Estimation of forecasting for Annual water usage Linear Regression Model of Machine Learning

Aman Dwivedi, Tannu Jindal, Aarya Verma, Karan Kushwaha,Nidhi Lal

Dept. of Computer Science and Engineering

IIIT Nagpur, India

amanrevenge2811@gmail.com

tannujindal1106@gmail.com

aaryaverma112@gmail.com

karankushwaha671@gmail.com

nidhi.2592@gmail.com

***Abstract*— Due to the limited natural water resources and the increase in population, managing water consumption is becoming an increasingly important subject worldwide. In this paper, we present and compare time series model and linear regression model that are able to predict water demand for annual water usage in baltimore. The study presents a reliable approach for long-term forecasting of water demand. The purpose of this study is to provide a convenient and reliable method for long-term forecasting of urban water demand while reducing the prediction uncertainty with limited number of features.**

The model is developed predicts the annual water usage based on water consumption of previous years.In order to evaluate the accuracy of the prediction, of linear regression outputs were compared with results time series model. Findings indicate that Linear regression model is an appropriate solution for long-term water demand forecasting. Furthermore, it can reduce uncertainties and significantly increase the accuracy of the long-term forecasting. These models achieve a high accuracy with a limited set of input features

***Keywords*— Linear regression model, Time series model.**

I. Introduction

Water use has been increasing worldwide by about 1% per year since the 1980s[1]. Global water demand is expected to continue increasing at a similar rate until 2050, accounting for an increase of 20 to 30% above the current level of water use, mainly due to rising demand in the industrial and domestic sectors. Over 2 billion people live in countries experiencing high water stress, and about 4 billion people experience severe water scarcity during at least one month of the year[2].

Drought, population growth, and increases in per capita consumption will give rise to a nearly global water crisis. On the other hand, water is a non-renewable commodity and makes the problem more complicated. In this situation,efforts should be made to optimize water consumption and prevent possible conflicts and quarrels to dominate water resources in the future. Furthermore, uneven distribution, either in time or resources, high evaporation rate,contamination of water resources, shortage of accurate information, price discordance, and social problems are yet other reasons that necessitate an accurate plan in order to optimize the efficiency of water resources' usage.

In this way, an estimation of future water demand can help decision-makers to take necessary measures according to the possible crisis and limitations. Forecasting domestic water demand and understanding its influencing factors are among the important steps in water crisis control and management.

II. Related work

Water demand forecasting is an active area of research as water consumption keeps increasing, especially in urban areas. In (House-Peters and Chang 2011), the benefits of both short-term and long-term predictions are reviewed. Short-term prediction is defined as the prediction of daily or monthly water consumption that is necessary to support operational decisions. Long-term prediction spans several years and depends on the forecasted growth of geographical regions. The authors of (House-Peters and Chang 2011) suggest that economic factors such as water price and household income are key features in water demand prediction. Similarly,(Arumugam et al. 2017) analyzed the spatio-temporal patterns of water consumption across the US and concluded that the efficiency of water-usage is higher in urban areas.

Regression models have traditionally been used in the prediction of water consumption levels. Some of the early research in the field (Morgan and Smolen 1976), (Hansen

and Narayanan 1981) used temperature and rainfall as weather features for these regression models. In (Morgan and Smolen 1976), regression models based on climatic indicators such as temperature and precipitation and potential evapotranspiration (sum of evaporation and plant transpiration)minus precipitation were developed. In (Hansen and Narayanan 1981), a multivariate regression model that includes the daily water demand of Salt Lake City, average temperature, total precipitation, and percentage of daylight hours as input features was proposed.
The importance of weather features for water demand forecasting was discussed in (Bakker et al. 2014). The authors tested three models with and without the inclusion of weather features. The results show that the models which take into consideration weather features outperformed the models without weather features.

III. Proposed Work

Linear Regression: Linear regression analysis is employed to predict the worth of a variable based on the value of another variable. The variable you wish to predict is termed as the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
Linear regression is employed in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.
In this study, we had to discover how to forecast the annual water usage in with Linear regression.
we had to use scipy environment,including:

- SciPy
- NumPy
- Matplotlib
- Pandas
- scikit-learn
- statsmodels

The problem is to predict annual water usage.The dataset provides the annual water usage in urban center from 1885 to 1963, or seventy nine years of knowledge.The values are within the units of liters per capita per day, and there are seventy nine observations.We should develop a check harness to research the information and evaluate candidate models.The dataset isn't current therefore we've got to figure on updated dataset.

This involves two steps:

1. Defining a Validation Dataset.
2. Developing a way for Model Evaluation.

2.1.Validation Dataset: The dataset is not current.Therefore, we will pretend that it is 1953 and withhold the last 10 years of data from analysis and model selection.

2.2.Model evaluation involves two elements:

- Performance Measure : We will evaluate the performance of predictions using the root mean squared error (RMSE).We can calculate the RMSE using the helper function from the scikit-learn library mean_squared_error() that calculates the mean squared error between a list of expected values (the test set) and the list of predictions. We can then take the square root of this value to give us a RMSE score.
- Test Strategy: We can split the dataset into train and test sets directly. We're careful to always convert a loaded dataset to float32 in case the loaded data still has some String or Integer data types.Using Numpy and Python code ,we can write code for test harness.

Persistence: The baseline prediction for time series forecasting is called the naive forecast, or persistence , here the previous time step is used as the prediction for the observation at the next time step.

Data Analysis: In this section, we will look at the data from four perspectives:

1.Summary Statistics.

2.Line Plot.

3.Density Plots.

4.Box and Whisker Plot.

Model Validation: It is validating and finalizing the model. It includes three steps.

1.Finalize Model: Train and save the final model.

2.Make Prediction: Load the finalized model and make a prediction.

3.Validate Model: Load and validate the final model.

MSE

In statistics, the mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated

values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

RMSE

The RMSE for training and test sets should be very similar if we wanted to build a really good model. If the RMSE for the test set is much higher than that of the training set, it is very likely that it will be very badly fit over the data i.e. it is a model that tests well in sample, but has very little predictive value when tested out of the sample.

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of the variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

IV. Result and Discussion

The best measure of model fit depends on the researcher's objectives, and more than one are often useful. The statistics discussed above are applicable to regression models that use OLS estimation. Many types of regression models, however, such as mixed models, generalized linear models, and event history models, use maximum likelihood estimation. These statistics are not available for such models.

Figure 1 provides a general view of the data set and parameter values considered in this study. We may observe that small water usage values were increasing in general every year. In this study , our predicted values gave RMSE of 6.197 which is very less compared to previous studies that were using time series model for analysis of similar data.In our study our MSE and R2_score were 38.407 and 0.6221 respectively.

Obviously, for the cases where parameter took lower values, a deeper analysis would be needed to determine the reasons (although the errors presented small values in the overall outcomes).

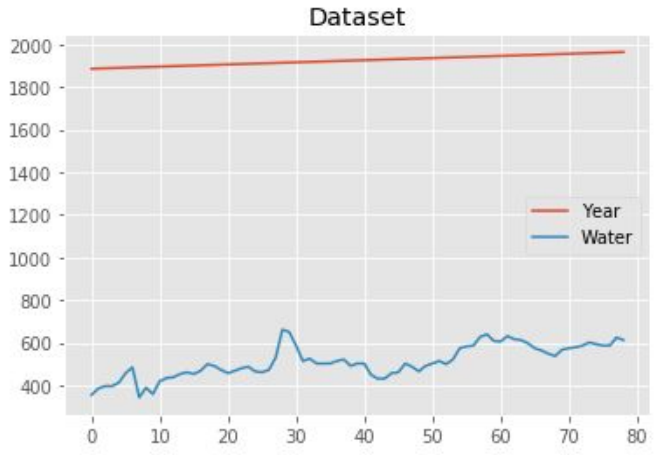

Figure 1

Discussion:

Based on the accurate results obtained in the general cases by the linear regression method applied here, we concluded that the method was suitable for annual water usage predictions. However, there were also some predictions where the values obtained were not so close to the actual measurements or that behaved worse than we expected.This can be clearly seen in figure 2 ,where during few years water usage decreased and increased significantly.To understand these sudden changes we needed more feature in dataset.

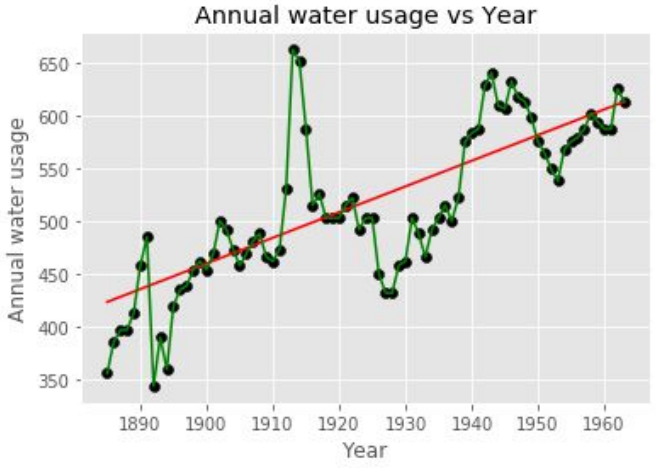

Figure 2

Figure 2 shows the previous work of who used time series model to forecast the annual water usage.This model has a Rmse value of 21.975.

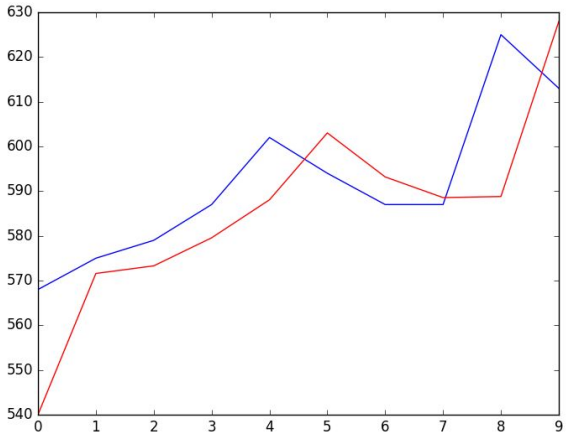

Figure 3:Plot of Forecast for Validation Dataset(previous work)

V. CONCLUSIONS

Knowing the amount of water demand is one of the important and effective factors in water resources management. To date, many reviews have been carried out in the field of water demand forecast but most of these studies have focused on the short-term forecast and time series model.In this study, we predicted the annual water usage with high level of accuracy compared to previous models. Linear Regression was used to predict the long-term water usage. For this purpose, to predict the water usage, effective variables for the period 1885 – 1963 were collected from the annual water usage data of Baltimore. Although the results obtained in the study were quite convincing and promising, there were some drawbacks that we would like to handle in the future. Water usage cannot be predicted based on previous year consumption,in future we would like to study a bigger dataset with other input features like population,salary,gender,age etc.

VI. REFERENCES

1)Water issues in developing countries - Wikipedia https://en.wikipedia.org/wiki/Water_issues_in_developing_countries

2) UN Water Publications https://www.unwater.org/publications/world-water-development-report-2019/

3) A long-term prediction of domestic water demand using preprocessing in artificial neural network

4) Sadegh Behboudian, Massoud Tabesh, Maliheh Falahnez had and Farrokh Alavian Ghavanini

5) Azadeh, A., Ghaderi, S. F. & Sohrabkhani, S. Forecasting electrical consumption by integration of neural network, time series and ANOVA. Appl. Math. Comput. 186, 1753–1761.

6) Babel, M. S., Das Gupta, A. & Pradhan, P. A multivariate econometric approach for domestic water demand modeling: an application to Kathmandu, Nepal. Water Resour. Manage. 21, 573–589.

7) Chen, H. & Yang, Z. F. Residential water demand model under block rate pricing: a case study of Beijing, China. Commun. Nonlinear Sci. 14, 2462–2468.

8) 41 S. Behboudian et al. | Long-term prediction of domestic water demand Journal of Water Supply: Research and Technology—AQUA | 63.1 | 2014

9) Firat, M., ErkanTuran, M. & Yurdusev, M. A. Comparative analysis of neural network techniques for predicting water consumption time series. J. Hydrol. 384, 46–51.

10) Greene, W. H. Econometric Analysis, 6th edn. Prentice-Hall, Upper Saddle River, NJ.

11) Kostas, B. & Chrysostomos, S. Estimating urban residential water demand determinants and forecasting water demand for Athens metropolitan area, 2000–2010. South-Eastern Eur. J. Econ. 1, 47–59.

12) Msiza, I. S., Nelwamondo, F. V. & Marwala, T. Water demand forecasting using multi-layer perceptron and radial basis functions. In: Proceedings of the IEEE International Conference on Neural Networks, Orlando, FL, 12–17 August, pp. 13–18.

13) Mylopoulos, N., Vagiona, D. & Fafoutis, C. Urban water pricing in sustainable water resources management: a socioeconomic study. World Acad. Sci. Eng. Technol. 54 (1), 547–553.

14) Renwick, M. E. & Archibald, S. O. Demand side management policies for residential water use: who bears the conservation burden. Land Econ. 74 (3), 343–360.

15) STATE OF MICHIGAN DEPARTMENT OF NATURAL RESOURCE www.michigan.gov/dnr/ RESEARCH REPORT 1995 Use of Multiple Linear Regression to Estimate Flow Regimes for All Rivers Across Illinois, Michigan, and Wisconsin. RR2095 April 2011 Paul W. Seelbach, Leon C. Hinz, Michael J. Wiley, and Arthur R. Cooper

Rumelhart, D. E., Hinton, G. E. & Williams, R. J.

16) Learning internal representations by error propagation. In: Parallel Data Processing (D. Rumelhart & J. McClelland, eds). Vol. 1, Chapter 8, MIT Press, Cambridge, MA, pp. 318–362.

17) Water Demand Forecasting Based onStepwise Multiple Nonlinear RegressionAnalysisAbdulkadir Yasar, Mehmet Bilgili &Erdogan Simsek Arabian Journal for Science andEngineering ISSN 1319-8025Volume 37Number 8 Arab J Sci Eng (2012) 37:2333-2341DOI 10.1007/s13369-012-0309-z