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 week in advance which will help commuter to plan his/her days better. This will lead us to have a better work-life balance, less traffic, better ecosystem in terms of pollution, fuel saving and more productivity.
This paper will be using metro city traffic data set for research and analysis. It's a time series dataset and hence 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) will be analyzed. Along with this research paper will also analyze a comparative study on conventional time series and deep learning algorithm (i.e. long short-term memory) deep learning algorithm.

Previous Work

Traffic is a worldwide problem since last couple of decades. Hence, a lot of researchers/scholars have put tremendous efforts to solve this problem. Therefore, it's not possible to include all their work in this paper. However, this paper will include most recent and closely associated work in this field.

In the previous studies, traffic prediction 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, KNN (k-nearest neighbor) and Deep neural network etc. Yunxiang Liu (Prediction of Road Traffic Congestion Based on Random Forest) suggested to solve this problem using Random forest algorithm. Study evidences that the Timeseries data can be analyzed and predicted better with time series specific algorithms such as Auto Regressive Moving Average Model (ARMA), Auto Regression Integration Moving Average Model (ARIMA), and Long Short-Term Memory (LSTM) etc. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. 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 such as but not limited to Parkinson, lung daises and asthma, skin cancer. Similarly, rest of the algorithms are also getting used to solve various real time business problems.

Yanguang Cai, A K-nearest neighbor locally search regression algorithm for short-term traffic flow forecasting analyzes a traffic forecasting using conventional K-nearest. Dataset always have some outliers; these outliers will impact the k nearest neighbors' outputs which may have a bad impact on the prediction value. The K-Nearest Neighbor (K-NN) method is a non-parametric regression and classification methodology. Alireza Eskandarinia, Comparison of Neural Network and K-Nearest Neighbor Methods in Daily Flow Forecasting has performed a comparison to the KNN with Neural Network on the time series dataset and concluded that neural network is more agile and robust for the new (non-trained) data compared to KNN.

Another research paper Ziwen Leng, Short-term forecasting model of traffic flow based on GRNN is using artificial neural network. It's a second-generation algorithm and hence its predictions capabilities are very accurate. It uses simple pass learning and hence there is no back propagation in GRNN. The only limitation is about its size. The network size grows very rapidly which requires high computation resources.

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

Motivation

Today, number of vehicles are continuously increasing. In 2017, United States of America (USA) had 263 million motors and this contributed to 811 motor vehicles per 1,000 inhabitants compared to 798 vehicles per 1000 persons in 2013.

This rapid growth introduced many disorders to the society and the ecosystem. A few of them are pollution growth (air and noise), traffic congestion and fuel loss. Traffic congestion should be considered a worldwide problem and government bodies need to start taking more measurable and concrete actions. Apart from environmental disorders, long and hectic commute time adds to human stress level which is resulting in loss of productivity and work life imbalance.

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

Research question

How accurately we can forecast the traffic volume in Minneapolis, Minnesota for next seven days? Does traffic flow pattern have a dependency on national holidays?

Proposed Methodology

Aim: -

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

Road Traffic congestion have become one of the major concerns for the commuter. In the United States of America, more than 14 million people are spending an hour or more travelling to work. This paper is a comparative analysis of various methodologies of traffic prediction. This specific research paper has been developed using Minneapolis, Minnesota traffic data. However, the approach is generic and can be extended to any other metro city.

The idea of building a robust and smart traffic predictor is to 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 or week in advance.

Objective: -

- To analyze the traffic congestion pattern for a metro city between years 2013 to 2018.
- To identify the influence of national holiday on the United States traffic flow.

This research paper will explore "Time of Day (TOD)" methodology for traffic data analysis. Times of Day divides a timeframe into several smaller intervals and tries to find optimal model for each interval. This paper will apply the soft computing technique like Neural network and classic time series algorithm on the TOD data to predict the traffic.

Data and methods used for analysis*:* -

This research paper will use publicly available dataset from UCI Machine learning Repository. As per the UCI site 'The UCI Machine Learning Repository is a collection of databases, 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 I-94) traffic volume. I-94 connects Minneapolis and St Paul. Hourly weather features and holidays included in the traffic dataset.

Traffic data set has hourly records from January 01, 2013 9:00 AM to September 09, 2018 09:00 AM (~5 Years). Overall data quality is good however there are a few minor issues. Data cleansing steps will be documented 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 time series related dataset 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 is recommended when data is growing longitudinal (horizontal) to the time axis.
2) *National holidays* – Data is 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 –

   $$Y_t = t_t + s_t + \varepsilon$$

   Here t_t represents global trend, s_t 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. Stationarity analysis is an essential step for the time series prediction and analysis.

2) *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: -
The end product will be 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. It is available on a public domain and hence no approval is required.

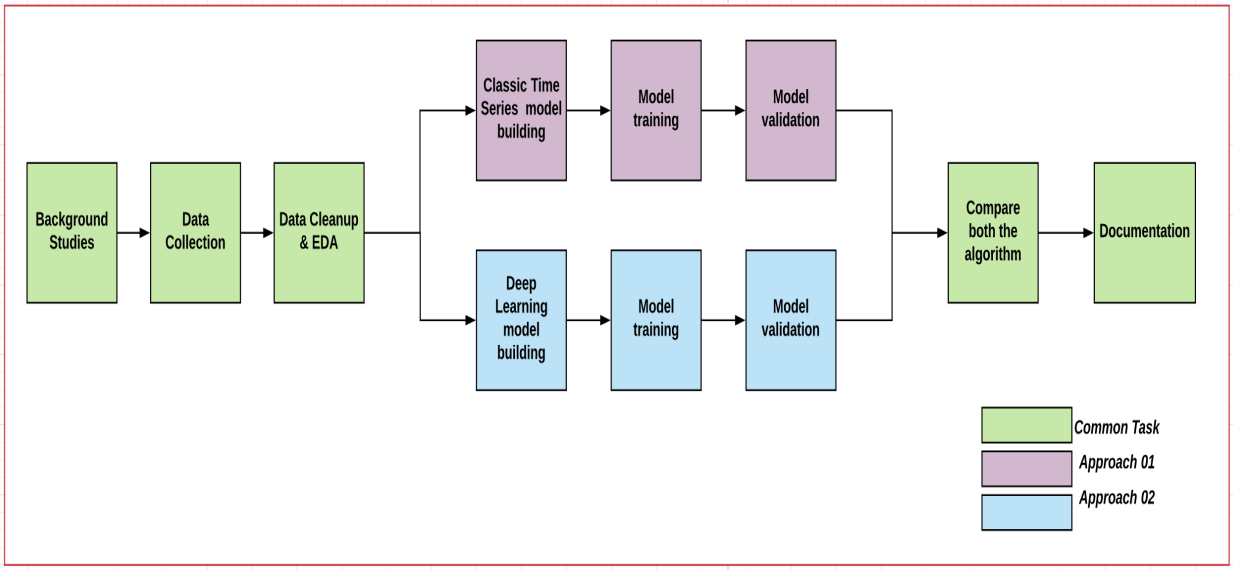
- Is coding an issue? Do you have access to the right coding library?
  I am very comfortable with computer programming. In this research paper, I will be using Python and I have experience with working on Python. I do have all the access to the required resources.

- Do you have access to the software?
  Yes

- What if your method will not work out? Is there such a risk?
  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 workaround associated with the possible risk.

- Are there other risks? Too small dataset? Too much missing data?
  Traffic dataset is optimal in size.
  Other risk - Deep learning models (especially 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

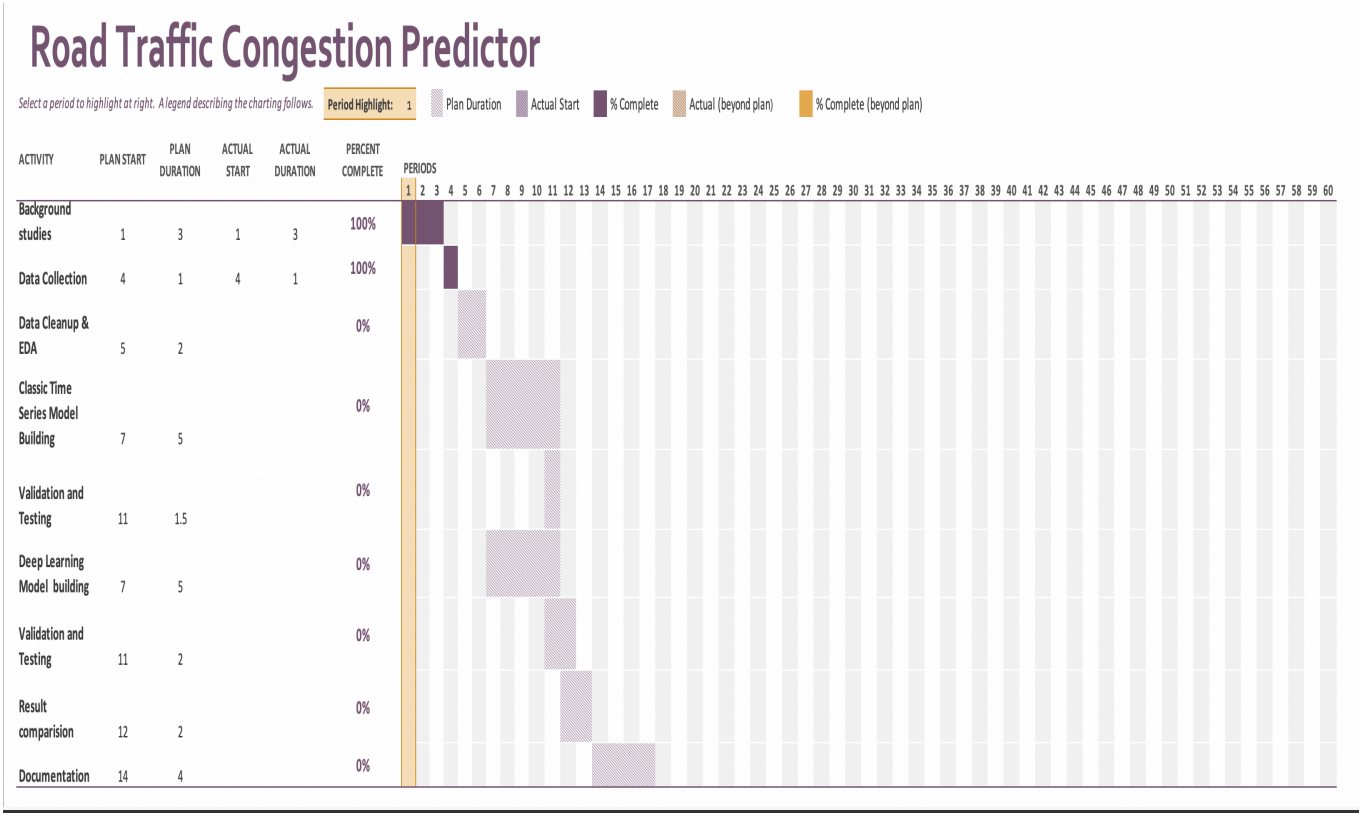
Timelines

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

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

Gantt Chart: -

In the below Gantt chart Time (X access Or Horizontal access) is in weeks and Y access (Vertical access is representing the activities).



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
- 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
- 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
- Haipeng Lu, Fan Yang A Network Traffic Prediction Model Based on Wavelet Transformation and LSTM Network
- Yanguang Cai, Helie Huang, Hao Cai , Yuanhang Qi A K-nearest neighbor locally search regression algorithm for short-term traffic flow forecasting
- 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
- Yunxiang Liu, Hao Wu Prediction of Road Traffic Congestion Based on Random Forest
- Ziwen Leng, Junwei Gao, Yong Qin, Xin Liu, Jing Yin Short-term forecasting model of traffic flow based on GRNN
- Journal - Alireza Eskandarinia, Hadi Nazarpour, Mehdi Teimouri and Mirkhalegh Z. Ahmadi Comparison of Neural Network and K-Nearest Neighbor Methods in Daily Flow Forecasting

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