2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)

Prediction of Temperature using Linear Regression

Sindhu P. Menon
Dept. of CSE,
Global Academy of Technology,
Bengaluru
sindhu33in@gmail.com

Ramith Bharadwaj
Dept. of CSE,
Global Academy of Technology,
Bengaluru

Pooja Shetty
Dept. of CSE,
Global Academy of Technology,
Bengaluru

Prajwal Sanu
Dept. of CSE,
Global Academy of Technology,
Bengaluru

Sai Nagendra
Dept. of CSE,
Global Academy of Technology,
Bengaluru

***Abstract*—An urban heat island (UHI) is an urban area or metropolitan area that is significantly warmer than its surrounding rural areas due to human activities. The project aims to showcase the effect of UHI using temperature as the independent variable with pollution and population as the dependent factor variables. Using the Time Series analysis we obtained the trend in temperature, population and pollution. Using Multiple Linear Regression we have predicted the temperature based on the factors. The accuracy of the predicted values is depicted by comparing the predicted and measured values of the years 2013-2016. Hence proving that the predicted value is accurate and measures must be taken to prevent this or the temperature will increase to an unlivable extent.**

Keywords—UHI, Time series; Linear Regression; Predictive analysis; Temperature prediction

I. INTRODUCTION

An urban heat island (UHI) is an urban area or metropolitan area that is significantly warmer than its surrounding rural areas due to human activities.[1][2][3] The temperature difference usually is larger at night than during the day, and is most apparent when winds are weak. UHI is most noticeable during the summer and winter. The main cause of the urban heat island effect is from the modification of land surfaces. Waste heat generated by energy usage is a secondary contributor. As a population center grows, it tends to expand its area and increase its average temperature. The less-used term heat island refers to any area, populated or not, which is consistently hotter than the surrounding area.

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series analysis can be applied to real-valued, continuous data, discrete numeric data, or discrete symbolic data. A number of different notations are in use for time-series analysis. A common notation specifying a time series X that is indexed by the natural numbers is written

$$X = \{X_1, X_2, ...\} \qquad (1)$$

Another common notation is

$$Y = \{Y_t: t \in T\} \qquad (2)$$

where T is the index set.

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function.

II. STUDY AREA: BENGALURU CITY

Greater Bangalore (77°37'19.54'' E and 12°59'09.76'' N) is the principal administrative, cultural, commercial, industrial, and knowledge capital of the state of Karnataka with an area of 741 sq. km. Bangalore city administrative jurisdiction was widened in 2006 by merging the existing area of Bangalore city spatial limits with 8 neighbouring Urban Local Bodies (ULBs) and 111 Villages of Bangalore Urban District (Ramachandra and Uttam Kumar, 2008; Sudhira et al., 2007). Thus, Bangalore has grown spatially more than ten times since 1949 (69 square kilometers) and is a part of both the Bangalore urban and rural districts (figure 1). Now, Bangalore is the fifth largest metropolis in India currently with a population of about 7 million (figure 2). The mean annual total rainfall is about 880 mm with about 60 rainy days a year over the last ten years. The summer temperature ranges from 18° C – 38° C, while the winter temperature ranges from 12° C – 25° C. [4]

Thus , Bangalore enjoys a salubrious climate all round the year. Bangalore is located at an altitude of 920 meters above mean sea level, delineating four watersheds, viz. Hebbal, Koramangala, Challaghatta and Vrishabhavathi watersheds.

978-1-5386-2361-9/17/$31.00 ©2017 IEEE

The undulating terrain in the region has facilitated creation of a large number of tanks providing for the traditional uses of irrigation, drinking, fishing and washing. This led to Bangalore having hundreds of such water bodies through the centuries. Even in early second half of 20th century, in 1961, the number of lakes and tanks in the city stood at 262 (and spatial extent of Bangalore was 112 sq km). This led to Bangalore having hundreds of such water bodies through the centuries. Even in early second half of 20th century. However, number of lakes and tanks in 1985 was 81 (and spatial extent of Bangalore was 161 sq km) shown in Figure 1.

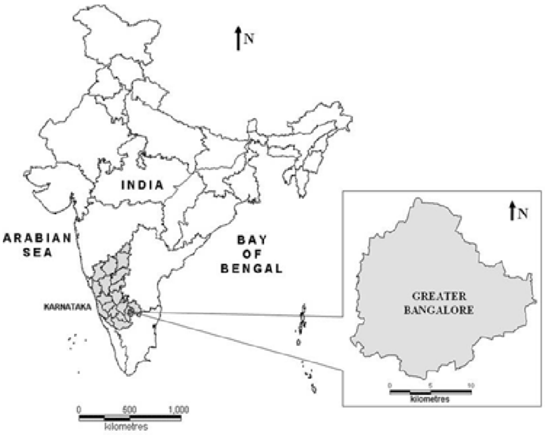

Figure 1: Study area Bengaluru

III. LITERATURE SURVEY

A. Studies on Time Series

- Mathew et al. (2016) have presented the concept of Linear Time Series model (LST) which is used to predict the temperature of Jaipur city. They have further pressed on the fact that urbanization is one of the major leading factors affecting the ecosystem, and thus a major contributing factor for UHI [2].
- Sharma & Sisodia, (2014) have illustrated the use of fuzzy time series in temperature prediction. They have specified that the length of intervals is important to the fact that it can affect the forecasting accuracy rate [3].
- Ali et al. (2012) have focused on the importance on the data filtering concept for the proper and accurate implementation of time series analysis. They have pointed out that noisy data will become a nuisance while using time series and pre-processing of data is required [4].

B. Studies on Urban Heat Island Effect (UHI)

- Zhongli & Hanqiu, (2016) have explained the effect of UHI in detail and its effect on urban cities in comparison with the surrounding areas. Using Fuzhou, China city as an example they have showcased the severity in which UHI is making the cities a furnace [5].
- Lucena et al. (2015) have explained the UHI effect in the city of Brazil in detail using the Land-Surface temperature as the parameter. They have proved that inn other places there is an increase of 2°C but in UHI there is an increase by 4.4°C [6].
- Li et al. (2014) have explained the UHI source and sink in detail by comparing the LST of the regions. They proved and have come to a conclusion that the 70% of the UHI source i.e. the city must be considered as future reference for future city planning [7].

C. Studies on Regression

- Kavitha et al. (2016) have explained in depth the wide use and importance of linear regression model. They have showcased the accuracy of regression model over the time series analysis, since the former takes into account the dependent factors [8].
- Shen et al. (2016) have presented in their paper that linear regression is very accurate in predicting the usefulness of reviews with Yelp as an example. They have thus proved the accuracy and usefulness of regression analysis in various fields including e-commerce [9].
- Arjun et al. (2014) have presented the upper hand of multivariate regression since it takes into account multiple dependent factors. They have used the multivariate model to predict the wind speeds and obtained accurate results thus proving the models wide application range and also its accuracy [10].

IV. TIME SERIES

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data [5][6][7]. Time series forecasting is the use of a model to predict future values based on previously observed values.

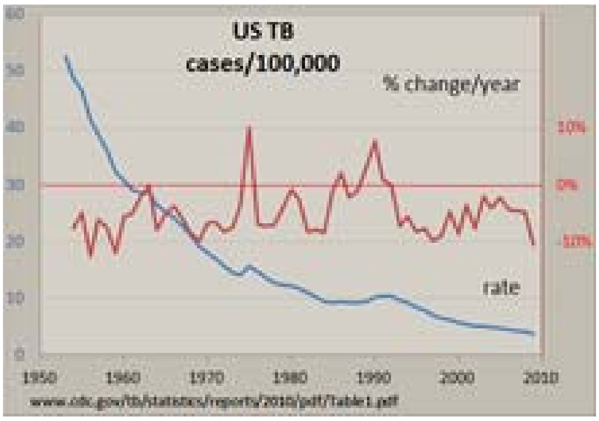

Figure 2: A Time Series example

Methods for time series analysis may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and wavelet analysis; the latter include auto-correlation and cross-correlation analysis. In the time domain, correlation and analysis can be made in a filter-like manner using scaled correlation, thereby mitigating the need to operate in the frequency domain. An example of time series graph is shown in Figure 2.

V. Regression

Regression analysis [8] is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function.

Linear regression is an approach for modeling the relationship between a scalar dependent variable *y* and one or more explanatory variables (or independent variables) denoted *X*. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression[9][10]. Figure 1.3 depicts a simple example of linear regression model graph which has one independent variable.

The general formula for regression is:

$$Y = Z\beta + € \qquad (3)$$

Suppose we have a sample of size n. As before, the design matrix Z has dimension n×(r + 1). But now:

$$Y_{n\times m} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \vdots\vdots & \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{(1)} & \mathbf{Y}_{(2)} & \cdots & \mathbf{Y}_{(p)} \end{bmatrix}, \qquad (3)$$

where $Y_{(i)}$ is the vector of n measurements of the i^{th} variable. Also,

$$\beta_{(r+1)\times m} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix} = \begin{bmatrix} \beta_{(1)} & \beta_{(2)} & \cdots & \beta_{(m)} \end{bmatrix}, \qquad (4)$$

where $\beta_{(i)}$ are the (r+1) regression coe□cients in the model for the i^{th} variable. Finally, the p n−dimensional vectors of errors €(i),i = 1,...,p are also arranged in an n×p matrix

$$\epsilon = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1p} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{np} \end{bmatrix} = \begin{bmatrix} \epsilon_{(1)} & \epsilon_{(2)} & \cdots & \epsilon_{(p)} \end{bmatrix} = \begin{bmatrix} \epsilon_1' \\ \epsilon_2' \\ \vdots \\ \epsilon_n' \end{bmatrix}, \qquad (5)$$

The above values are integrated into the formula 5 to obtain the Y value i.e. temperature.

Thus, by implementing multiple linear regression we obtain the predicted temperature value based on the factors i.e. pollution and population.

VI. Existing System

In the existing system the land surface temperature of Jaipur city [2] is predicted. They have implemented linear time series model to predict the temperature. Prediction was only for 2 years i.e. they have only a one-year model and a two-year model. The time series analysis implemented is an analysis that takes into account only the trend. The input parameters used are EVI and elevation. Thus, the existing system does not take into account dependent factors while predicting the temperature. It also doesn't give values beyond 2 years which is not very useful when taking environmental steps keeping in mind the future outcomes.

VII. Proposed System

In our system, we propose to predict the temperature of Bengaluru city. We have considered Bengaluru because it is known for is cool weather and is also called the summer capital of Karnataka. The city has also seen rapid urbanization due to the development of IT industry. Currently it is the 5th most populous state in the country and 18th most populous in the world and the vehicular population is increasing by a minimum of 10% every year. Hence, we decided to consider Bengaluru for our project with pollution and population as the dependent factors.

In our project we have implemented time series as well as regression analysis to predict the temperature. The project aims at predicting the temperature up to the year 2050. Currently, we have collected the data from 2005 to 2016 of temperature, pollution and population. The time series analysis is used to obtain the trend in temperature, pollution and population. Then, we implement the regression analysis to predict the temperature with pollution and population as the dependent factors.

The system also aims at improving the accuracy of the prediction if necessary. The accuracy of the predicted values is proved by comparing the measured values and the predicted values of temperature of the years 2013 to 2016. Thus, we propose from the results that certain steps have to be taken towards saving the city from very hot future.

VIII. ARCHITECTURE

Figure3 represents the various phases of our project

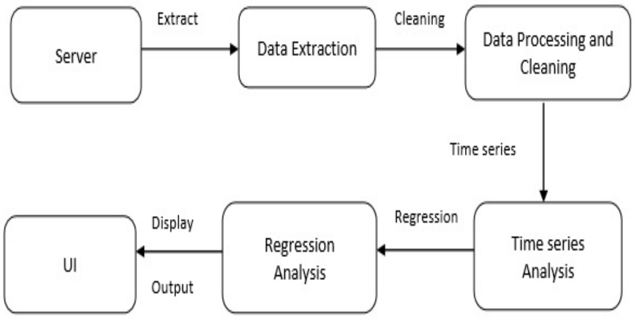

Figure 3: Phases of the proposed system

IX. IMPLEMENTATION

Server: The server contains all the event handlers that will be executed based on the input given by the user in the UI. The server takes care of the entire back end activities such as which graph to be plotted and shown in the UI for the chosen tab, which tab should be shown when the user selects a particular input, and so on. Thus, the server takes care of the overall handling of the system.

Data Extraction: The first step in our project is to collect the data of temperature, pollution and population. We have collected the data of pollution from the Central Pollution Control Board website and stored it in csv files. The data of population is collected from the Bengaluru census website and stored in the form of csv file. The temperature data is collected from https://en.tutiempo.net/ which is a website where the temperature data of various cities around the world is collected from official sites and made available. All the csv files are placed in a common folder and extracted individually into separate variables in our project.

Data Processing and Cleaning: The next step in the project is processing of the data extracted. The data of pollution and population is obtained year wise, hence it is directly sent for data cleaning. The temperature data is obtained day wise and the data is stored in separate files for each month from 2005 to 2016. We read the data from all the files, then combine and store it in a common file for easy processing and access. The extracted data is then processed for any anomalies, if found it is immediately solved. The data extracted may also contain erroneous data or no data i.e. -, 0, and so on which will result in inaccurate results during the implementation of prediction models. All such data is cleaned and is made as <NA> i.e. non-applicable so that it doesn"t interfere in the implementation of the analysis models.

Time Series Analysis: The time series model is applied on the cleaned data. Using time series analysis we obtain the trend in the temperature, pollution and population. Thus, this step sets the base line for the prediction of the temperature. The corresponding data is stored in different variables and sent to the UI where the data is represented in a user-friendly graphical format.

Regression Analysis: Multiple linear regression analysis is implemented on the data obtained from time series, where temperature is the independent variable, pollution and population are the dependent factors. The module upon implementation stores the result in a variable with data such as formula used, R mean square value, intercept value, and so on. Upon calling the predict function we obtain the predicted values of temperature. The result set is sent to the UI where the results are represented graphically.

UI: In this module the UI components are defined and the corresponding event handlers are defined. The UI consists of a sidebar panel where the user can provide input, and there is a main panel where the data is represented in the form of graphs in various tabs. The data from 2005 to 2016 can be viewed on a monthly, semi-annual or year basis. The results of the time series and regression models are also represented graphically. Finally there is a comparison tab where the graph shows the accuracy of the prediction by comparing the measured versus the predicted values of the years from 2013 to 2016.

X. DATA AND RESULTS

The temperature data of Bengaluru city of the years 2005 – 2016 is obtained. The pollution data is also obtained from the Pollution Control Board website. The pollution data is collected from the Bengaluru census website and the corresponding data is cleaned and processed.

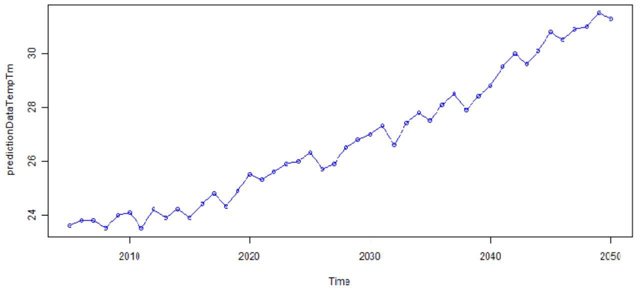

Figure 4: Trend in temperature

By implementing the time series analysis we obtain the trend in temperature, pollution and population and we plot the graph to show the trend. The Figures 4, 5,6 show the graph of temperature, pollution and population trend up to the year 2050.

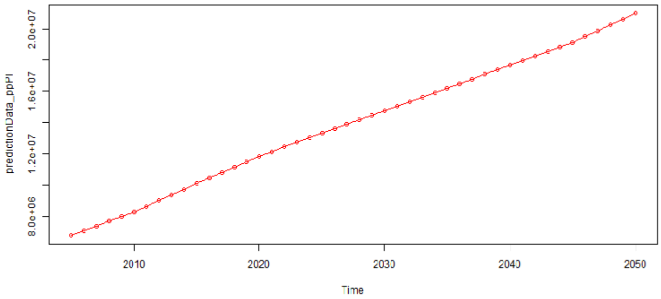

Figure 5: Trend in Pollution

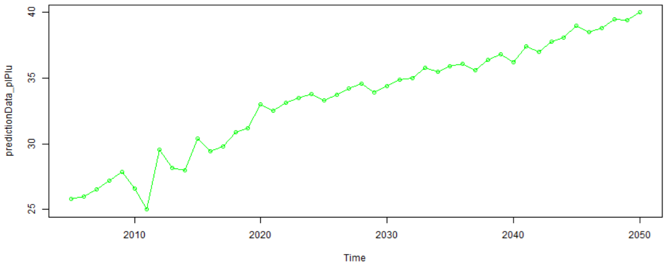

Figure 6: Trend in Population

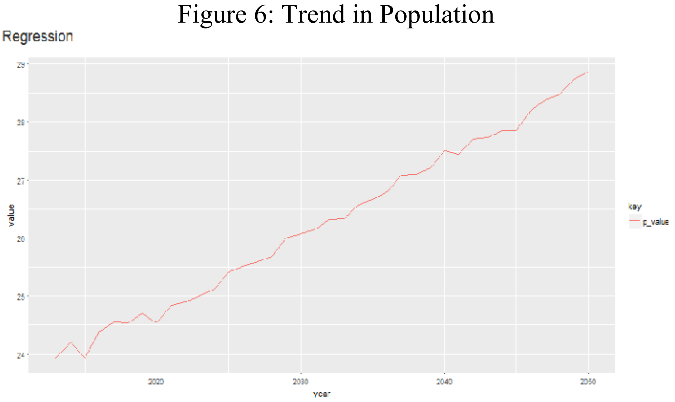

Figure 7: Regression result graph

The above data set is passed on to the regression algorithm where temperature is the independent variable that is being predicted and pollution and population are the dependent factors. The corresponding result obtained is plotted on a graph to have clear understanding and a graphical perspective which can be seen in the Figure 7.

The accuracy of the above prediction is proved by comparing the measured value and the predicted value of the years 2013-2016. The comparison can be seen the Figure 8 and in Table 1.

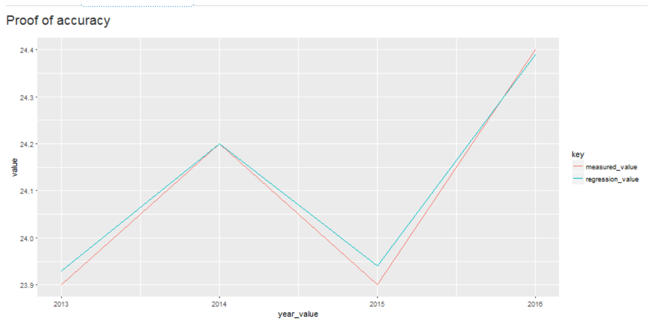

Figure 8: Comparison graph

TABLE I. Comparison of measured and predicted temperature

Year	Measured value	Predicted value
2013	23.9	23.92
2014	24.2	24.20
2015	23.9	23.94
2016	24.4	24.38

Every year the temperature of the cities is increasing at a higher pace than the surrounding area, this isn't causing any problem at present. But, in the future this will definitely affect us hence certain measures must be taken now so as to have a better future.

The major problem of the city is the rapid inflow of population due to development and job opportunities. This huge inflow of humans into a small city will result large apartments, more vehicular traffic, more pollution and many other factors. Thus, the major cause is this massive population growth of a city. This will also result in development of various areas to provide the necessary resources such as housing, jobs, recreational activities, road network and so on. This development not only leads to higher pollution, it also leads to the destruction of nature which actually keeps the environment balanced.

Another major cause is the air pollution. Especially in cities the vehicular traffic and industries are huge in number and as a result cause a lot of air pollution. The air index within the city center is so bad in majority of the cities that actually people are breathing out cleaner air than the air that they are breathing in. Hence, in our project to predict the temperature we have used population and pollution as the dependent factors.

XI. Conclusion

Our work shows that regression is more precise and accurate compared to time series analysis. Since urbanization is a major cause for increase in population and pollution of a city, we have predicted the temperature based on these factors. By looking at the comparison graph between the measured values and the predicted values of the years 2013-2016, we can come to the conclusion that the implemented analysis is accurate based on the currently available statistics.

Thus, from the results obtained we have come to the conclusion that unless certain environmental friendly steps are not taken, the urban cities will create an oven like effect, thus leading to very hot city where in the future it will be impossible to live without external aids.

Certain innovative ideas must be implemented so that the environment is balanced and the city also progresses towards the future. Some steps such as in house gardens, roof top gardens, balcony gardens etc. will help to increase the greenery as well improve the air condition in this concrete jungle of a city. Rain water harvesting, using renewable resources such as solar water heaters, solar and wind powered homes, electric cars must be used more in the cities to reduce the impact on the environment which will help us have a better and less hot future.

Future work on this project can be using more dependent factors to predict the temperature so that we obtain a more accurate result. Integrate the project with an IoT system which uses sensors to take live inputs of temperature and pollution, upload to the database from which this system can retrieve the data actively and predict the results i.e. temperature in more active way.

REFERENCES

[1] Zhongli L, Hanqiu, "A study of Urban heat island intensity based on "local climate zones": A case study in Fuzhou, China", In Earth Observation and Remote Sensing Applications (EORSA), 2016 4th International Workshop , pp: 250-254, July 2016.

[2] Lucena de, A.J, Faria Peres L, Rotunno Filho O.C, Almeida França, J.R., " Estimation of the urban heat island in the Metropolitan Area of Rio de Janeiro-Brazil", In Urban Remote Sensing Event (JURSE), 2015 Joint pp: 1-4, March 2015

[3] Li L.G, Wang H.B, Zhao Z.Q, Cai F Zhao, X.L, Xu S.L, "Characteristics of urban heat island (UHI) source and sink areas in urban region of Shenyang", In Earth Observation and Remote Sensing Applications (EORSA), 2014 3rd International Workshop , pp: 62-66, June 2014.

[4] Ramachandra T.V, Kumar U, " Greater Bangalore: Emerging urban heat island", GIS development, Vol 14, Issue 1, pp:86-104,2010

[5] Mathew A, Sreekumar S., Khandelwal S., Kaul N, and Kumar R, Prediction of Land-Surface Temperatures of Jaipur City Using Linear Time Series Model", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol 9, Issue 8, pp:3546-3552.

[6] Sharma Y, Sisodia, S, " Temperature Prediction Based on Fuzzy Time Series and MTPSO with Automatic Clustering Algorithm", In Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium , pp: 101-105, December 2014

[7] Ali A, Ghazali R, Ismail L.H, " The wavelet filtering in temperature time series prediction", In Uncertainty Reasoning and Knowledge Engineering (URKE), 2012 2nd International Conference, pp: 153-157, August 2012.

[8] Kavitha S, Varuna S, Ramya R., " A comparative analysis on linear regression and support vector regression", In Green Engineering and Technologies (IC-GET), 2016 Online International Conference pp: 1-5, 2016.

[9] Shen R, Shen J, Li Y, Wang H., " August. Predicting usefulness of Yelp reviews with localized linear regression models", In Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference, pp: 189-192, 2016

[10] Arjun N.N, Prema V, Kumar D.K, Prashanth P, Preekshit V, Rao K.U, "Multivariate regression models for prediction of wind speed", In Data Science & Engineering (ICDSE), 2014 International Conference, pp: 171-176, mk, August 2014.