Thesis Proposal

Markus Kunej, 1005351897

October 28, 2022

Background

Air pollution is the greatest environmental burden to human health. It contributes to over eight million premature deaths globally every year, of which 14,400 are in Canada [1]. Two ways people cause air pollution are constant use of an automobile for transportation and high energy consumption. Thankfully, both of these can be reduced through urban densification, the process of increasing the density of people living in urban areas. When implemented along with accessible public transit, urban densification reduces the distance people need to travel for work and leads to a lower average energy consumption due to the efficiency of high-density buildings [2]. It is also associated with a reduced land take, which allows for greater availability of land for agriculture, nature, and biodiversity.

Urban densification often requires a major construction overhaul. This is especially true for projects that require deep pits to be dug, such as for high-rise buildings or underground transportation. These plans can lead to construction lasting up to a decade long [3]. As a result of the heavy machinery, increased congestion on the roads, and the release of particles (like dust) into the air, major construction sites lead to an increase in air pollutant levels. A case study in Qingyuan, China, found the daily concentration of total suspended particulates (TSP) near a construction site increased by 42.24%. It also saw a 16.27% increase in particulate matter 2.5 ($PM_{2.5}$), which refers to particles that are 2.5 microns or less in width [4]. $PM_{2.5}$ pose the greatest health risks, since due to their small size, they can lodge deeply into the lungs [5]. So, while there are long-term benefits to urban densification, the short-term decline in air quality caused by construction sites is worrying.

Air quality is popularly measured with the US Air Quality Index (AQI). This index considers the following pollutants: Ozone (O_3), $PM_{2.5}$, PM_{10}, Carbon Monoxide (CO), Sulfur Dioxide (SO_2) and Nitrogen Dioxide (NO_2) [6]. Weather forecast providers display the AQI value for a given location, usually for an entire city. However, due to environmental and location-based factors, such as major construction, the given AQI value may not be representative of all sub-regions within a given location [4].

There has been past analysis between air quality and construction sites. In 2022, a study was released showing correlation analysis between air pollutants and construction sites in Hangzhou, China. They used non-linear regression models to forecast future air pollutant levels [7]. Some limitations from this study I hope to address are:

a) The lack of consideration for other factors like wind direction, temperature, humidity, etc.
b) The relatively poor performance of the models

Hypothesis and Proposal

Due to the widely scoped AQI values issued by weather forecast providers and an increase in air pollutants near major construction sites, I believe residents living nearby are receiving lower-than-actual AQI forecasts, and are unknowingly having their health put at risk.

Using the data collected near the major construction site at Yonge & Eglinton in Toronto, I will develop two models to better understand the correlations between environmental factors and air quality, and more accurately forecast future air pollutant levels near major construction sites.

Data

The data is being provided by 10 measuring devices near the major construction site of Yonge & Eglington. It is considered a major construction site since a new light rail line, the Eglinton Crosstown, is being built, as well as multiple high-rise buildings. The devices measure the following pollutant levels: NO, NO_2, CO, CO_2, O_3, PM_1, $PM_{2.5}$, PM_{10}, and the AQI / Air Quality Health Index (AQHI). They also measure the following environmental factors: temperature, relative humidity, air pressure, wind speed, wind direction, and noise levels. The devices capture and upload data every minute and were installed on July 22nd, 2022.

Methods

I plan to use two different modelling techniques: a simpler statistical one and a more complex deep-learning (DL) model. The reasoning for these two approaches is that the statistical model can serve as a baseline for the deep-learning model. It requires less computations and is generally more interpretable. Should the prediction accuracy be poor, then it suggests more intricate dependencies between the variables exist, and that a DL model should be used next. While it would be harder to interpret interdependencies, a DL model should be able to achieve higher prediction accuracy because of its ability to form complex dependencies between inputs. Predictions will also be compared to historical AQI forecasts provided by different weather services for the area, noting any variances.

Since the measurements are taken every minute, and many of the data variables exhibit clear trends over a period of time (temperature, wind speed, etc.), this should be considered a time series forecasting problem, thus requiring time series models. There is also more than one time-dependent variable, and I suspect there exists interdependencies between them. This turns the problem into a Multivariate Time Series (MTS) one, therefore needing a vectorized model.

Statistical Model: Vector Auto Regressive Integrated Moving Average (VARIMA)

VARIMA is the vectorized version of ARIMA, a widely used time series forecasting technique. It combines an Autoregressive model, which considers past values of the variable, a Moving Average model, which accounts for past forecast errors, and an Integration step, which differences out seasonal trends and patterns [8].

First, use Granger Causality Test to investigate causality of each data variable with each other. If none of the variables are causal, then this indicates it is a good candidate for VARMA modelling [8]. Next, check whether each variable is stationary or not using the augmented Dickey-Fuller (ADF) [8]. There will likely be non-stationary data with this project, since there are seasonal trends and patterns with variables like temperature, wind speed, etc. If a variable is not stationary, difference it. Finally proceed with the model. Dependencies between variables can be explored afterwards. One technique is to compute the Impulse Response Function (IRF), which traces the effects of an innovation shock to one variable on the response of all variables in the system [8].

Deep Learning Model: Long Short-Term Memory (LSTM) Network

A LSTM network is essentially a more sophisticated version of a Recurrent Neural Network (RNN), which is considered the “standard” network for handling time-series data, due to its use of a feedback loop (using prior outputs as inputs).

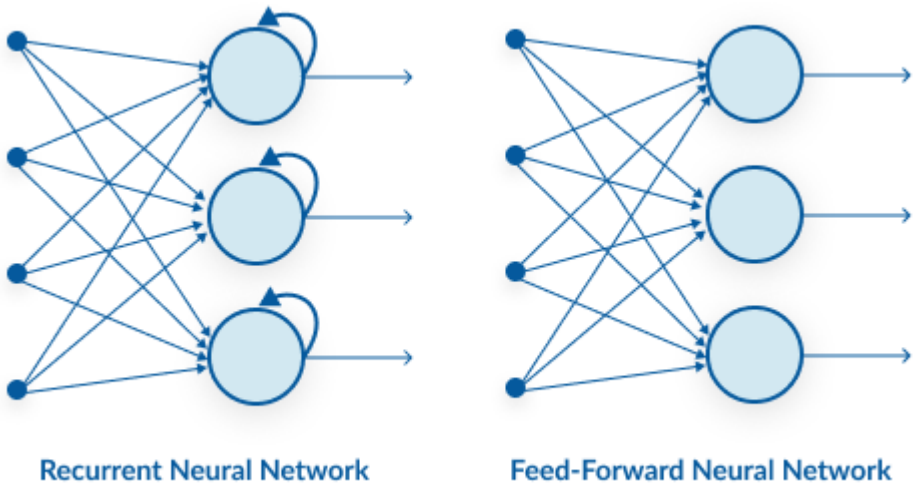

Figure 1 Comparison between a RNN and FNN [9]

The difference is LSTMs include a "memory cell" that can remember information for longer periods of time. This lets the network learn longer-term dependencies. LSTMs also deal with the vanishing and exploding gradient problem introduced by RNNs, by including new "input" and "forget" gates. These gates allow for better preservation of long-range dependencies [10][11].

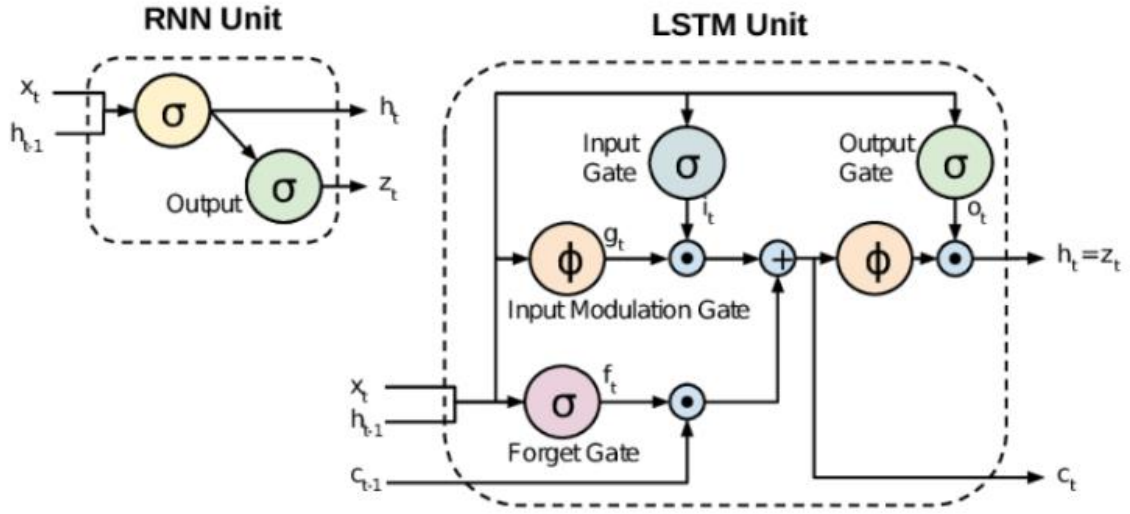

Figure 2 A LSTM unit is a more sophisticated RNN unit [10]

The downside to using a deep learning model is the lack of transparency within the network. It is much more complex compared to VARIMA, making it harder to extrapolate dependencies/importance of individual input variables. However, there are still methods to gain "some level" of understanding of the weights given to each variable. A LSTM is also more computationally expensive compared to VARIMA, but it can result in a better prediction, especially if the problem is quite complex with many different interdependencies [10][11].

References

[1] *Evans Research Group*. [Online]. Available: https://www.labs.chem-eng.utoronto.ca/evans/research/linking-emissions-and-health/. [Accessed: 22-Oct-2022].

[2] J. Teller, "Regulating urban densification: What factors should be used?," *Buildings and Cities*, vol. 2, no. 1, pp. 302–317, 2021.

[3] M. Draaisma, "Ontario premier says construction underway on New Toronto subway line | CBC news," *CBCnews*, 27-Mar-2022. [Online]. Available: https://www.cbc.ca/news/canada/toronto/ontario-line-official-breaking-ceremony-toronto-doug-ford-john-tory-1.6399282#. [Accessed: 22-Oct-2022].

[4] H. Yan, G. Ding, H. Li, Y. Wang, L. Zhang, Q. Shen, and K. Feng, "Field evaluation of the dust impacts from construction sites on surrounding areas: A city case study in China," *Sustainability*, vol. 11, no. 7, p. 1906, 2019.

[5] "Health consequences of air pollution on populations," *World Health Organization*, 15-Nov-2019. [Online]. Available: https://www.who.int/news/item/15-11-2019-what-are-health-consequences-of-air-pollution-on-populations. [Accessed: 24-Oct-2022].

[6] "Technical assistance document for the reporting of Daily Air ... - airnow," *AirNow*, 2018. [Online]. Available: https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf. [Accessed: 24-Oct-2022].

[7] H. Li, A. Cheshmehzangi, Z. Zhang, Z. Su, S. Pourroostaei Ardakani, M. Sedrez, and A. Dawodu, "The correlation analysis between air quality and construction sites: Evaluation in the urban environment during the COVID-19 pandemic," *Sustainability*, vol. 14, no. 12, p. 7075, 2022.

[8] X. Chen, "A multivariate time series modeling and forecasting guide with python machine learning client for SAP HANA," *SAP Blogs*, 12-Jul-2021. [Online]. Available: https://blogs.sap.com/2021/05/06/a-multivariate-time-series-modeling-and-forecasting-guide-with-python-machine-learning-client-for-sap-hana/. [Accessed: 15-Oct-2022].

[9] A. Pai, "Ann vs CNN vs RNN: Types of neural networks," *Analytics Vidhya*, 19-Oct-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/. [Accessed: 15-Oct-2022].

[10] A. Tripathi, "What is the main difference between RNN and LSTM: NLP: RNN VS LSTM," *Data Science Duniya*, 18-Jul-2022. [Online]. Available: https://ashutoshtripathi.com/2021/07/02/what-is-the-main-difference-between-rnn-and-lstm-nlp-rnn-vs-lstm/. [Accessed: 16-Oct-2022].

[11] Mao76, "What is the difference between LSTM and RNN?," *Artificial Intelligence Stack Exchange*, 23-Feb-2020. [Online]. Available: https://ai.stackexchange.com/questions/18198/what-is-the-difference-between-lstm-and-rnn. [Accessed: 16-Oct-2022].