# MULTIVARIATE TIME SERIES ANALYSIS

A six week internship report

By

**Nagendra Pal singh**

National Institute of Technology

Hamirpur

Under the Supervision of

**Dr. Praveen Tripathy**

(Associate Professor)

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI

# Acknowledgement

This project has seen contributions from various individuals. It has been an honor to work under my mentor, Dr Praveen Tripathy, Associate Professor, IIT Guwahati. I am extremely thankful to him for his support and mentorship throughout the project.

# Contents

## 1. Abstract

This report discusses data-driven predictive models for the energy use of appliances. Data used include measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures. Statistical models were trained with repeated cross validation and evaluated in a testing set

a. LSTM
b. Facebook prophet
c. Regression Algorithms
d. Support vector machines
e. Classifies gradient boosting
f. XGBoost

From the wireless network, the data from the kitchen, laundry and living room were ranked the highest in importance for the energy prediction. The prediction models with only the weather data, selected the atmospheric pressure (which is correlated to wind speed) as the most relevant weather data variable in the prediction.

## 2. Introduction

The use of time series data for understanding the past and predicting future is a fundamental part of business decisions in every sector of the economy and public service. Retail businesses need to understand how much inventory stocking do they need to have next month; power companies need to know whether they should increase capacity to keep up with demand in the next 10 years; call centers need to know whether they should be hiring new staff anticipating higher call volumes — all those decision-making requires forecasting in the short and long-term, and time series data analysis is an essential part of that forecasting process.

Time series analysis accounts for the fact that data points taken over time may have an internal structure ( such as autocorrelation, trend or seasonal variation) that should be accounted for. The understanding of the appliances energy use in buildings has been the subject of numerous research studies. Since appliances represent a significant portion (between 20% and 30%) of the electrical energy demand. Thus prediction models of electrical energy consumption in buildings can be useful for a number of application: to detect abnormal energy use patterns, to be a part of an energy management system for load control, to predictive control applications where the loads are needed, for demand side management (DSM) and demand side response (DSR) and as an input for building performance simulation analysis.

The electricity consumption in domestic buildings is explained by two main factors: the type and number of electrical appliances and the use of the appliances by the occupants. Naturally, both factors are interrelated. The domestic appliances used by the occupants would leave traceable signals in the indoor environment near the vicinity of the appliance, for example: the temperature, humidity, vibrations, light and noise. The occupancy level of the building in different locations could also help to determine the use of the appliances. In this work, the prediction was carried out using different data sources and environmental parameters (indoor and outdoor conditions). Specifically, data from a nearby airport weather station, temperature and humidity in different rooms in the house from a wireless sensor network and one sub-metered electrical energy consumption (lights) have been used to predict the energy use by appliances.

## 2.a. Dataset Description

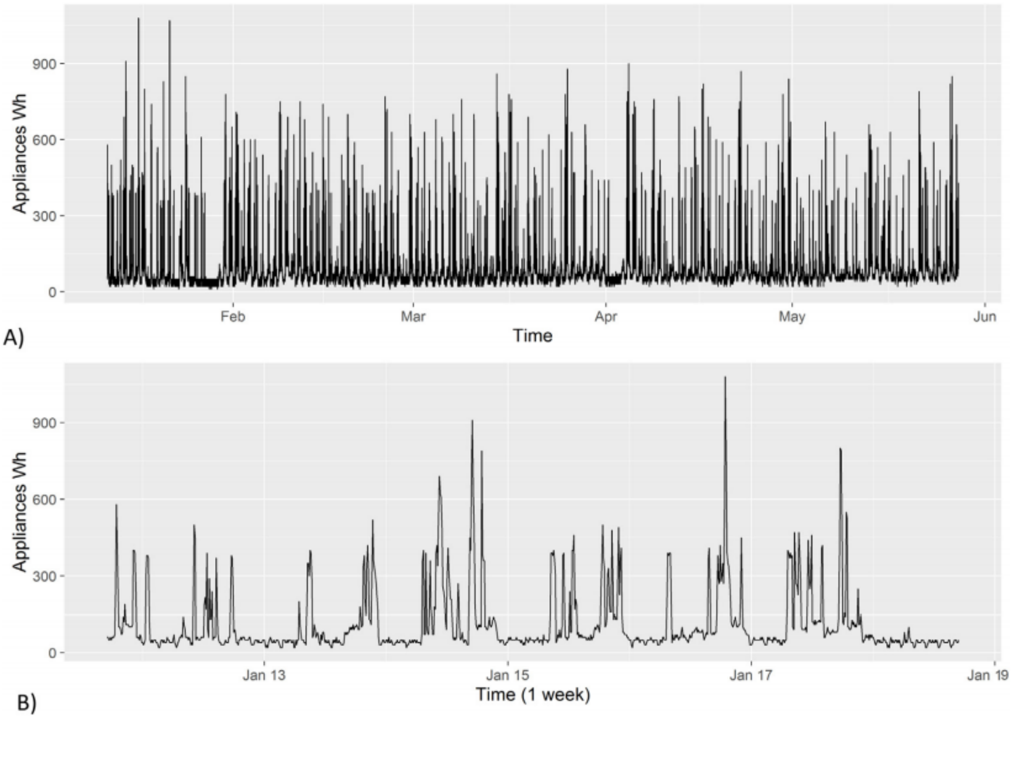

Fig 1. (A) Appliances energy consumption measurement for the whole period, (B) A closer look at the first week of data

The dataset is a time series from the month of jan to june. The combined data set is split in training and test validation, 75% of the data is used for the training of the models and the rest is used for testing.

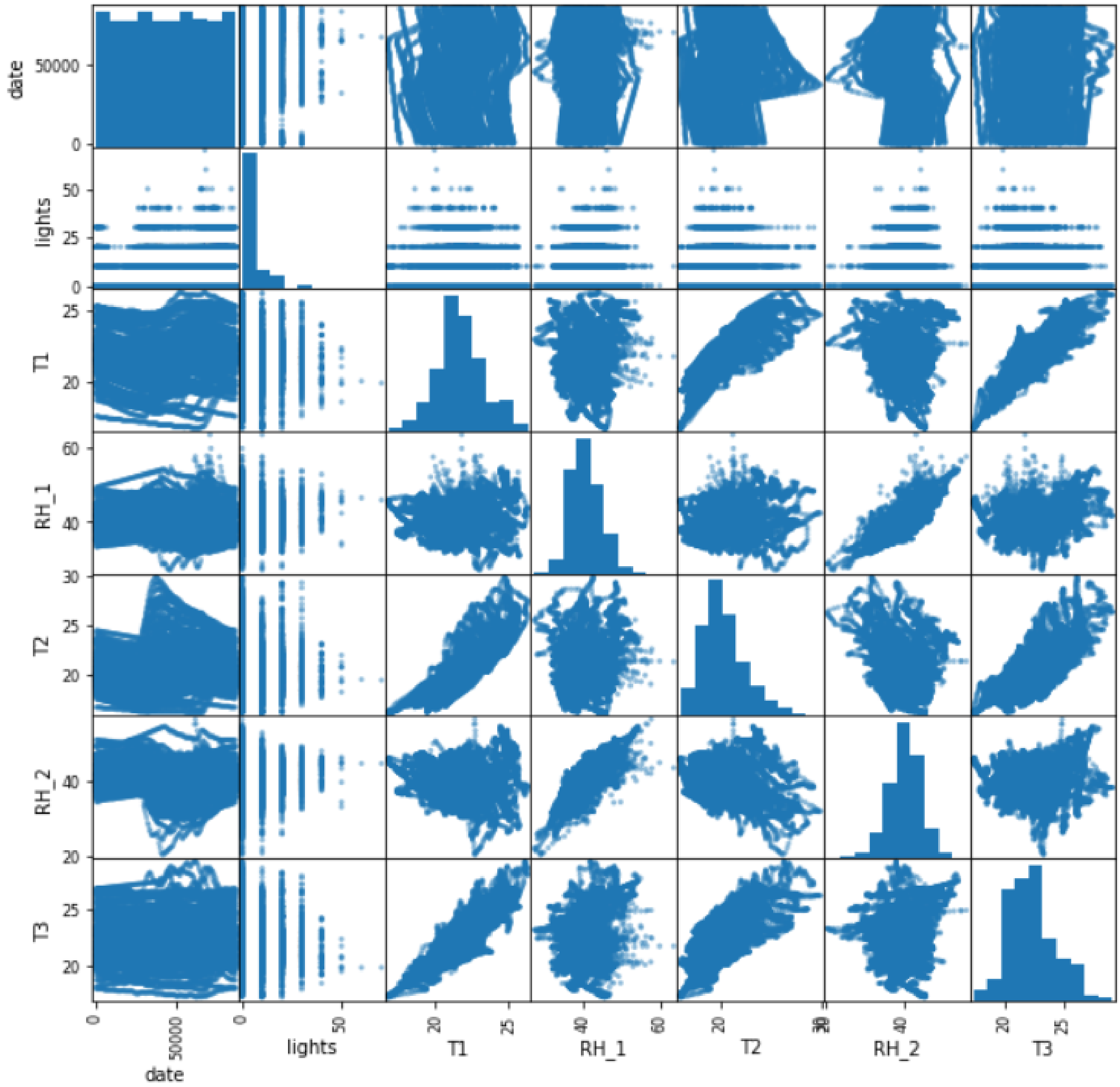

**Fig2. Relationship between the energy consumption of appliances with: lights, T1, RH1, T2, RH2, T3, RH3. T1 and RH1 correspond to the kitchen conditions; T2 and RH2 correspond to the living room conditions**

Some insights from the above plot:

1. T1,RH_1, T2,RH_2,T3 are normally distributed. No extra preprocessing needed.
2. Lights seem to be skewed. Need preprocessing
3. Date seems to be same for all the examples

4. Each pair of RH and T (RH_1&RH_2 or T1&T2) are positively correlated. This makes sense as well as each T is the temperature in different regions of the house.So, they will vary together. Similar for the RH as each of them is the Humidity in different regions of the house.

## 3. Multivariate Analysis and forecasting

A **Multivariate** time **series** has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. This dependency is used for forecasting future values.

### Time Series Analysis using LSTM

Time series prediction problems are a difficult type of predictive modeling problem. Unlike regression predictive modeling, time series also adds the complexity of a sequence dependence among the input variables.

A powerful type of neural network designed to handle sequence dependence is called recurrent neural networks. The Long Short-Term Memory network or LSTM network is a type of recurrent neural network used in deep learning because very large architectures can be successfully trained.

The Long Short-Term Memory network, or LSTM network, is a recurrent neural network that is trained using Backpropagation Through Time and overcomes the vanishing gradient problem.

As such, it can be used to create large recurrent networks that in turn can be used to address difficult sequence problems in machine learning and achieve state-of-the-art results.Instead of neurons, LSTM networks have memory blocks that are connected through layers.

A block has components that make it smarter than a classical neuron and a memory for recent sequences. A block contains gates that manage the block's state and output. A block operates upon an input sequence and each gate within a block uses the sigmoid activation units to control whether they are triggered or not, making the change of state and addition of information flowing through the block conditional.

There are three types of gates within a unit:

- **Forget Gate**: conditionally decides what information to throw away from the block.
- **Input Gate**: conditionally decides which values from the input to update the memory state.
- **Output Gate**: conditionally decides what to output based on input and the memory of the block.

Each unit is like a mini-state machine where the gates of the units have weights that are learned during the training procedure.

You can see how you may achieve sophisticated learning and memory from a layer of LSTMs, and it is not hard to imagine how higher-order abstractions may be layered with multiple such layers.

The results of using LSTM on our data were quite good. The model was able to predict with mape 15%.

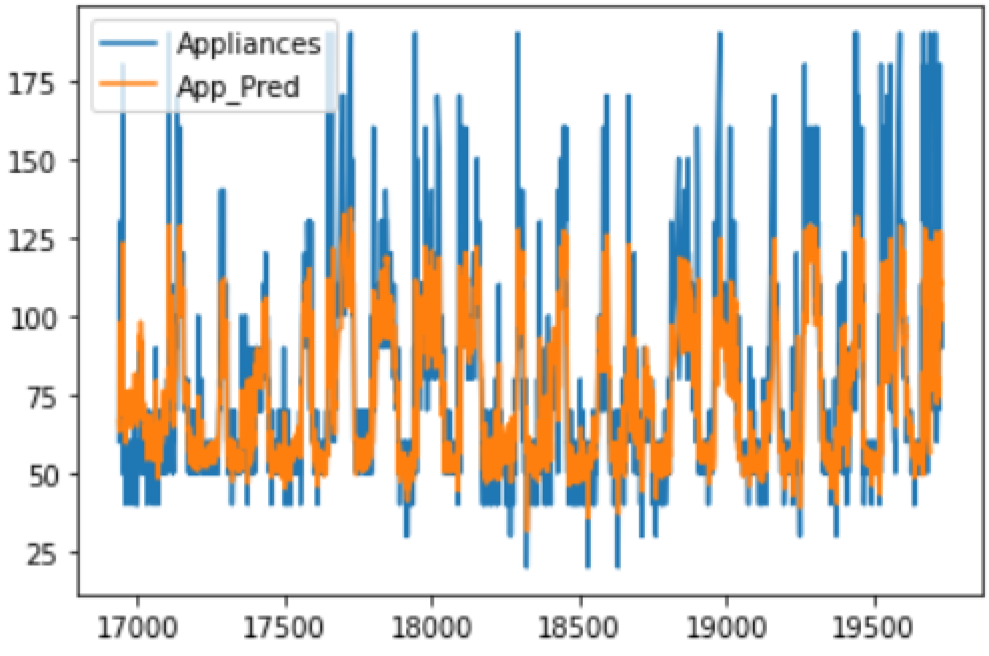

**Fig3. Actual vs predicted value**

## Time series analysis with facebook prophet

Time series data can be difficult and frustrating to work with, and the various algorithms that generate models can be quite finicky and difficult to tune. This is ***particularly*** true if we are working with data that has multiple seasonalities. In addition, traditional time series models like SARIMAX have many stringent data requirements like stationarity and equally spaced values. Other time series models like Recurrent Neural Networks with Long-Short Term Memory (RNN-LSTM) can be highly complex and difficult to work with.

Facebook Prophet is an open-source algorithm for generating time-series models that uses a few old ideas with some new twists. It is particularly good at modeling time series that have multiple seasonalities and doesn't face some of the above drawbacks of other algorithms. At its core is the sum of three functions of time plus an error term: growth `(t)`, seasonality `s(t)`, holidays `h(t)` , and error `e_t` :

$$y(t) \;=\; g(t) \,+\, s(t) \,+\, h(t) \,+\varepsilon$$

So, Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. This model predicts with mape 0.1% to 0.4%.

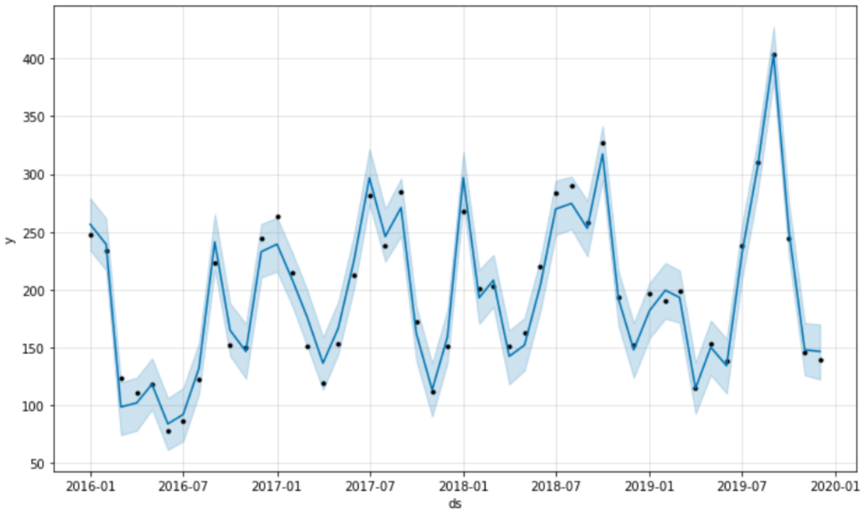

**Fig.4 Actual vs predicted value using prophet**

**Performances of other machine learning algorithms**

| **Model** | **Mean Squared Error** |
|---|---|
| Logistic Regression | 9345.76 |
| Random Forest Regression | 5188.23 |
| Support Vector Machines | 10394.54 |
| Classifier Gradient boosting | 6886.64 |
| XGBoost | 6828.62 |

Except Random forest regression all the other implementations were overfitting. Hence the Random forest reg is best of them all for this kind of dataset.

## Conclusion and future work

The LSTM can be considered good when the data to be processed is large enough to train its powerful neural network. The statistical data analysis has shown thought-provoking results in both the exploratory analysis and in prediction models. The models with only weather data ranked the pressure as the most important weather variable, followed by outdoor temperature, outdoor relative humidity. The possible explanation for why the pressure has strong prediction power may be related to its influence on the wind speed and higher rainfall probability which could potentially increase the occupancy of the house.

The results can be improved using the realtime data of physical quantities like temperature, humidity, wind speed rather than predicting them. The real time data can be obtained from weather forecasting agencies or manually installing the iot sensors in the house. Future work could also include considering weather data such as radiation and precipitation. Also occupancy and occupant's activity information could be useful to improve the prediction and find its relationship with other parameters. The wireless sensors could also measure co2 and noise to help in the prediction and to track the occupant's movement from room to room and time spent in each root.

The data and the processing scripts implemented in this report is made available on public repository: neon0047/TimeSeriesAnalysis (github.com)

## References

1. Time series - Wikipedia
2. Wang, M. L. (2018). Advanced Multivariate Time Series Forecasting Models. *Journal of Mathematics and Statistics*, *14*(1), 253-260. https://doi.org/10.3844/jmssp.2018.253.260
3. Multivariate time series forecasting | by Mahbubul Alam | Towards Data Science
4. End to End Multivariate Time Series Modeling using LSTM - YouTube
5. Khodabakhsh A., Ari I., Bakır M., Alagoz S.M. (2020) Forecasting Multivariate Time-Series Data Using LSTM and Mini-Batches. In: Bohlouli M., Sadeghi Bigham B., Narimani Z., Vasighi M., Ansari E. (eds) Data Science: From Research to Application. CiDaS 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 45. Springer, Cham. https://doi.org/10.1007/978-3-030-37309-2_10