Prediction of Road Traffic Congestion

Amit Gupta

MSC Data Science Research Proposal

**Proposed Research**

Time is the most precious resource that human being has. According to a survey an average American is spending 2340 mins in traffic and conditions are worst in many well-developed cities like Los Angeles, where on an average commuter are spending 6120 hours in traffic jams.

These non-productive hours could be used for many novel and humanitarian work. The goal of this research is to predict the traffic month in advance which will help commuter to plan his/her days & weeks well in advance. This will lead us to have a better work-life balance, less traffic and more productivity.

I will be using Minneapolis traffic volume data set for this analysis and research. Used dataset is a time series dataset and hence I will be using various time series algorithm like - Autocorrelation function (ACF), Partial autocorrelation function (PACF), Auto Regressive Moving Average Model (ARMA) and Auto Regression Integration Moving Average Model (ARIMA). Along with this I will perform comparative analysis on conventional time series and deep learning algorithm i.e. long short-term memory deep learning algorithm. Here are few key goals of this paper -

* Increase the efficiency of the existing roads by better traffic flow planning.
* Increase the productive hours and work life balance efficiency of human beings by reducing commute time.

**Previous Work**

Traffic is a worldwide problem and hence a lot of research and effort has been gone in this field. Hence, it’s not possible to include all the research in this paper. However, this paper will include most closely recent work in this field.

In the previous studies, traffic perdition problem has been solved by various methodologies. Most of the scholars/researchers have used state of art technology for the data collection and applied various soft computing techniques like random forest algorithm, KNN (k-nearest neighbor) and Deep neural network etc. [Prediction of Road Traffic Congestion Based on Random Forest](https://ieeexplore.ieee.org/document/8283291) is using Random forest algorithm to forecast the traffic density. Study evidences that the Timeseries data can be analyzed and predicted better with time series specific algorithms like Auto Regressive Moving Average Model (ARMA), Auto Regression Integration Moving

Average Model (ARIMA) etc. Long Short-Term Memory (LSTM) is a special kind of neural network that make predictions according to the data of previous times. The use of these methods has been extended to other fields as well. For example, Neural network has been used extensively in medical science and complex business scenarios. Parkinson, lung daises and asthma, skin cancer etc are just a few elementary examples. Similarly rest of the algorithms are also getting used to solve the various real time business problems.

[A K-nearest neighbor locally search regression algorithm for short-term traffic flow forecasting](https://ieeexplore.ieee.org/document/8321530) is using conventional K-nearest. The K-Nearest Neighbor (K-NN) method is a non-parametric regression and classification methodology. However, this algorithm’s accuracy needs to be closely monitored for the high dimensional feature dataset. Few researches has been performed to compare the KNN with Neural Network on the time series dataset and observed that neural network is more agile to the changes compare to KNN. ([Comparison of Neural Network and K-Nearest Neighbor Methods in Daily Flow Forecasting](https://scialert.net/fulltextmobile/?doi=jas.2010.1006.1010)). However, this hypothesis is vary based on data and its purity.

Another research paper [Short-term forecasting model of traffic flow based on GRNN](https://ieeexplore.ieee.org/document/6561614) is using artificial neural network. It’s a second-generation algorithm and hence its predictions capabilities are very accurate. However, this paper only talks about short term traffic forecasting. Hence, commuter will not be able to plan his day/week very well in advance.

This research paper, will explain the classic time series model and deep neural network (long short-term memory) model. Also, this will compare the brief output of both the approaches.

**Motivation**

Today number of vehicles are continuously increasing. In 2017, only USA had 263 million motors. This contributed to 811 [motor vehicles](https://en.wikipedia.org/wiki/Motor_vehicle) per 1,000 inhabitants compared to 798 vehicles per 1000 persons in 2013.

This rapid growth adds many disorders to the society like increase in pollution (air and noise), traffic congestion and fuel loss. Clearly traffic congestion needs to be considered a word wide problem and government bodies needs to start taking more concreate actions compare to last decade. Apart from environmental disorder long and hectic commute time adding human stress level which is resulting a productivity loss and work life imbalance.

Successful traffic congestion prediction will help commuter to plan his/her work in a week advance and result a less traffic on the road, less pollution (air & noise), less fuel wastage & most importantly less time to commute.

**Research question**

* How accurately can we forecast the traffic volume in Minneapolis, Minnesota for next 7 days? Does this pattern have any dependency on national holidays and climate condition?

**Proposed Methodology**

*Aim: -*

The aim of this study is to develop and apply a ‘scientific approach’ for improving the traffic congestion in metro cities.

This specific research paper has been developed using Minneapolis, Minnesota traffic data. However, this can be extended to any other metro city. In today’s world, rapid population growth, road traffic and congestion has become one of the major concerns for the commuter. In United States of America more than 14 million people are spending an hour or more traveling to work. This paper is a comparative analysis of various methodologies of traffic prediction.

The idea of building a robust and a smart traffic predictor is to help predict and alert various government bodies and individuals to take an appropriate measure well in advance. Additionally, this will help commuter to plan his/her day in a week advance.

*Objective: -*

To analyze the traffic congestion pattern for a metro city between year 2013 to 2018.

To identify the correlation between the United states national holiday and traffic flow.

This research paper will explore Time of day (TOD) methodology for traffic data analysis. TOD control divides a day into several intervals and tries to find optimal controls for each interval. This research paper will also use soft computing technique like Neural network along with classic time series algorithm to predict the traffic.

*Data* and methods used for analysis*: -*

This research paper will use publicly available dataset from UCI Machine learning Repository. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The chosen dataset is having an hourly inter-state 94 (aka I94) traffic volume. Roughly, I-94 connects Minneapolis and St Paul, MN. Hourly weather features and holidays included for impacts on traffic volume.

This data set is having almost 5 years hourly records from January 01, 2013 9:00 AM to September 09, 2018 09:00 AM. Overall data quality looks good apart from few minor issues. I will be documenting those steps under Data Processing and Cleansing section.

I will refer this dataset as ‘traffic dataset’ in this paper.

Mostly time series related business scenarios dataset has less data inconsistency compared to other datasets. Few of the timeseries related dataset example are Amazon US stock price or number of passengers for delta British Airways.

Traffic dataset cleansing steps –

1. *Missing records* – January 2015 to June 2015 records are missing. To impute the synthetic value, I will be using Last observation carried forward (LOCF) over the other available methodologies like mean imputation, median imputation or mode imputation. Mostly, the Last observation carried forward (LOCF) imputation method recommended when data is growing longitudinal (horizontal) to the time axis.
2. *National holidays* – Data are not formatted properly for the US National holidays.

*Proposed Machine Learning & Artificial Intelligence Algorithms: -*

1. *Classic Machine Learning Approach -*

Most often time series analysis looks very similar to simple regression (or ANN), the results obtained by regression alone are not enough for making an accurate forecast due its dynamic nature i.e. global and local trends. Traffic dataset is also a time series dataset with multiple features like national holiday, temperature, rain in past one hour, snow in past one hour, weather description and traffic volume. The objective is to find out the global and local trends in it and eliminate the white noise.

Here is a quick snapshot of the various internals’ factors in a time series record –

Yt= tt  + st + ε

Here tt represents global trend, st depicts Stationarity+ local trends and ε is denoted as white noise. In the notation, white nose is unpredictable and global trend and Stationarity can be predicted. This could have been introduced due to multiple reasons, such as issues with sensor, manual data entry mistake etc.

In this paper we will try to eliminate unpredictable element (aka white noise or ε) from the traffic dataset and try to find out the global trends along with seasonality (if any) in a dataset. Stationarityanalysis is an essential step for the time series prediction and analysis. A time series is called “Stationary”, if it’s time invariant. Which indicates that two different parts of time series will have the same Statistical measures like mean, median, variance and co-variance etc. A STS can have neither a global trend nor a seasonal component present in it.

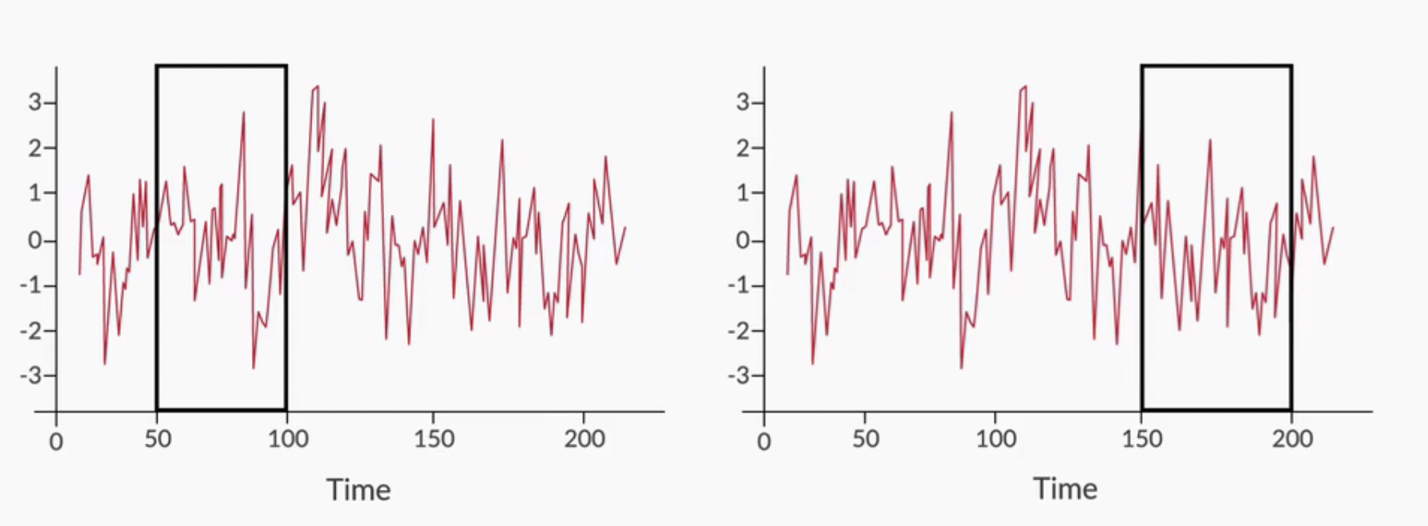


fig-01

In most of the practical scenarios its really difficult to maintain same Statistical measures and hence we will consider the weak Stationary rather Stationary.

This research paper will use following methods to find out Stationarity-

1. ACF - Autocorrelation function
2. PACF - Partial autocorrelation function
3. ADF - Augmented Dickey–Fuller

A sample ACF plot is depicted in fig-02. In this diagram, correlation value is **between the blue lines** (the upper and lower limits of the confidence interval), you can say that it isn't significantly different from zero; and hence, you can **take it to be zero.**

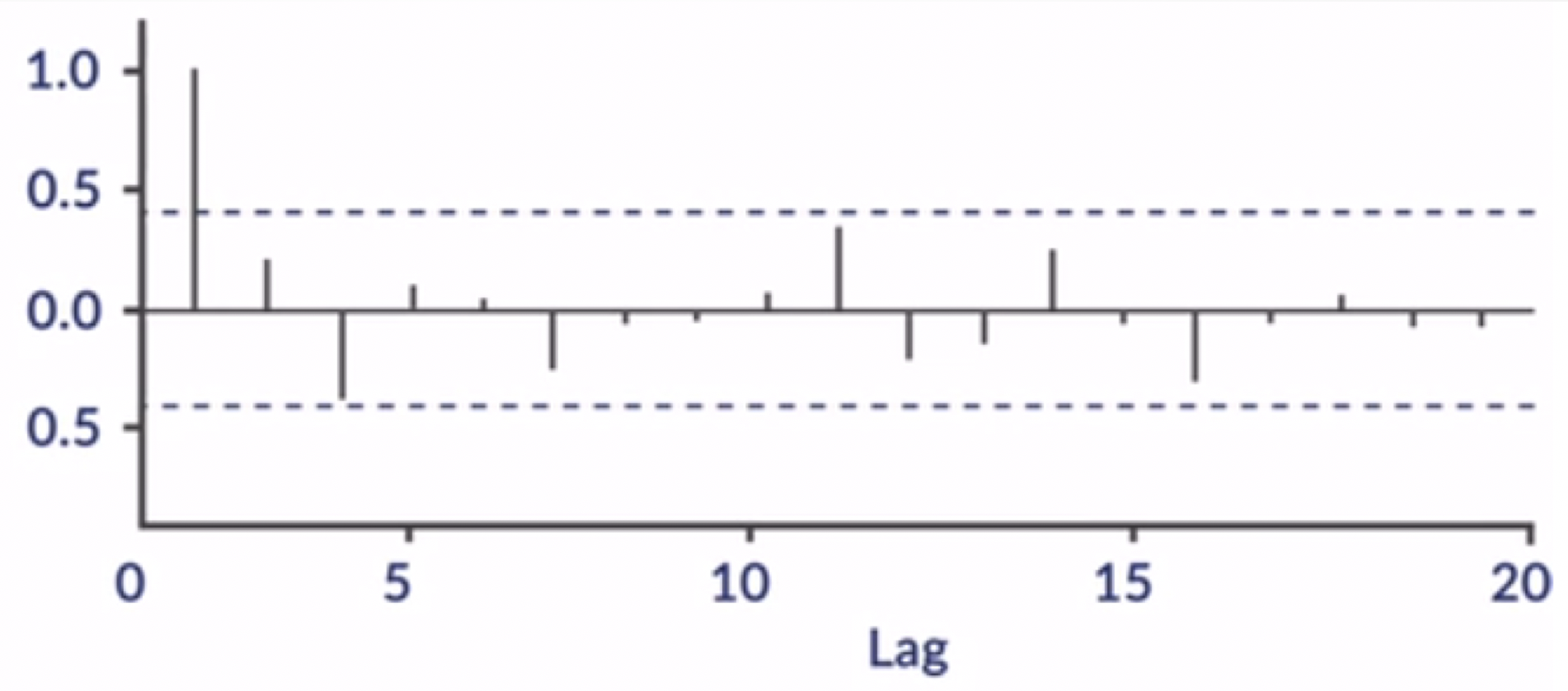


fig-02 (pure white noise representation through ACF)

1. *Deep Learning (LSTM) Approach -*

LSTMs are a very promising solution to sequence and time series related problems.

They are specially designed neural network to work with **sequential data,**i.e. data where natural notion of a 'sequence' is very import such as (sequences of words, sentences etc.), videos (sequences of images), speech etc. LSTM have been able to produce state-of-the-art results in fields such as natural language processing, computer vision, and time series analysis and hence this research paper will use LSTM to predict the future traffic. LSTM network is scientifically proven methodology to overcome the Gradient Descent problem.

*Required Software: -*

* Python
* Python packages for deep learning (Keras, TensorFlow, Statsmodels and Pmdarima)
* Microsoft Word
* Microsoft Excel

*Output: -*

I am expecting a working model which will be able to predict the Minneapolis traffic precisely and accurately. Additionally, this model will be able to depict the correlation between the national holidays and the traffic volume.

This research paper will use adam/ rmsprop and root mean square error (RMSE) for

optimizer and loss function respectively for metric collections and quality control.

**Project Plan and Risk or Contingency Plan**

* What are the ethical/legal risks?

*NA*

Are there any risks? And if yes, then what is the plan to mitigate them?

*NA*

* Will you collect data on people? If yes, do you have ethical approval?

*NA*

* Will you use existing data? If yes, do you need and do you have approval?

*Yes, I will be using the existing data. Its available on public domain and hence No Approval Required.*

* Is coding an issue? Do you have access to the right coding library?

*I am very comfortable with computer programming. In this research paper I will be using Python and I have working experience in it. I do have all the access to the required resources.*

* Do you have access to the software?

*Yes*

* What if your method will not work out? Is there such a risk?

*I believe the above hypothesis should work well with the classical Time Series approach and with the Deep leaning (LSTM) model. In case of any contingency, we need to relook traffic dataset and try to remove the anomalies (if any minor) to make it suitable for analysis.*

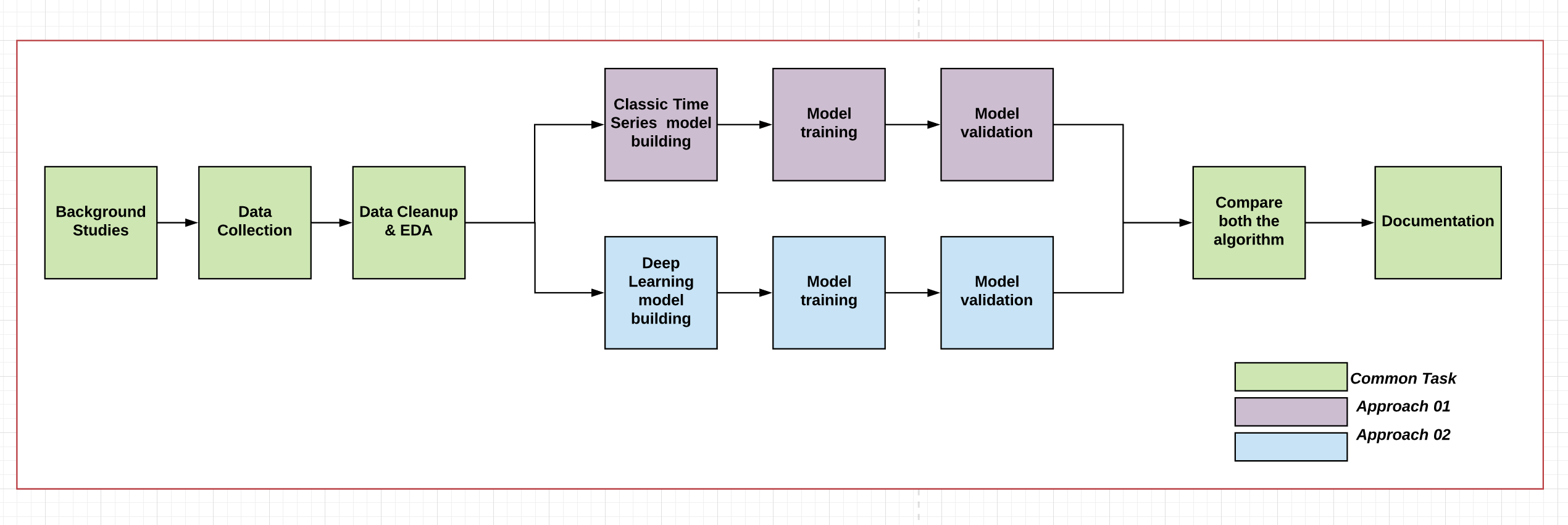
*If still model is not performing well, it would be an indicator of pure white noise and in that scenario, we need to explore a new traffic dataset and I am very much aware of this risk.*

* Are there other risks? Too small dataset? Too much missing data?

*Traffic dataset is optimal in size.*

*Other risk - Deep learning models (specially LSTM) is a compute heavy operation and hence it will need a sophisticated hardware to run these algorithms.*

* Can use a PERT chart, perform the Critical Path Analysis



* What is the contingency plan in case of any risks?

*Mentioned above*

* Can use flowcharts/diagrams with proper references

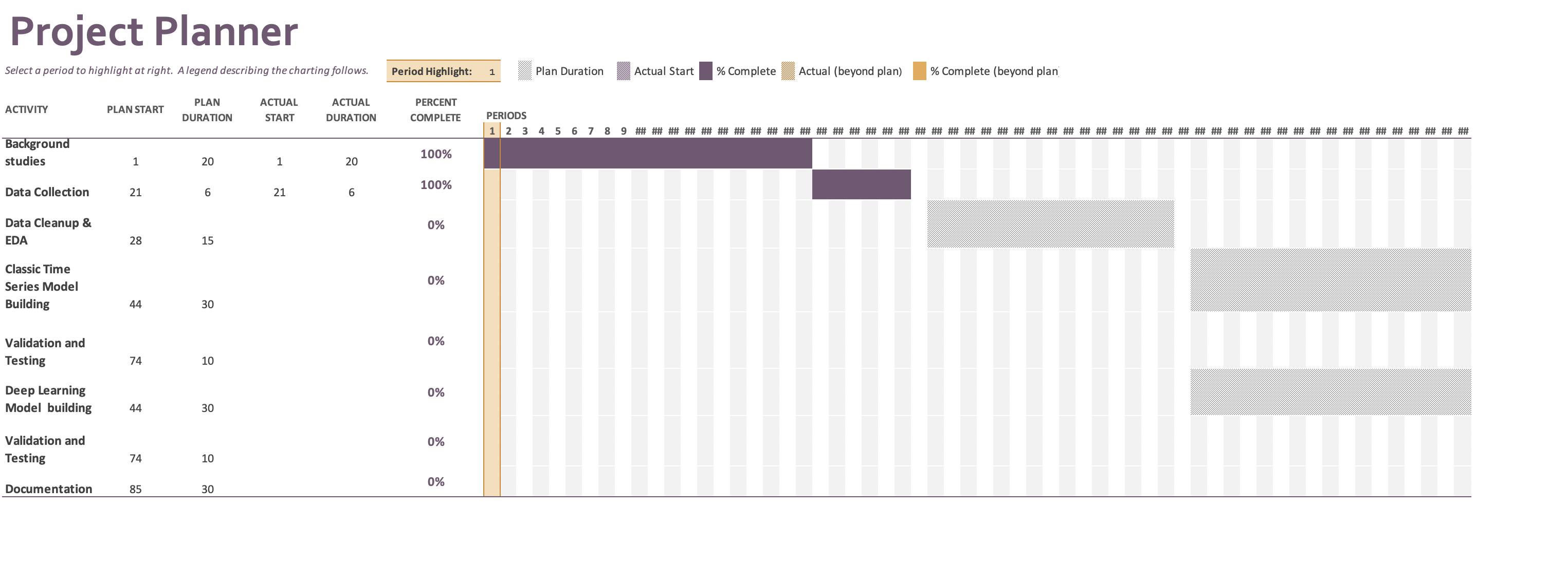
*NA*

**Timelines**

*work breakdown structure: -* Here are the list of independent tasks, which needs to be performed as part of this research project.

1. Background studies
   1. Literature review
2. Data Collection
3. Selecting the right algorithm
4. Implement the Model
   1. Classic Time Series Model Building
   2. Deep Neural Network Model Building
5. Validate & Test the results
6. Documentation

*Gantt Chart: -*



**Reference**

* [http://statistics.brussels/figures/did-you-know/how-much-time-on-average-did-people-spend-stuck-in-traffic-jams-in-brussels-during-2017#.XWN8yJMzbyU](http://statistics.brussels/figures/did-you-know/how-much-time-on-average-did-people-spend-stuck-in-traffic-jams-in-brussels-during-2017)
* <https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test>
* <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
* [http://ai.dinfo.unifi.it/paolo//ps/tnn-94-gradient.pdf](http://ai.dinfo.unifi.it/paolo/ps/tnn-94-gradient.pdf)
* <https://skymind.ai/wiki/lstm>
* <http://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume>
* Jonny Evans, Ben Waterson, Andrew Hamilton [Forecasting road traffic conditions using a context-based random forest algorithm](https://www.tandfonline.com/doi/full/10.1080/03081060.2019.1622250?scroll=top&needAccess=true)
* Haipeng Lu, [Fan Yang](https://ieeexplore.ieee.org/author/37085862746) [A Network Traffic Prediction Model Based on Wavelet Transformation and LSTM Network](https://ieeexplore.ieee.org/document/8663884/authors#authors)
* Yanguang Cai, Helie Huang, Hao Cai , Yuanhang Qi [A K-nearest neighbor locally search regression algorithm for short-term traffic flow forecasting](https://ieeexplore.ieee.org/document/8321530)
* Pedro Lopez-Garcia, Enrique Onieva, Eneko Osaba, Antonio D. Masegosa, Asier Perallos [A Hybrid Method for Short-Term Traffic Congestion Forecasting Using Genetic Algorithms and Cross Entropy](https://ieeexplore.ieee.org/abstract/document/7331637)
* Yunxiang Liu, Hao Wu [Prediction of Road Traffic Congestion Based on Random Forest](https://ieeexplore.ieee.org/document/8283291)
* Ziwen Leng, Junwei Gao,  Yong Qin, Xin Liu, Jing Yin [Short-term forecasting model of traffic flow based on GRNN](https://ieeexplore.ieee.org/document/6561614)
* Journal - Alireza Eskandarinia, Hadi Nazarpour, Mehdi Teimouri and Mirkhalegh Z. Ahmadi [Comparison of Neural Network and K-Nearest Neighbor Methods in Daily Flow Forecasting](https://scialert.net/fulltextmobile/?doi=jas.2010.1006.1010)

**Keywords and abbreviations**

1. Statistical stationarity
2. White noise
3. ACF - Autocorrelation function
4. PACF - Partial autocorrelation function
5. AR model- Auto Regressive model
6. MA model – Moving Average
7. ARMA model – Auto Regressive Moving Average Model
8. ARIMA Model – Auto Regression Integration Moving Average Model
9. SARIMA – Seasonality ARIMA
10. TSA – Time series analysis
11. LOCF - Last observation carried forward
12. LSTM – Long short-term Memory
13. DL – Deep learning
14. ANN – Artificial Neural network
15. STS – Stationary Time series
16. ADF - Augmented Dickey–Fuller
17. RNN – Recurrent neural network