Python Solutions

M. Tsaqif Wismadi
S2431645

Report: Machine Learning for Traffic Modelling in Eixample District, Barcelona

February 2022

UNIVERSITY OF TWENTE. | ITC

1. Introduction

1.1. Background

Traffic congestion is an emerging global problem that many cities face nowadays. This problem occurs in the developing world as well as in the developed world. In the region of Asia Pacific, traffic congestion has intensified pollution levels and provoked an increase of respiratory-related disease up to 35% (Climate & Clean Air Coalition, 2015). Meanwhile, in Western countries, traffic congestion has been recognized as a daily burden on personal happiness and harms the quality of life (Hays, Olds, & Spence, 2016). Responding to these situations, global cities have come up with various traffic interventions to improve their traffic situations. One of the common types of interventions is traffic modification. This type of intervention attempts to eradicate congestion by changing street directions or street access without adding any new roads (Topirceanu et al., 2016). Even though this intervention often manages to remove traffic congestion from a certain area, sometimes, it also results in road congestion elsewhere (The Jakarta Post, 2019).

Adding to the consequences of traffic modification, the urban street pattern also influences the overall traffic congestion of a city (Wu, Hu, Jiang, & Hao, 2021). For example, it is found that street patterns with multiple junctions can improve traffic flow efficiency (Tsekeris & Geroliminis, 2013). But, tortuous long streets without sufficient junctions are recorded to be more congested than those without this characteristic (Zhao & Hu, 2019). On top of that, a rigid grid pattern critically increases congestion density due to its hindrance on junction's turning and U-turning (Wu et al., 2021). Realizing these correlations, in this study, street network parameters have to be considered as one of the factors that influence traffic congestion and the performance of traffic modification.

Considering all of the existing research gaps, this study aims to assess the performance of traffic modification policy, while considering urban street patterns as a driving factor of traffic congestion by using machine learning that gains its data from sensor-generated input (Figure 2).

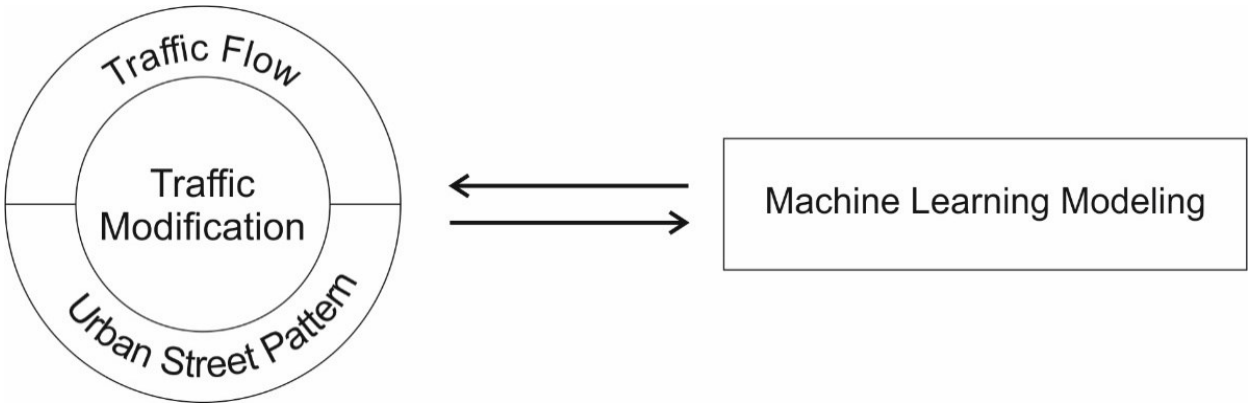

Figure 1. Conceptual Framework: Using Machine Learning to Understand The Interaction between Street Pattern and Traffic Flow

1.2. Study Area

In order to develop the proposed model, a case study needs to be chosen. In a brief, the chosen study area will be a city that meets the criteria of a) has an ongoing traffic modification policy, b) has a distinctive urban street pattern, and c) has adequate traffic sensor infrastructure all over its area. On

that account, Barcelona has an ongoing traffic modification policy called the 'superblock', it has distinguishable street patterns and also has publicly available real-time traffic data. These conditions make Barcelona a suitable city for the study case.

To give a clearer picture, Barcelona's superblock is a policy that aims to minimize the presence of cars by restricting traffic mobility in the city. As can be seen in Figure 3, The superblock is designed to limit motorized vehicle movement in some dedicated areas. Meaning, inside the dedicated areas, existing streets will be converted into urban parks and pedestrian-friendly zones. Hopefully, through this approach, the city can reclaim urban spaces from traffic and become more livable (Ajuntament de Barcelona, 2014).

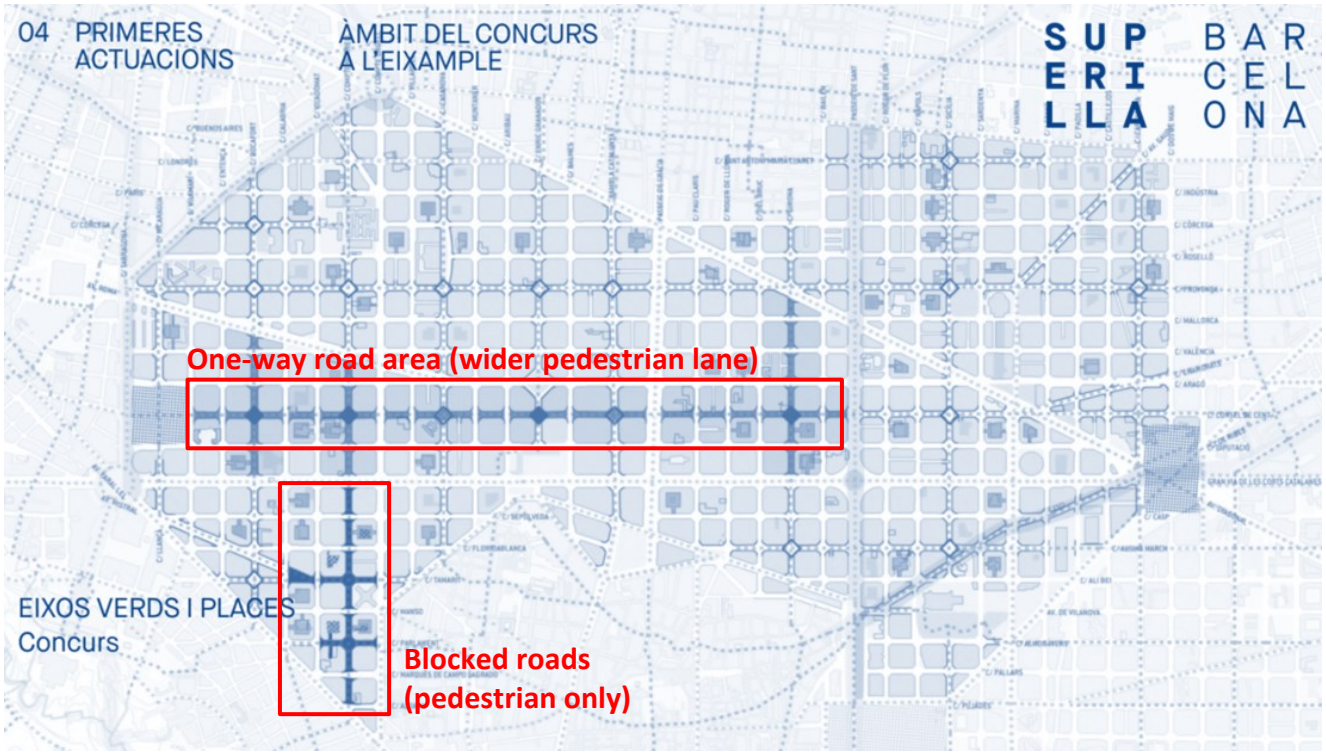

Figure 2. Dedicated Areas for Superblock Policy (Ajuntament de Barcelona, 2014)

Currently, the superblock has only been implemented in The District of Eixample. This is because the district has a distinctive variation of street patterns with a dominant grid that can be easily observed. Correspondingly, this study will not observe the whole city of Barcelona but will only focus on The District of Eixample as its study area.

2. Materials and method

2.1. Description of dataset

To generate a decent machine learning model, a big dataset is required to be fed into the algorithm (Alwosheel, van Cranenburgh, & Chorus, 2018). In this case, Barcelona has an adequate amount of traffic dataset that is generated continuously every 30 minutes. This resulting dataset is produced by sensors installed under the asphalt located near intersections that measure variations in the magnetic field caused by the passage of vehicle metal masses (Open Data Barcelona, 2017). Other than the congestion data, each street sensor is also equipped with street network parameters that represent

the street pattern conditions around it (300-meter buffer from the street sensor). Furthermore, this dataset is also openly available to the public and can be easily downloaded in CSV format through *opendata-ajuntament.barcelona.cat*. To be more detailed, the traffic dataset contains attributes such as the street section code, description of the section (street name), coordinates of the street section, various street network parameters, and the current congestion status (Table 1).

Table 1. Attribute Description of Barcelona Traffic Dataset

Attribute	**Description**
Street_name	Name of the street where the sensor belongs in
Long	Longitude coordinates of the street sensor
Lat	Latitude coordinates of the street sensor
idTram	Street sensor identifier
L_section	The total length of road within the 300-meter buffer of each sensor
N_vertices	Total number of junctions (vertices) within the 300-meter buffer of each sensor
Blocked_r	The total length of blocked road within the 300-meter buffer of each sensor
Half_r	The total length of one-way road within the 300-meter buffer of each sensor
Congestion	Congestion status (0 = no data, 1 = very fluid, 2 = fluid, 3 = dense, 4 = very dense, 5 = congestion, 6 = cut)
NR	The ratio of the number of vertices with connections (nodes) to the total number of vertices within the 300-meter buffer of each sensor
CR	The ratio of the number of vertices without connections (cul de sac) to the total number of vertices within the 300-meter buffer of each sensor
TR	The ratio of the number of T-junctions to the total number of junctions within the 300-meter buffer of each sensor
XR	The ratio of the number of cross junctions to the total number of junctions within the 300-meter buffer of each sensor
Stan_LS	Standardized value of the 'L_section' attribute (0-1 scale)
Stan_NV	Standardized value of the 'N_vertices' attribute (0-1 scale)
Stan_BR	Standardized value of the 'Blocked_r' attribute (0-1 scale)
Stan_HR	Standardized value of the 'Half_r' attribute (0-1 scale)

The indices that are used in this study are adopted from the quantitative classification of street patterns by the measurement of network features (Han, Sun, Yu, Song, & Ding, 2020). Although the 'L_section' and the 'N_vertices' are seen as tentative indices, the other indices are essential for pattern determination and work in duality. It means that the sum of NR and CR is equal to 1, while the sum of TR and XR is equal to 1.

2.2. Methodology

This study will be conducted in four major steps, namely 1) initial mapping and data preparation, 2) exploratory spatial data analysis, 3) selecting a machine learning method for a traffic modelling task, and 4) building a machine learning model (Figure 3).

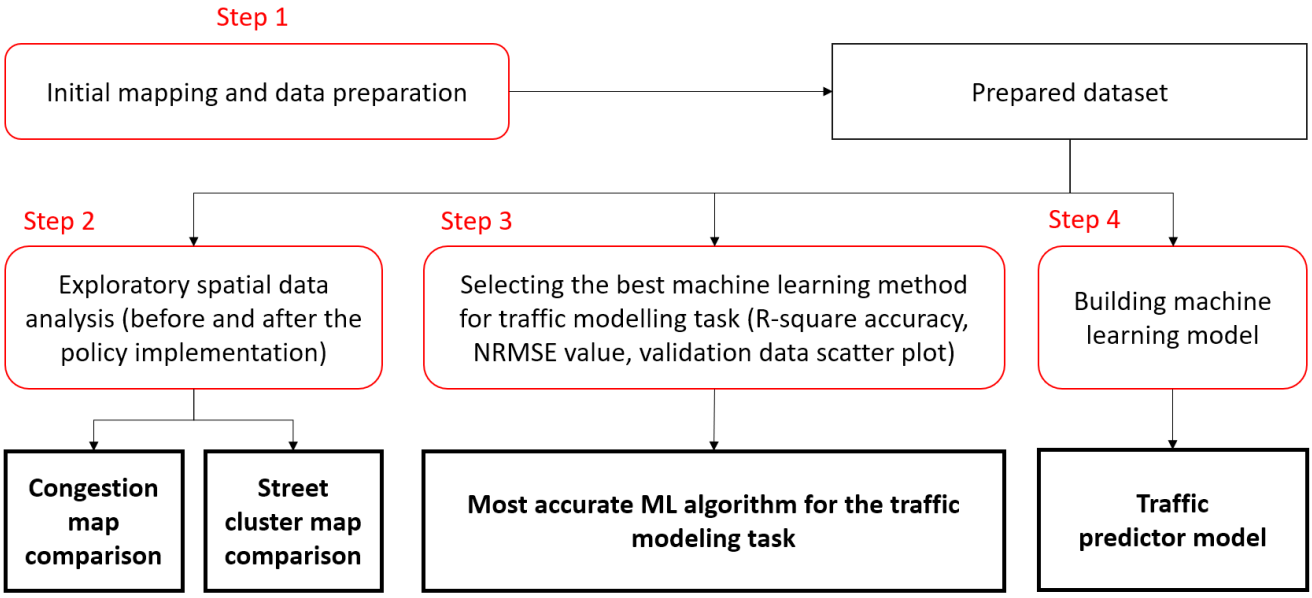

Figure 3. Methodological Framework of The Study

Initial mapping and data preparation

In this stage, the retrieved data will be sorted and spatially plotted. Through this step, the distribution of sensor locations can be observed. This is necessary to understand whether there is enough sensor to covers the study area. Other than understanding the spatial distribution, this step is also necessary to match the coordinate reference of each dataset. Finally, once the data is plotted and coordinately projected, the program can be overlayed neatly in one spatial frame, and thus the prepared dataset is produced.

Exploratory spatial data analysis

For this particular study, there will be only two tasks of ESDA to be conducted, namely congestion mapping and street cluster mapping. The congestion mapping will be done in a straightforward spatial mapping procedure, using the 'congestion' attribute on the dataset, the program will visualized its value using manual classification where 0 = no data, 1 = very fluid, 2 = fluid, 3 = dense, 4 = very dense, 5 = congestion, 6 = cut. Each value will be mapped under the purple to red color scale where the higher the value, the redder and bigger the point will be.

In terms of cluster mapping, the analysis will be conducted with the K-means clustering method. This method will be fed on various indices that the dataset possesses (in this case it will include all street parameters) and groups each datapoint into k-number of classes based on the mean value of the given indices.

Selecting a machine learning method for a traffic modelling task

A machine learning approach is considered as a modern statistical method from a perspective focusing on two interrelated questions: how can one construct computer systems that automatically improve through experience and what are the fundamental statistical computational information laws that govern all learning systems (Jordan & Mitchell, 2015). A learning problem can be defined as the

problem of improving some measure of performance when executing some task, through some type of training experience.

Three quantitative performance measures were calculated to assess the performance of each machine learning method: R-square, normalized- RMSE (NRSME), and the scatter plot of validation data. The R-square accounts for the total variance of the model that can be explained by the predictors, the NRMSE measures the differences between the predicted values by a model or an estimator and the observed values, whilst the scatter plot gives a visual confirmation on how the predicted values from the model compared to the actual value on the field (Lundberg, Johnson, & Stewart, 2021).

The performance evaluation was carried out to three differing algorithms: decision tree (DT), linear regression (LR), and random forest (RF). The best model was selected from the algorithm showing the highest R-square, the lowest NRSME, and the most diagonally-plotted scatter plot. In addition to the model performance, the relative importance of each variable included in the model will also be examined. The relative importance was measured using the Feature Importance (FI) scores (Rogers & Gunn, 2005). The FI scores represent the relative weight of each variable to the model.

Building machine learning model

Once the machine learning method is selected, a machine learning predictor will be built. The predictor will be programmed to ask for street parameters input from the user, after that, it will proceed to predict the traffic congestion level based on the given inputs. The prediction that it gives will be based on the learning pattern that it grasps from its training with the given Barcelona data.

3. Results and discussion

3.1. Initial Mapping

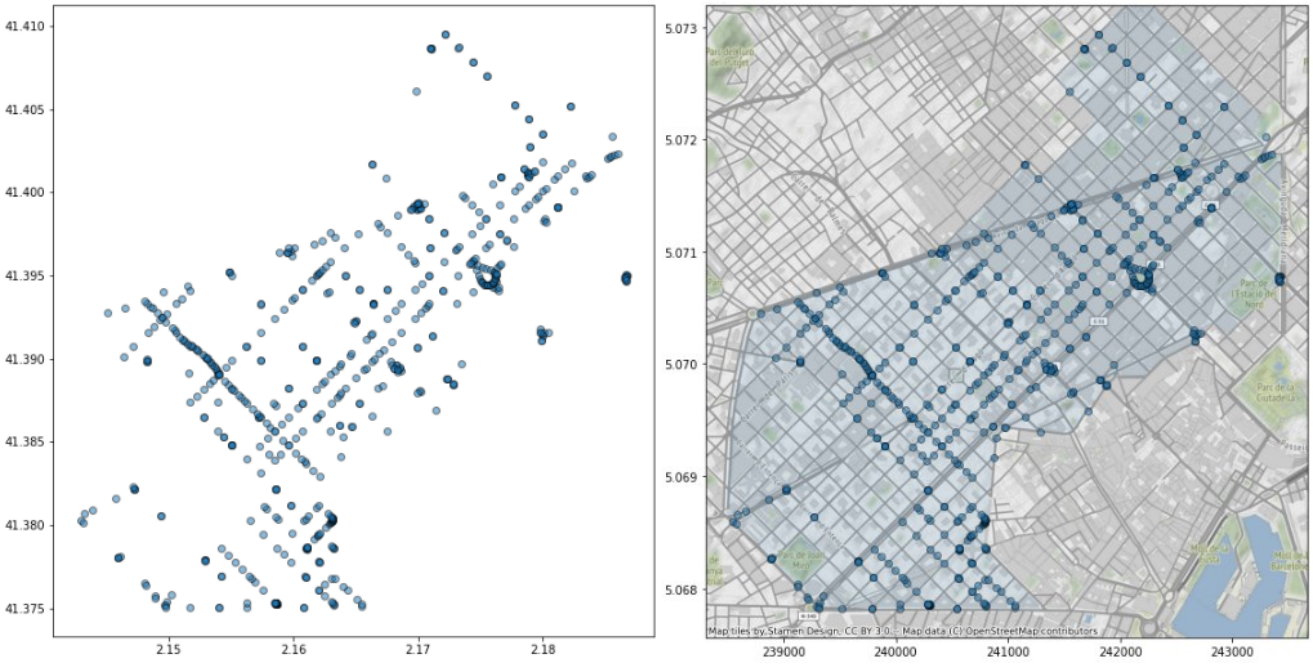

Figure 4. Traffic Sensor Location Distribution on The District of Eixample, Barcelona

From initial mapping, it can be observed that the place of traffic sensors is somewhat adequate to cover the whole study area of the Eixample District (the translucent blue area on the map). Furthermore, in the district center, the sensor placement is very dense that some road sections could even have more than one sensor between junctions. However, it can also be seen that the east side and the north side of the district are quite hollow compared to the rest of the district. Fortunately, this will not affect the study since the superblock intervention areas are not located in that hollow sides (Figure 2).

3.2. Exploratory spatial data analysis

From the comparison of the traffic map, it is quite apparent that the congestion where road-blocking is being implemented (red circle) is decreasing from the average level of 3 (dense traffic) to the average level of 2 (fluid traffic). However, on the new one-way area (red rectangle), the traffic has been significantly increased from the congestion level of 1 (very fluid), to a little bit over level 2 (fluid). Assumably, from this phenomenon, we can interpret that in terms of congestion, dedicating a fully pedestrian area is better than making a one-way road with a wider pedestrian lane (Figure 5). Most probably, by widening the pedestrian lane, there is a decrease in capacity but an increase in traffic volume, and hence it creates a slower traffic flow.

Additionally, an interesting finding emerged when k-means clustering was performed (Figure 6). When all explanatory variables are considered, the superblock policy has changed the street network cluster. Assumably, this is because all selected explanatory variables in this project are network parameters which to some extent will capture the intervention of superblock. Furthermore, the k-means clustering method also captures the difference between one-way road changes and blocked road areas. The one-way area (red rectangle) is clustered as a group post superblock policy (group 2), whilst the blocked road area is also captured as another group (group 1). The result of cluster map comparison shows that superblock does significantly change the street pattern dynamic in Eixample. Referring to the traffic modification theory, the changes of the street pattern will also change the congestion level dynamic (Wu et al., 2021), and thus, the condition is appropriate to be modeled in a machine learning task.

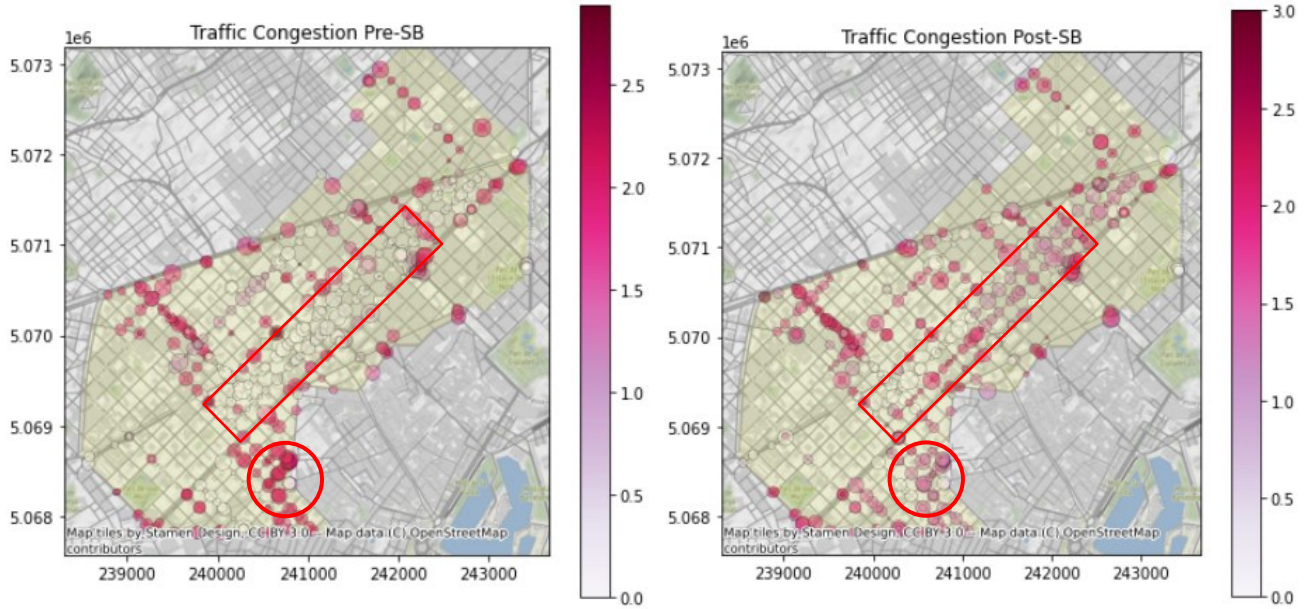

Figure 5. Congestion Map Comparison

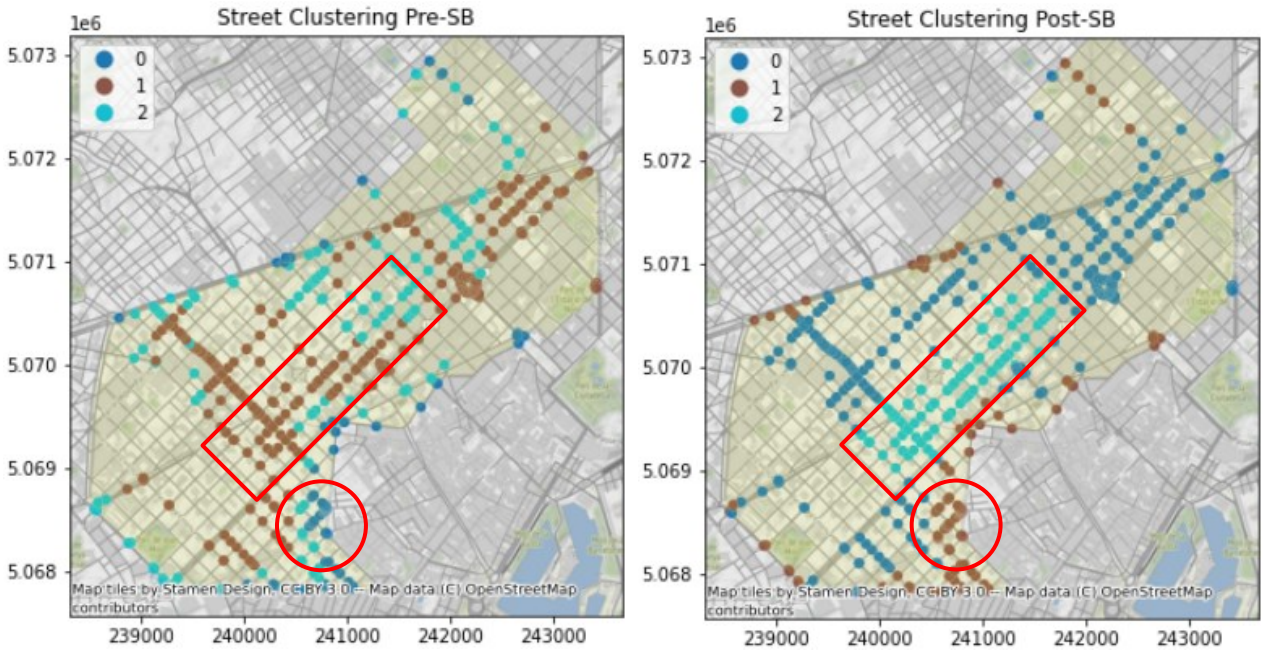

Figure 6. Street Cluster Map Comparison

3.3. Machine learning performance

The performance of three machine learning algorithms was summarized in Table 2 and Figure 7. Here, it can be observed that street network parameters do has some correlation with traffic congestion. It can also be observed, that random forest outperformed other algorithms. In most cases, machine learning can be considered powerful with the ability to produce relatively high performance (R-square more than 0.8) (Raghuram, Akshay, & Chandrasekaran, 2016). However, the result in this study shows a much lower rate with a range of R-square 0.08 to 0.56.

Additionally, modeling performance is dependent on the nature of the model and the data used in the model. In this study, we exploited both numerical and categorical variables in the modeling. The fact that random forest outperformed two other algorithms, can possibly be attributed to several reasons. First, random forest (RF) is an algorithm designed to handle both numerical and categorical data, and accordingly, more variables can be inputted for the modeling, and consequently can explain variations of traffic congestion better. Second, as opposed to the decision tree (DT), which is also able to handle both numerical and categorical data, the RF model structure is more complex and therefore, can potentially represent the dynamics of traffic congestion. Lastly, merely considering street parameters to predict traffic congestion is inadequate. Numerous studies show that traffic congestion levels can be complex as it can be modulated by varying factors such as land use, demographic density, and bottleneck effect (Sarzynski, Wolman, Galster, & Hanson, 2016).

Table 2. Summary of model performance from three algorithm methods

Measurement	Linear Regression (LR)	Decision Tree (DT)	Random Forest (RF)
R-square accuracy	8%	45%	56%
NRMSE value	0.25	0.2	0.18

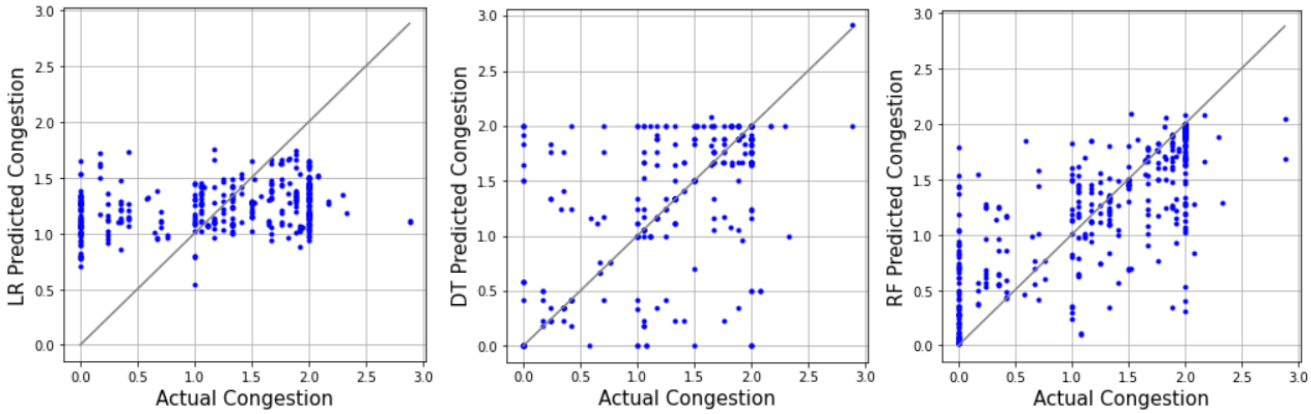

Figure 7. The Scatter Plot of Validation for Each Algorithm Method

4. Conclusion

In summary, this study manages to assess the performance of traffic modification policy of superblock, while considering urban street patterns as a driving factor of congestion. It shows that there is a correlation between street patterns and traffic congestion, and therefore using network parameters to model congestion is a valid approach. However, for better accuracy, it should also be combined with other non-network variables.

Furthermore, the study also deduced that even though creating a pedestrian-only area can reduce traffic congestion, widening a pedestrian lane and converting a two-way road into a one-way lane might increase traffic congestion. In addition, the study has also shown that when it comes to traffic modeling, random forest consistently outperforms the other methods. Assumably, this is caused by its ability to handle a more complex task that includes both numerical and categorical data.

On a bigger picture, the project of modeling traffic congestion with machine learning has improved our understanding regarding how essential it is to design and manage the configuration of the urban street. From what we have observed, it can significantly affect the level of traffic congestion.

Reference

Ajuntament de Barcelona (2014). Urban Mobility Plan of Barcelona. Barcelona.

Alwosheel, Ahmad, van Cranenburgh, Sander, & Chorus, Caspar G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*, 167–182.

Climate & Clean Air Coalition (2015). Air pollution measures for Asia and the Pacific,. Retrieved from https://www.ccacoalition.org/en/content/air-pollution-measures-asia-and-pacific.

Han, Baorui, Sun, Dazhi, Yu, Xiaomei, Song, Wanlu, & Ding, Lisha (2020). Classification of urban street networks based on tree-like network features. *Sustainability (Switzerland)*, *12*(2).

Hays, Tanna, Olds, Preston, & Spence, John (2016). Happiness and Traffic: An Analysis of Long Term Effects. Denver.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Lundberg, Ian, Johnson, Rebecca, & Stewart, Brandon M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory: *https://doi.org/10.1177/00031224211004187*, *86*(3), 532–565.

Open Data Barcelona (2017). Traffic State Information by Sections,. Retrieved from https://opendata-ajuntament.barcelona.cat/data/en/dataset/trams.

Raghuram, M. A., Akshay, K., & Chandrasekaran, K. (2016). Efficient User Profiling in Twitter Social Network Using Traditional Classifiers. *Advances in Intelligent Systems and Computing*, *385*, 399–411.

Rogers, Jeremy, & Gunn, Steve (2005). Identifying Feature Relevance Using a Random Forest. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3940 LNCS*, 173–184.

Sarzynski, Andrea, Wolman, Harold L., Galster, George, & Hanson, Royce (2016). Testing the Conventional Wisdom about Land Use and Traffic Congestion: The More We Sprawl, the Less We Move?: *http://dx.doi.org/10.1080/00420980500452441*, *43*(3), 601–626.

The Jakarta Post (2019). Odd-even traffic policy trial is a success,. Retrieved from https://www.thejakartapost.com/news/2016/08/26/odd-even-traffic-policy-trial-is-a-success-police.html.

Topirceanu, Alexandru, Iovanovici, Alexandru, Cosariu, Cristian, Udrescu, Mihai, Prodan, Lucian, & Vladutiu, Mircea (2016). Social cities: Redistribution of traffic flow in cities using a social network approach. *Advances in Intelligent Systems and Computing*, *356*, 39–49.

Tsekeris, Theodore, & Geroliminis, Nikolas (2013). City size, network structure and traffic congestion. *Journal of Urban Economics*, *76*(1), 1–14.

Wu, Chao Yun, Hu, Mao Bin, Jiang, Rui, & Hao, Qing Yi (2021). Effects of road network structure on the performance of urban traffic systems. *Physica A: Statistical Mechanics and its Applications*, *563*, 125361.

Zhao, Pengjun, & Hu, Haoyu (2019). Geographical patterns of traffic congestion in growing megacities: Big data analytics from Beijing. *Cities*, *92*, 164–174.