Congestion predictor

Summary of 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. I want to predict the traffic well in advance which will help commuter to plan his/her day & 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 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 ACF, PACF, ARMA and 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.
- Reduces traffic volume at a particular section.
- Increase the productive hours and work efficiency of human being

Research question

- What effect does national holidays have on the Minneapolis traffic?

  OR
- What is the traffic volume pattern in Minneapolis, Minnesota? Does this pattern have any relation with national holidays and climate condition?

- **Why are you doing this research????**

Details of the research project

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

However, this specific research paper has been developed using the data of Minneapolis, Minnesota. In today's world with increasing population, traffic and congestion has become one of the major concerns for humans with total commute time rising rapidly. In 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 well in advance.

Data

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 in next section (under Data Processing).

I will refer this dataset as 'traffic dataset' in this paper.

Processing of data and the analysis

Data Processing

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.

Research Analysis

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 (global and local trends).

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 (ε) from the given data set and try to find out the global trends along with seasonality (if any).

Stationarity analysis 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.

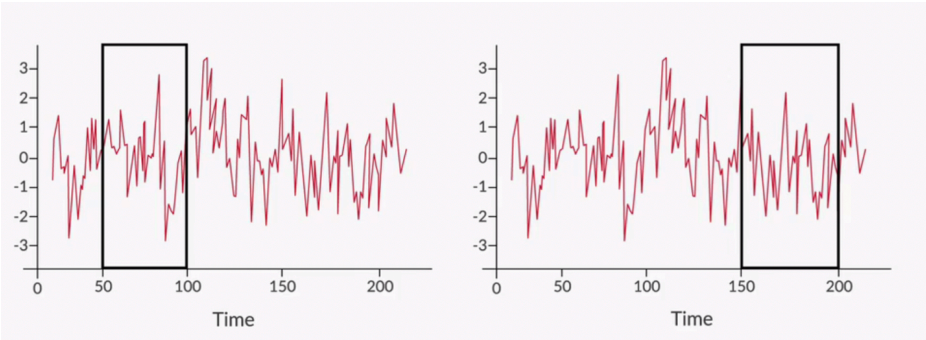


fig-01

In most of the practical scenarios its really difficult to maintain same Statistical measures and hence we will consider the weak Stationary rather Stationary. Weak Stationary preserves the pairwise operations. However, in this research paper we will continue to call weak Stationary as Stationary. In the below example group of five values will have the same Statistical measures.

Timestamp	Values
T_1	230
T_2	280
T_3	490
T_4	190
T_5	250
T_6	320
T_7	380
T_8	210
T_9	220
T_{10}	290
T_{11}	286
T_{12}	158
T_{13}	400
T_{14}	230
T_{15}	260

LEG 1	
Timestamp	Values
T_1	230
T_2	280
T_3	490
T_4	190
T_5	250
T_6	320
T_7	380
T_8	210
T_9	220
T_{10}	290
T_{11}	286
T_{12}	158
T_{13}	400
T_{14}	230
T_{15}	260

LEG 2	
Timestamp	Values
T_1	230
T_2	280
T_3	490
T_4	190
T_5	250
T_6	320
T_7	380
T_8	210
T_9	220
T_{10}	290
T_{11}	286
T_{12}	158
T_{13}	400
T_{14}	230
T_{15}	260

fig-02

Sample ACF plot is depicted in fig-03. In this diagram, correlation value is **between the blue lines** (that signify 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.**

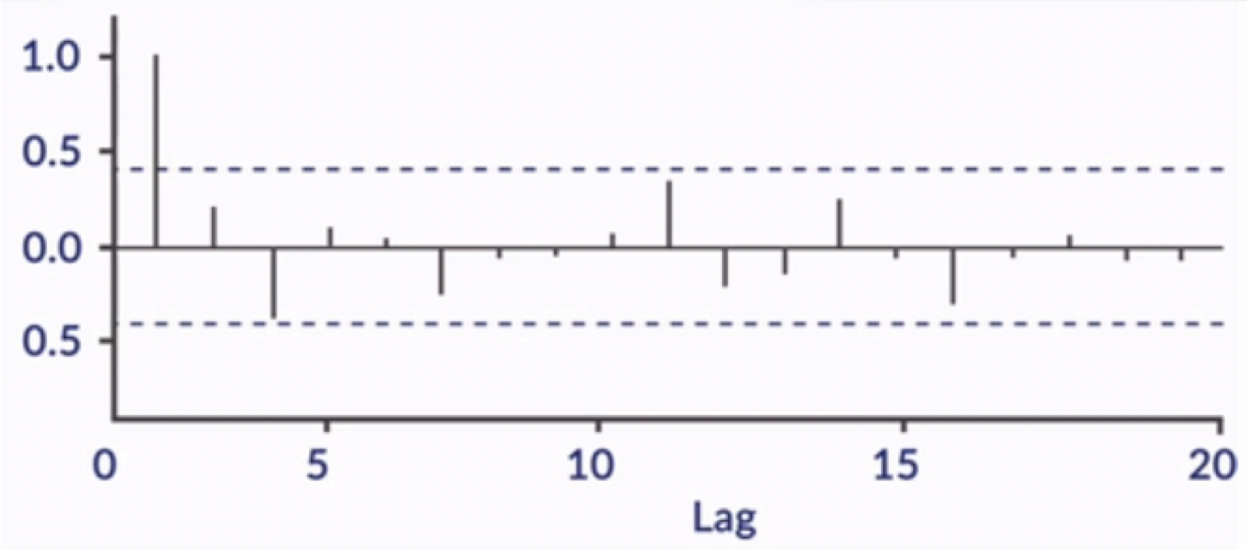


fig-03 (pure white noise representation through ACF)

Mostly White noise and Stationary time series looks very similar. White noise contains a series of value in which each value in the series is completely independent of every other value. If the series is a white noise, the values in it will belong to a gaussian (normal) distribution.

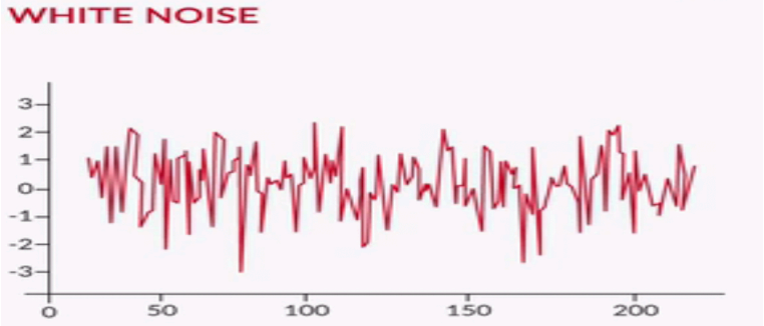


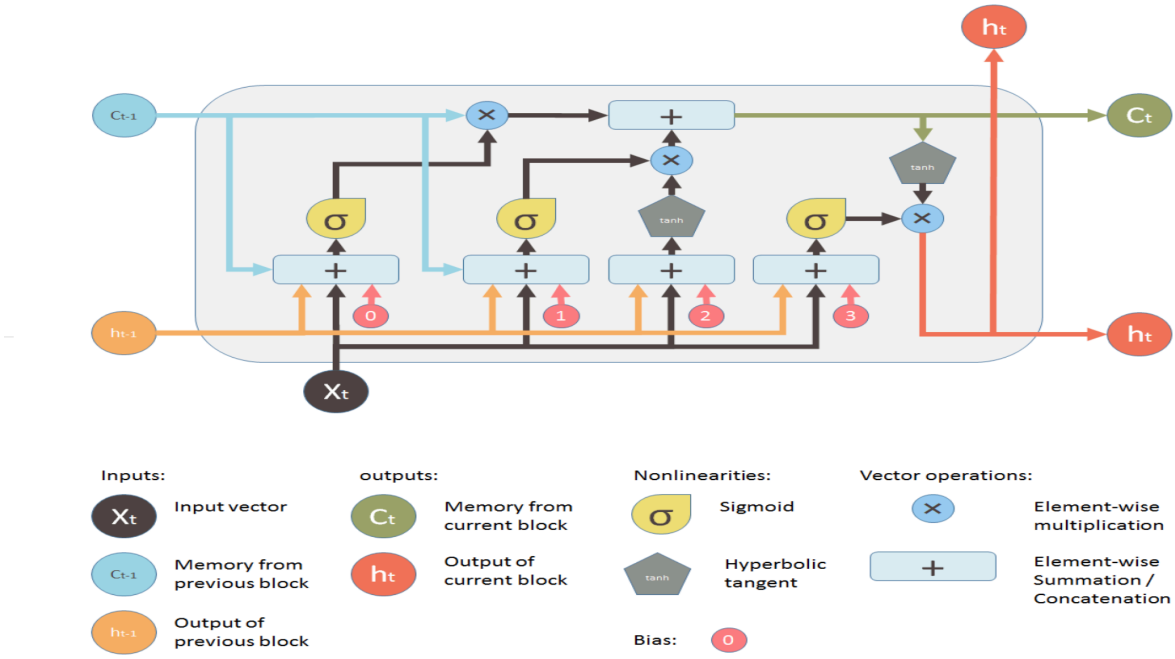
fig-03

To perform a Stationary test, this research paper will use ADF methodology, which is an extension of Dickey–Fuller test. Stationary test will be performed on the traffic volume feature to identify the existence of the underlaying trends and seasonality. Autocorrelation function and Partial Autocorrelation function methodology will be used to identify the model which will be the best fit for a given dataset.

Once we have model ready, I will introduce the exogenous variable holiday. This will generate a model based on traffic volume and holiday (national).

Deep learning model for time series prediction

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.



The fundamental to the LSTM is a cell state (C_t) the horizontal line running through the top of the diagram. This cell state flows across the network with minor liner interactions. LSTM network adds or eliminates information to the cell after performing the sigmoid operation on Input and Output Vector. This layer is called as Forget layer.

$$f_t = \alpha\,(W_f.[h_{t-1},x_t]+b_f)$$

In the above-mentioned weighted sum formula, x_t is traffic volume vector which will be fed to this neural network. h_{t-1} is representing output of t-1 timestamp.

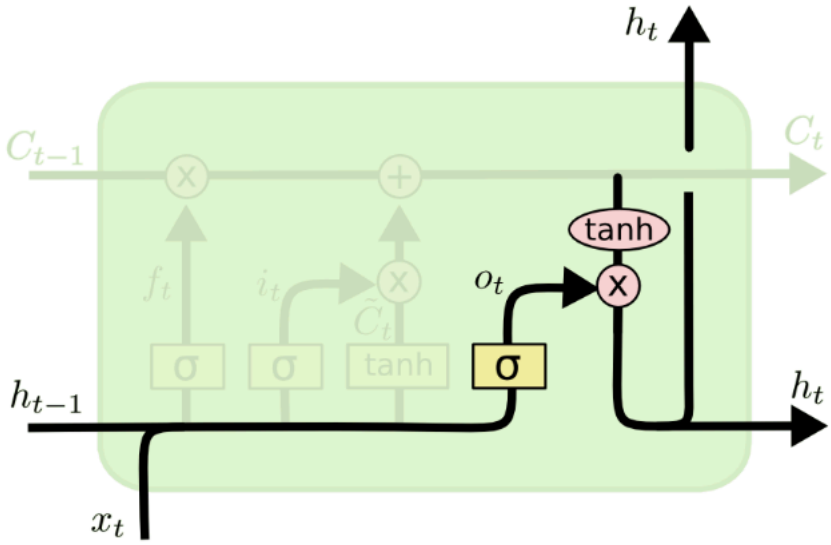
Input gate layer, updates the old cell value (if required) C_{t-1} into a new Cell state C_t.

$$i_t = \alpha\,(W_i.[h_{t-1},x_t]+b_i)$$

$$Ç_t = \tanh\,(W_C.[h_{t-1},x_t]+b_C)$$

$$C_t = f_t * C_{t-1} + i_t * Ç_t$$

Finally, LSTM generates output which is based on cell state and weighted sum of input and output of previous block.



$$o_t = \alpha\,(W_o.[h_{t-1},x_t]+b_o)$$

$$h_t = o_t * \tanh\,(C_t)$$

This research paper will use multiple neural network layers (deep neural network) to predict the future traffic. Also, I am planning to explore adam/ rmsprop and root mean square error for optimizer and loss function respectively.

Required Software

Python
Python packages for deep learning (Keras, TensorFlow, Statsmodels and Pmdarima)
Microsoft word

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.

Project Plan and Risk or Contingency Plan

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.

Deep learning models (specially LSTM) is a compute heavy operation and hence it will need a sophisticated hardware to run these algorithm.

References

- 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

Keywords and abbreviations

1. Statistical stationarity
2. White noise
3. ACF - Autocorrelation function
4. PACF - Partial autocorrelation function (PACF)
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