Sharma, Lini Mathew, Chatterji s. "Weather Forecasting using Soft Computing and Statistical Techniques" . IJAREEIE. Vol.3 , Issue 7,

[4] Sohn T., Lee J.H., Lee S.H. and Ryu, "Statistical prediction of heavy rain in South Korea" Advances in Atmospheric Sciences, Vol. 22, 2005.

[5] Kannan, M. Prabhakaran S. and Ramachandran, P. "Rainfall forecasting using data mining technique". International Journal of Engineering and Technology, Vol. 2, No. 6, pp. 397-401, 2010.

[6] Chattopadhay S. "Multiplayer feed forward artificial neural network model to predict the average summer monsoon rainfall in India". Acta Geophysica, Vol. 55, No. 3, pp. 369-382, 2007.

[7] Hayati M. and Mohebi Z. "Temperature forecasting based on neural network approach". World Applied Science Journal, Vol. 2, No. 6, pp. 613-620,2007.

[8] Kal N., Jim C. and Moula C. "The inventory policy using ESWSO measure for the ARIMA lead-time demand and discrete stochastic lead-time". Journal of Academy of Business and Economics, Vol. 10, No. 2, 2010.

[9] Badmus M.A. and Ariyo O.S. " Forecasting cultivated areas and production of maize in Nigeria using ARIMA model". Asian Journal of Agricultural Sciences, Vol. 3, No. 3, pp. 171-176, 2011.

[10] Saima H., Jaafar J., Belhaouari S. and Jillani T.A. "ARIMA based Interval Type-2 Fuzzy Model for Forecasting". International Journal of Computer Applications, Vol. 28, No. 3, pp. 17-21,2011.

[11] Sharma M.A. and Singh J.B. "Use of Probability Distribution in Rainfall Analysis". New York Science Journal, Vol. 3, No. 9, pp. 40-49,2010