Spatial Analysis on Total Delayed Minutes of Toronto Transit Commission Subway Stations

GISC9318D4b

4/16/2018
KW Inc.
am13to

Karen Wong, B.Sc.
135 Taylor Road
Niagara-on-the-Lake, ON
L0S 1J0
Project Code: GISC9318-D4b

Mr. Ian D. Smith
Niagara College
135 Taylor Road
Niagara-on-the-Lake, ON
L0S 1J0

Dear Mr. Smith,

RE: SUBMISSION OF GISC9318 DELIVERABLE 4B

Please accept this document as the submission of GISC9318 Deliverable 4b.

The study was to perform two prediction techniques, Inversed Distance Weighted and Kriging to understand the geospatial relationship between the frequency of the data and their geographical locations. This study was done based on the retrieved data of delayed minutes at each Toronto Transit Commission subway stations from Open Data City of Toronto. Analysis was performed not just on the delay minutes, but also the location distributions of the subway stations. Data summary was done using R. Data analysis and predictions were performed using ArcGIS Geospatial Analysis extension.

Should you have any questions or concerns regarding on the project proposal please do not hesitate to contact me at 905-641-2252 or nc.kwong12@gmail.com. Thank you for your time and consideration. We look forward for your feedbacks or discussions regarding on the project proposal.

Regards,

Karen Wong, B.Sc.
KW/
Enclosure(s): WONGKGISC9318D4b

Table of Contents

List of Figures/Table

Karen Wong
GISC9318D4b
April 16, 2018

1.0 Introduction

Studying geospatial data using various data analytical methods can help us to interpret the data and to make prediction based on the data. The purpose of the deliverable is to use a set of data that are geographically corrdinated to perform prediction using two prediction techniques: Inversed Distance Weighted and Kriging. The two techniques perform predictions based on the distance between each points and also the frequeny (z value) of the data. In order to achieve the deliverable objectives, a set of data with x, y and z values are chosen. This study also provides a good idea of the correlation of the data and their locations. Studying the data with the statistical methods such as histogram and boxplot would be done to understand the data. Afterwards a prediction will be made with the dataset. The study topic is the total minutes delay at each subway station of the Toronto Transit Commission.

2.0 Study Area

As of February 2018, the Toronto Transit Commission (TTC) subway connects services for majority of the Greater Toronto Area (GTA) including Vaughan, Scarborough, Etobicoke and the city of Toronto. With high housing prices in the city of Toronto, many people chose to live at the surrounding cities such as Vaughan while commuting to the centre area of GTA for work though public transit. While public transit serves to increase efficiency by avoiding traffic congestions, delays especially during rush hours constantly occurred between TTC subway stations. This consistency of subway delays defeats its original purpose, also brings inconvienency on passengers. The study period is from September 2017 to January 2018.

There are 75 subway stations in total with four subway lines. General subway stations map is displayed in Figure 1. The five stations located at the top left corner of the map are new Toronto-York Spadina Subway Extension which were opened in December 15, 2017. According to the TTC Opeating Statistics in 2016, the busiest subway stations are Bloor-Yonge, St. George, Union, Finch, Dundas, Sheppar-Yonge, Eglinton and Kennedy stations (Toronto Transit Commission, 2017).

Toronto Transit Commission Subway Layout

Figure 1 TTC Subway Station Map

3.0 Data Collection

TTC subway station locations and the total minutes delayed at the reported stations would be the primary source of data in this study. All the data were retrived from the City of Open Data Catelogue website. The coordinates of the subway locations are available in text file and ESRI shapefile. TTC published a detailed record every month which included day, time, station name, delay code and minutes delayed to subway service etc. In order to get sufficient amount of data to make prediction, five months of data, September 2017 to January 2018, were retrieved from the Open Data City of Toronto website. The data and files are very organized and detailed therefore they can be easily interpreted.

There are subway stations that did not have any delays reported during the five-month-period. While longitude and latitiude has a lot smaller range on their data (standard deviation), the delayed minutes data has a wide range between the minimum and maximum values.

Statistical summary of each column is summarized in Table 1.

Statistic Summary	Longitude	Latitude	Delay in Minutes
Minimum	-79.54	43.64	0
Maximum	-79.25	43.79	928
Median	-79.39	43.68	150
Mean	-79.40	43.66	203.7
Standard Deviation	0.0688	0.0469	169.43

Table 1 General data statistic results

3.1 Longitude and Latitude (x and y values)

Some stations such as Kennedy station are the terminal station and also the interchange station with line 2 and the Scroborugh RT. For consistency purposes the delay minutes are combined and grouped by station name resulting longer delays than other stations. (Kennedy station had 928 minutes delay in total which is the longest delay in this dataset.

The x and y values of the data set will be the longitudes (Figure 2) and latitudes (Figure 3) of each subway station.

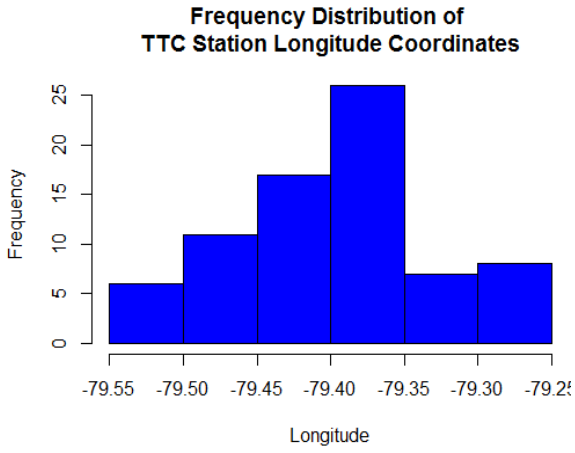


Figure 2 Frequency Distribution of Station Longitudes

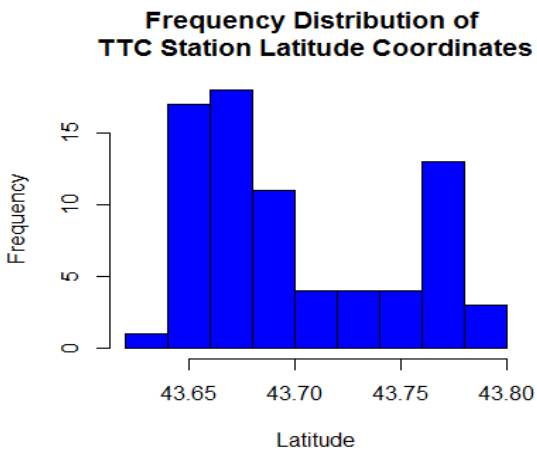


Figure 3 Frequency Distribution of Station Latitudes

The easternmost subway station is Midland Station at city of Scarborough and the most westernmost subway station is Kipling Station at Etobicoke. With the difference between the two stations, majority of the stations are located closed to the centre of Toronto.

The southernmost subway station is Union Station at downtown Toronto and the westernmost subway station is Vaughan Metropolitan Centre Station at Vaughan. With the difference between the two stations, more stations are located at the east side of GTA.

3.2 Delays in Minutes on TTC subway stations

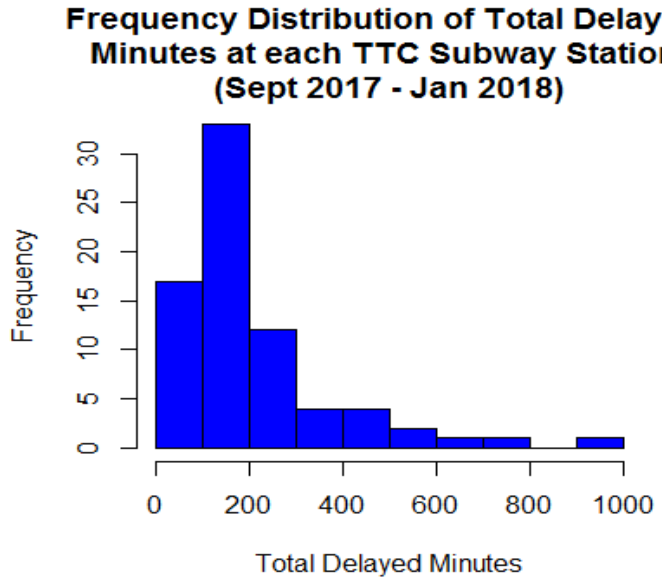
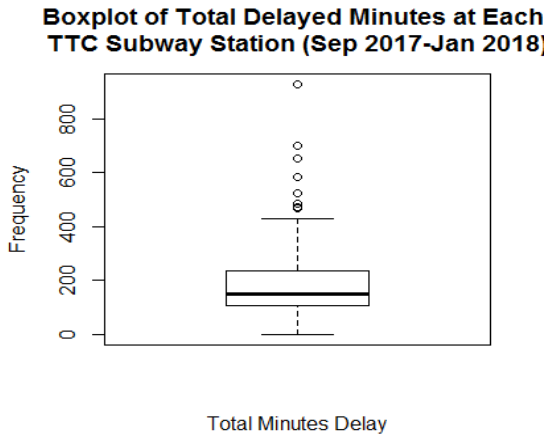


Figure 4 Frequency Distribution of Delayed Minutes

Majority of total delays in five months for each subway station were between 100-200 minutes which is matched with the median of the data. While more stations had 0-50 minutes delay during the study period, few stations have more delays than the other stations such as 600-900 minutes total in five months. The histogram (Figure 4) demonstrated that the graph is left skwered which is not evenly distributed. When a histogram does not have a normal distirbution there are potentials that outliers would appear. The outliers are present in this dataset (Figure 5) showing from the boxplot. The outliers are located at the top of the chart (the circles). The Q-Q plot (Figure 6) illustrated that majority of the data fall within 68% of the normal distribution.



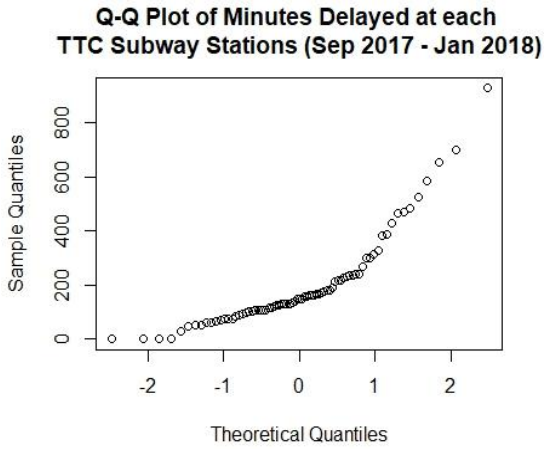


Figure 5 Boxplot of Minutes Delayed in each subway station (5 months)

Figure 6 Q-Q plot of Minutes Delayed at each subway station (5 months)

4.0 Methodology

Two different prediction methods were perform with the collected data: Inversed Distance Weighted and Kriging techniques using the ArcGIS Geospatial Analysis tools – Geospatial Wizard. The Q-Q plot demonstrated the data has a normal distribution because the data are located closed to the trend line. Hence no transformation was performed when data were processed in Geospatial Wizard (Figure 7).

4.1 Inversed Distance Weighted (IDW) Technique

The IDW technique uses a weighted system with exponential factor to make predictions. It provides a quick glance on prediction without prior information required. The technique weights the distance between points to perform prediction. Weighted value is higher when distance between the two locations is shorter. By applying exponential factor to the coorelation, the weight decreases dramatically with the increasing distance between the two locations. Power value was set as 2 to ensure no nagative values result in the prediction.

4.2 Kriging

Kriging interpolates the geospatial data using various statistical models such as probability and prediction etc. Unlike IDW, probability is associated with the prediction Kriging not only makes trends, it also considers measurement errors as well. Kriging method relies on notion of autocorrelation. Autocorrectlation is a function of distance. Kriging is more suitable for predicting normalized data.

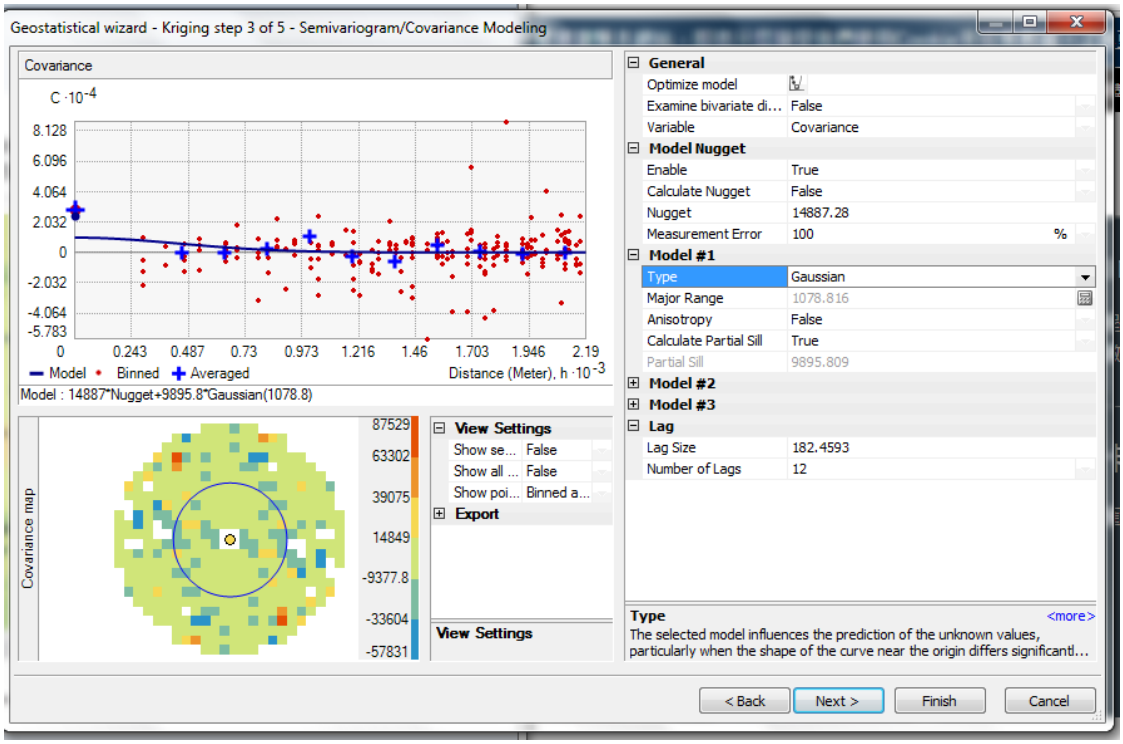


Figure 7 Geostatistical Wizard

5.0 Results

5.1 Total Minutes Delay

Figure 8 illustrated the distribution of total delayed minutes on each TTC subway station. The stations with labels displayed on the map are the ones that have longer delays compare to other stations. These stations are usually the terminal stations or the interchange stations of two subway lines.

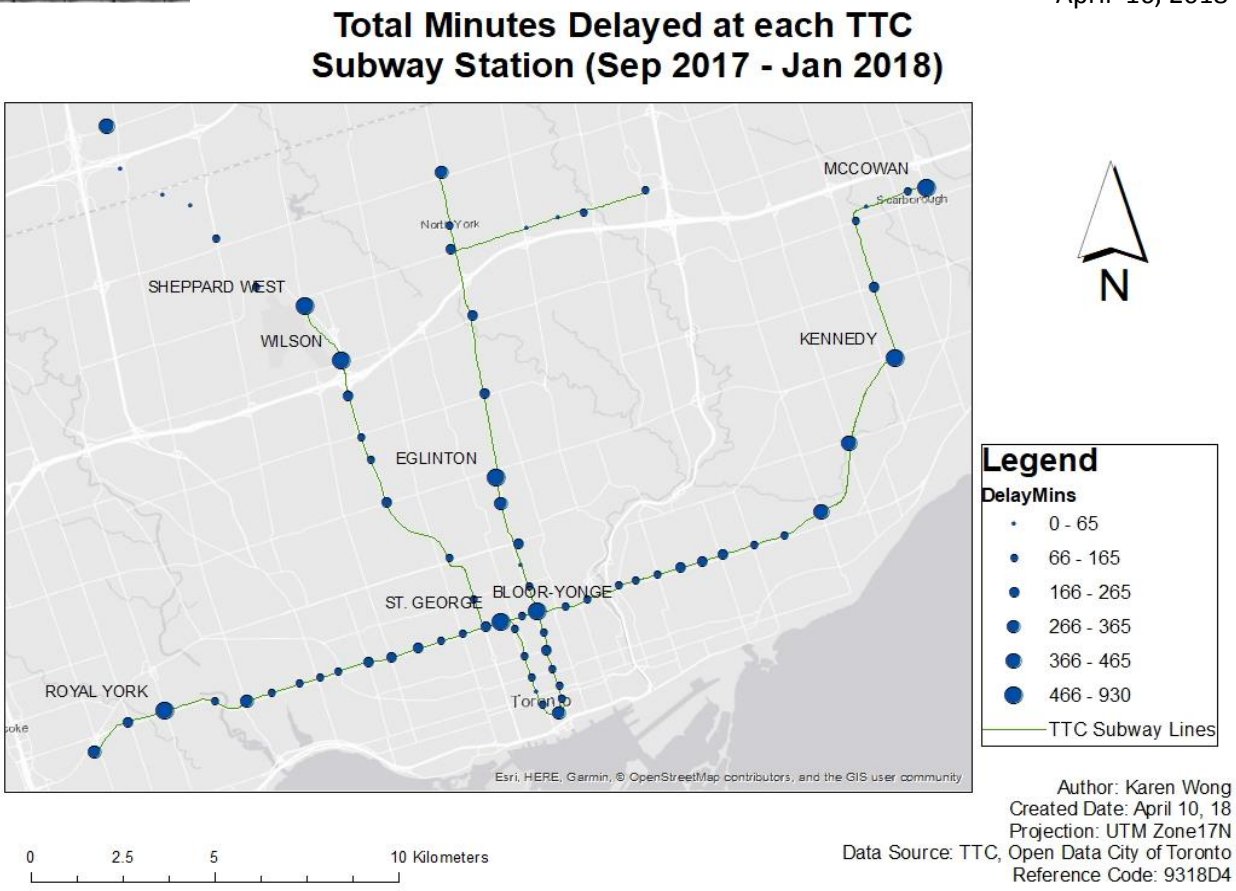


Figure 8 Total Minutes Delayed at each TTC Subway Station

5.2 IDW and Kriging Results

Using the two techniques introduced above, result using IDW (Figure 8) and Kriging (Figure 9) are shown below maps. The range of the result is divided by Natural Break in six classes to obtain an evenly distributed results.

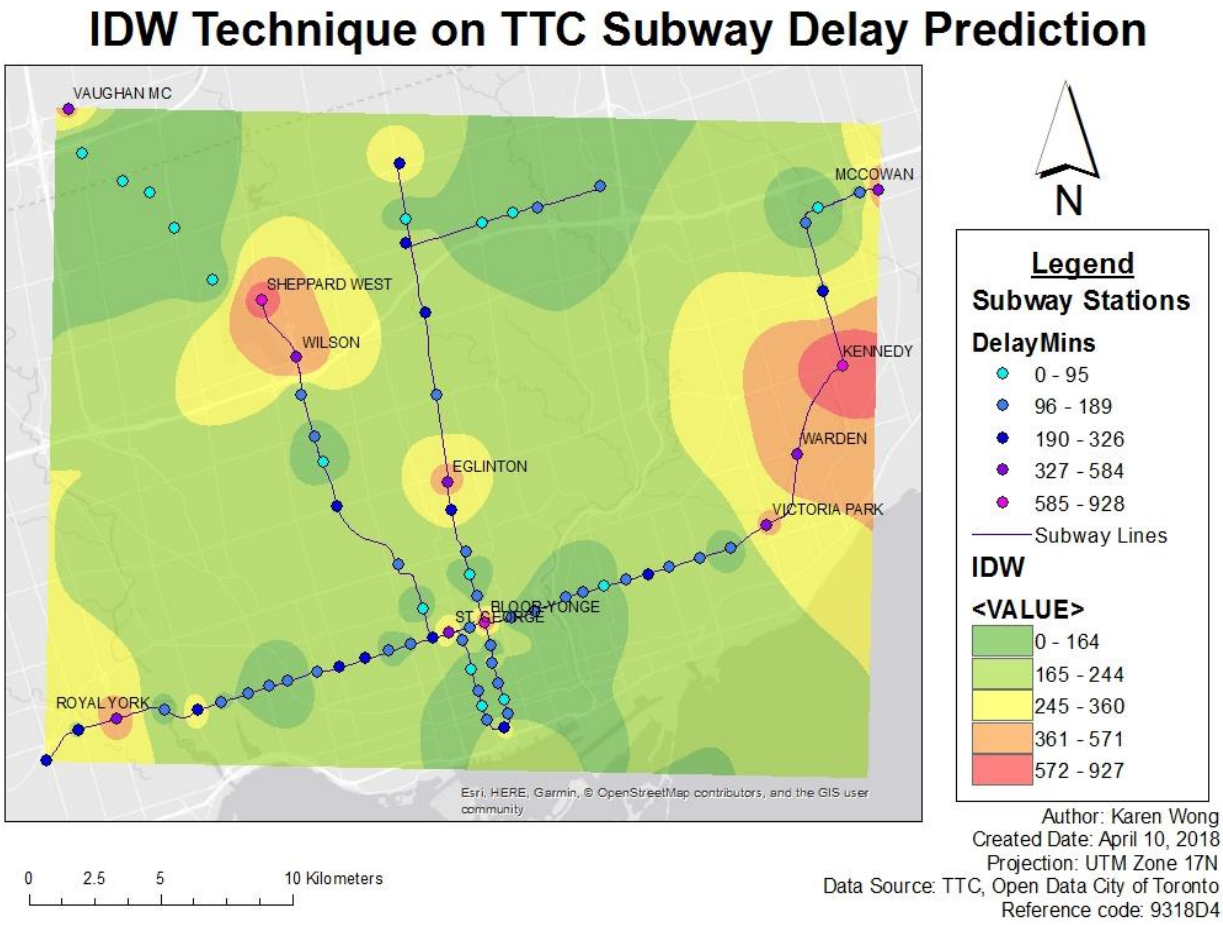


Figure 9 IDW of Total Minutes Delayedat TTC Subway Stations

There is a wide range of area in red colour (predicted delay time from 572 – 927 minutes) near Kennedy station. Areas around Shppard West, Royal York and Eglinton stations were predicted to have moderate delay time (361 – 571 minutes delay). Subway stations located at East side of the subway line near city of Scarborough would have a longer delay than the West side. In general, the predicted delay for most subway stations is less than 200 minutes, including the stations (blue dots) that had longer delays.

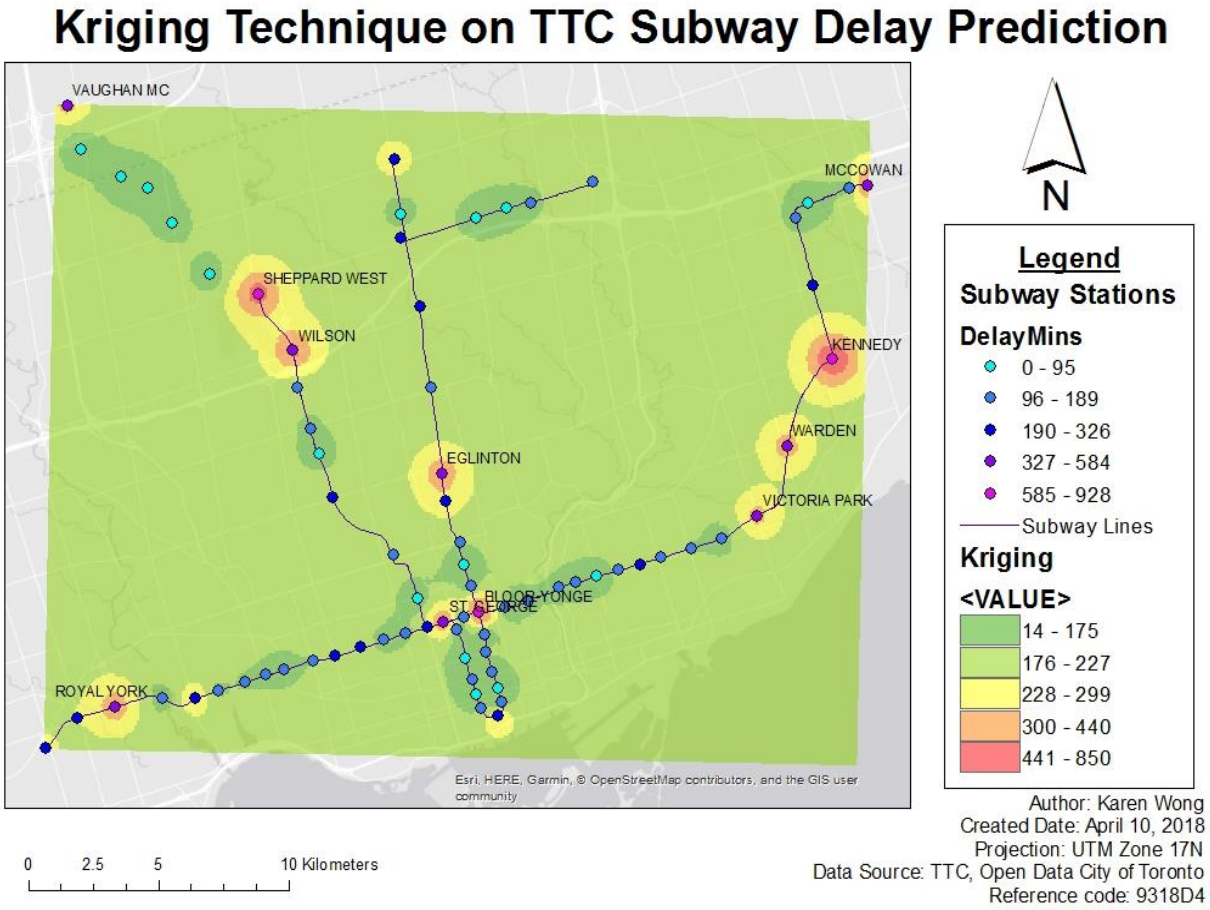


Figure 10 Prediction with Kriging of Total Minutes Delayed at TTC Subway Stations

The range of each station was not as connected to the neightbour stations comparing to the IDW result. Each range especially the stations with longer delay is fairly independent. Stations with moderate delays (range from 96 to 326 minutes delay) are predicted to have shorter delays. The range of values with the Kriging technique is also lower than the original values.

6.0 Interpretation

The prediction with IDW technique demonstrated that stations with longer delays will have larger range affecting neighbouring stations which would have longer delay time as well. While IDW weights z values (delay minutes) heavier when the distance between negihbouring points are shower. In this case subway stations that have longer delays have longer distance between the stations result a wider range. On the other hands for Kriging prediction, the range of the longer delayed stations is more concentrated while the range of shorter delayed stations seems to affect neighbouring stations to have shorter delay in prediction. Even though the buffered ranges vary from the technique results, both maps have similar range distributions on the stations. Combining both of the results it can be concluded that subway stations such as

Kennedy, Royal York, St. George and Bloor-Yonge will continue to have longer delays. Other than that the subway stations are predicted to have shorter delays in general based on the distance and z value anaylsis from IDW and Kriging. Both techniques demonstrated that location and distance of each subway station is slightly correlated to each other however they do not have high correlation.

7.0 Recommendation

The identified “ouliers” from the charts are cruial information that can be interpreted and applied in different ways. For example, the TTC could put more resources on maintenances or prepare more shuttle buses for these stations in case the delays are expected to be longer period of time. On the other hand, companies who consider to place their advertisments with the TTC can purposely place their ads on those “outlier” stations because the longer the delay is the higher chance passengers will look at the subway advertisments.

The total delayed minutes at each subway stations can be interpreted in many difference ways for business decisions depending on the nature of the business. When a set of data does not have a normal distribution, the data provides a lot of information on what is currently happening and changing. In this study the outliers provide essential information of how subway delays occur very often and how long a delay could last.

While IDW provides a wider range of prediction on what area will have more delays, the Kriging result provides a very concentrated range of prediction. Based on the maps comparison above, IDW technique produced a better prediction. Even thought delayed minutes might not have a strong correlation with the station distances, IDW provides a better idea on where the delay distributions will be.

8.0 Closure

IDW technique demonstrated the delay minutes on station affect a wider range on neighbouring stations while Kriging technique provided a more narrow range as result. With the aid of various graphs it has been demonstrated that the total minutes delayed at each TTC stations over four months were not evenly distribution as certain stations have more severe delay issues compare to the other. Those stations that had long delayed times are not the majority of the populations therefore they are the outliers of the dataset. In this scenario studying the length of delay at each TTC stations, the outliers have significant meanings to the data as those stations are the ones that TTC may be suggested to devlop strategy to reduce the length of the delays.

9.0 Bibliography

Toronto Transit Commission. (2017). *Operating Statistics*. Retrieved from Toronto Transit Commission: https://www.ttc.ca/About_the_TTC/Operating_Statistics/2016/section_one.jsp