# Exploring spatio-temporal patterns and potential factors of traffic congestion: a case study of New York City


## ABSTRACT

Traffic congestion has become a pressing urban issue alongside rapid urbanization, impacting various aspects of citizens' lives, from daily commutes and mental well-being to broader goals of urban sustainability. Understanding the underlying mechanisms of traffic congestion can inform effective solutions and policies. To reach this goal, a critical step is systematically identifying the geospatial factors that shape traffic patterns. However, existing studies focus primarily on the spatial and temporal characteristic analysis of traffic congestion with a lack of exploration of potential factors driving these spatio-temporal patterns. This study addresses this gap by analyzing a dataset of approximately 26.07 million average travel speed records for 100,206 road segments in New York City over one month. Firstly, based on the travel time index (TTI) and hierarchical clustering method, typical congestion patterns across city and borough levels are identified. Then, least squares regression analysis is applied to explore the significant factors related to traffic congestion. Finally, five non-linear function models are used to better characterize relationships between TTI and each significant factor, with Pearson correlation applied to address multicollinearity. The findings show four temporal and two spatial congestion patterns in New York City , with 12 out of 20 selected factors—categorized as Diversity, Density, Design, and Distance to transit—emerging as key determinants of congestion patterns. Building on these, the study proposes strategies to alleviate traffic congestion, providing actionable insights for optimizing traffic management and establishing a foundation for congestion prediction.



## CCS CONCEPTS

• Information systems → Geographic information systems;

• Data mining → Clustering.

## KEYWORDS

Traffic congestion, Spatio-temporal pattern, Travel time index, Least squares regression


## 1 Introduction

The rapid urbanization has boosted socio-economic development but also resulting in various urban challenges. Among them, traffic congestion stands out as a critical issue, affecting citizens' daily activities and mental health [1] [2]. To understand its underlying mechanism, it is key to explore spatio-temporal variations of traffic congestion and investigate the relationship between traffic congestion patterns and the associated potential factors. While previous studies have addressed such topic, they either mainly focus on city level with a lack of borough level analysis or conduct a study without a comprehensive analysis on what and how geospatial factors potentially shape traffic congestion.

To fill the gap, this study systematically explores spatio-temporal patterns of traffic congestion and the associated factors from the lens of transport geography. Taking New York City as a case study, an hourly average speed data set from Dec 1, 2018 to Dec 31, 2018 is analyzed for uncovering spatio-temporal characteristics, and an associated social sensing dataset is studied for factors exploration.

This study resulted in the following three contributions:

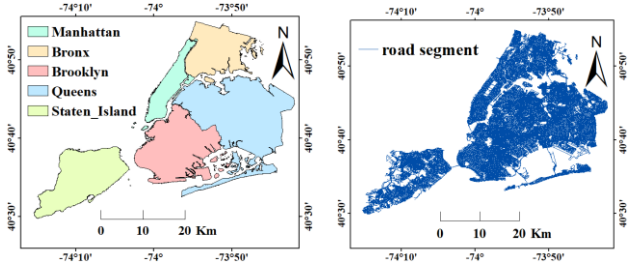
- Four temporal patterns and two spatial patterns at city and borough levels are uncovered by employing a hierarchical clustering method
- A multiple linear model based on least square method is established, linking the traffic congestion index with 16 significant influencing factors
- Non-linear fitting combined with Pearson correlation analysis is employed to elucidate the true relationships between congestion and these factors, providing a deeper insight into the key determinants of traffic congestion

The paper is structured as follows: **Section 2** describes the study area and dataset utilized in this study. **Section 3** outlines the methodology. **Section 4** presents the clustering results and identifies potential factors related to traffic congestion patterns. **Section 5** discusses the findings and concludes the study.

# 2 Materials

## 2.1 Study area

The New York City (NYC) is selected as the case study owing to its intricate and diverse road network, coupled with extensive access to traffic data, enabling a detailed analysis of congestion patterns. The city exhibits multicentricity characteristics, including five boroughs: Manhattan, Bronx, Brooklyn, Queens, and Staten Island (**Fig.1a**). A total of 110,704 road segments are selected and the road network system is comprised of expressways, primary roads, secondary roads, tertiary roads and branches (residential roads, service roads, living streets, tracks, etc.) (**Fig.1b**).




(a) Boroughs of NYC (b) Road Network (2018)

**Figure 1: Study area**

## 2.2 Dataset

*2.2.1 Average travel speed data.* In this study, 25,068,883 records of average travel speed data were obtained from the Uber Movement plaform covering 100,206 road segments in the period from Dec 1, 2018 to Dec 31, 2018 with an one-hour temporal resolution. Each record of the travel speed data contains three fields: recording time, road segment ID and average speed. Finally, a total of 2,221,916 records of Mondays are extracted, 9,680,290 records of the other four weekdays are extracted, 5,396,114 records of weekends are extracted and 7,770,563 of holidays are extracted.

*2.2.2 Social sensing data*. The five Ds (density, diversity, design, destination accessibility, and distance to the transit) are considered to shape urban forms and influence human behaviors [5]. Traffic congestion, in turn, can be regarded as a consequence of human travel patterns heavily reliant on automobiles, alongside complex urban structures [4]. In this study, we categorize 20 factors that may influence traffic congestion into four classes based on the Ds model, excluding destination accessibility due to data limitations. These four categories of factors are then used as explanatory variables to examine the relationship between urban form factors and traffic congestion. **(Table 1)**.

**Table 1: Four "Ds" built-environment variables**

| Four Ds | Independ variables | Symbol |
|---|---|---|
| Diversity | Number of social service facilities | $X_1$ |
| | Number of commercial facilities | $X_2$ |
| | Number of residential facilities | $X_3$ |
| | Number of recreational facilities | $X_4$ |
| | Number of transportation facilities | $X_5$ |
| | Number of educational facilities | $X_6$ |
| | Number of cultural facilities | $X_7$ |
| | Number of healthcare facilities | $X_8$ |
| | Number of bus stops | $X_9$ |
| Density | Population density | $X_{10}$ |
| | Building height | $X_{11}$ |
| | Area of parking lots | $X_{12}$ |
| Design | Road length | $X_{13}$ |
| | Road width | $X_{14}$ |
| | Number of street lights | $X_{15}$ |
| | Distance to the nearest bridge | $X_{16}$ |
| | Length of bike routes | $X_{17}$ |
| Distance to transit | Distance to the nearest bus stop | $X_{18}$ |
| | Distance to the nearest subway entrance | $X_{19}$ |
| | Distance to the nearest domestic airport | $X_{20}$ |

# 3 Method

## 3.1 TTI-based hierarchical clustering

To systematically understand the spatio-temporal patterns in the NYC, hierarchical clustering was performed on city and borough levels by using the constructed average hourly TTI matrix. The selection of features is a crucial step of clustering, as it is vital for measuring traffic congestion performance and identifying clustering patterns. In this study, travel time index is adopted to characterize traffic congestion, constructing clustering features[6]. The calculation formula cam be expressed as formula (1):

$$TTI = \frac{\bar{T}}{T_{free}} = \frac{\frac{L}{\bar{V}}}{\frac{L}{V_{free}}} = \frac{V_{free}}{\bar{V}} \quad (1).$$

$\bar{T}$ indicates the average travel time. $T_{free}$ indicates the free-flow travel time. L indicates the length of road segment. $\bar{V}$ indicates the average travel speed. $V_{free}$ indicates the free speed. Traffic congestion performance can be divided into five levels according to the official classification of TTI in AutoNavi map[3], as shown in **Table 1** in **Appendix**.

Assuming that there are N road segments in the road network, the TTI vector of road segment $n$ on one day can be expressed as $TTI_n = [tti_{n,1}, tti_{n,2}, \dots, tti_{n,T}] \in R^T$, where $tti_{n,t}$ denotes the average TTI value of a single day at the $t_{th}$ hour. In this study, T is set to 24 as the time interval is 1 hour.

Therefore, the TTI matrix for N road segments on one day can be acquired as $TTI = [TTI_1, TTI_2, \dots, TTI_N] \in R^{N \times T}$. Through a visualization for the TTI of every day, one can find that there is significant difference among Mondays, the other four normal weekdays, weekends and holidays. Therefore, four matrices are calculated respectively. ($TTI_{Mondays,}$ $TTI_{Normal\ weekdays}$, $TTI_{Weekends}$, $TTI_{Holidays}$)

Due to the effectiveness in terms of clustering and the ease of results visualization, There are two main approaches of hierarchical clustering[7]: agglomerative clustering and divisive clustering, depending on whether the dendrogram formation using a "bottom-up" agglomerative strategy or a "top-down" divisive strategy.The agglomerative approach[8] is applied to the TTI matrix to obtain clusters.

### 3.2 Exploration of potential factors

A regression model based on least squares regression [9] is initially applied to fit the relationship between TTI and the explanatory variables. In the regression analysis, the TTI is the average value of each road segment over the study period, representing the typical congestion level. Similarly, the potential factors are road segment-level attributes capturing various spatial and built-environment variables around each road. While the regression model can reflect the relationship between TTI and multiple factors, multicollinearity may exist among the explanatory variables. Hence, the Pearson correlation coefficient [10] is employed to a better discovery for the relationship between TTI and explanatory variables.

## 4 Results

### 4.1 Spatio-temporal patterns

The clustering dendrograms are shown in **Fig.2-Fig.6** in **Appendix**. The variation curve of the travel time index with time was shown in **Fig.2-Fig.7**. The horizontal axis presents the hours of a day, while the vertical axis presents the average value of TTI. It can be found that the spatio-temporal patterns of the five boroughs are consistent with that of the city, with four temporal patterns (i.e., Mondays pattern, normal weekdays pattern, weekends pattern and holidays pattern) and two spatial patterns. According to the classification of the TTI, the spatial pattern of cluster 1 can be defined as smooth with mild congestion at peak hours, while the spatial pattern of cluster 2 can be defined as congestion pattern.

The clustering results show that there are two evident peaks in the morning and evening on weekdays, while only evening peak exists on weekends and holidays. From the TTI value, we can also find that the traffic pressure on weekdays is stronger than that of non-weekdays. Although Mondays and normal weekdays exhibit morning and evening peak characteristics at the same time, there is a difference in the traffic congestion intensity, that is, the morning peak on Mondays is usually more congested than that of normal weekdays, while the evening peak is the opposite. This phenomenon indicates that commuting is indeed a main causation for traffic congestion in the NYC on weekdays. Similarly, there is also difference in congestion intensity between weekends and holidays. An interesting phenomenon is that the congestion intensity on weekends is usually weaker than that of holidays except for Manhattan. Possible reasons for this may include the decrease in commuting, tourism, local residents' activities and commercial activities on weekends.

To further gain insight into the traffic congestion performance of the five boroughs, the proportion of the TTI at different levels are summarized as shown in **Fig.8**. It can be seen that Manhattan and Brooklyn have the highest traffic pressure, with congestion state accounting for more than 50% of the total time. Especially, the congestion period of Manhattan on weekdays even reaches to 70%. Queens and Staten Island are in a state of free, smooth, and mild congestion for more than 70% and 80% of the time, respectively. These two boroughs have the lowest traffic pressure compared to the other three ones. Besides, the heavy congestion state on weekdays is about 1-2% higher than on non-weekdays, implying that the traffic pressure on weekdays is stronger than non-weekdays.

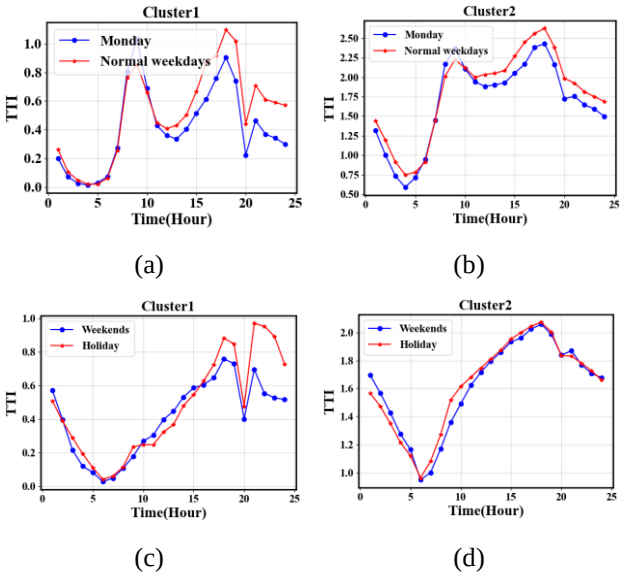
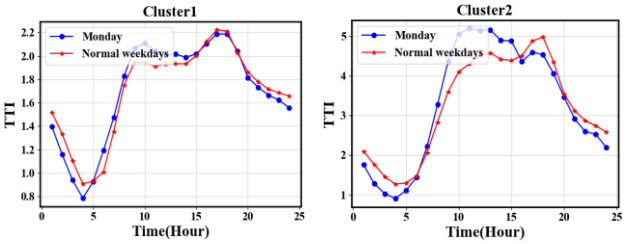
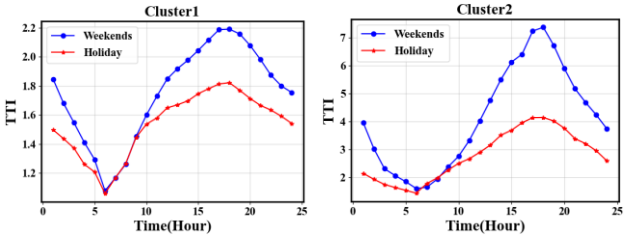



**Figure 2:** Spatio-temporal patterns of roads in NYC

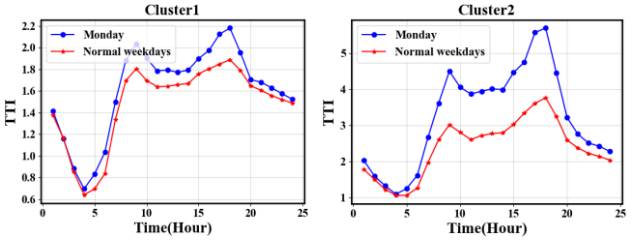



(a) (b)

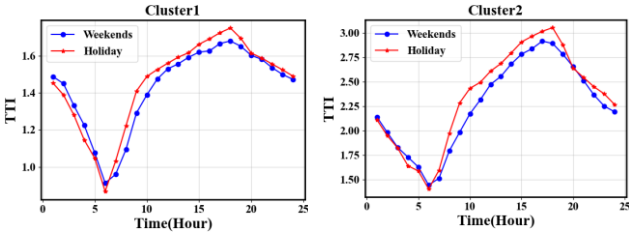



(c) (d)

**Figure 3:** Spatio-temporal patterns of roads in Manhattan

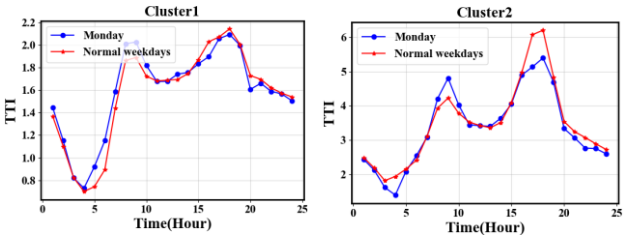



(a) (b)

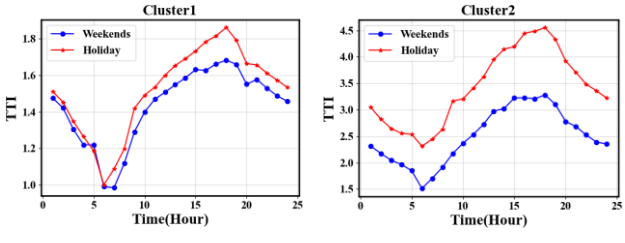



(c) (d)

**Figure 4:** Spatio-temporal patterns of roads in Brooklyn

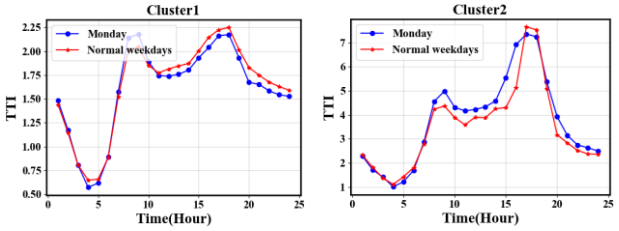



(a) (b)

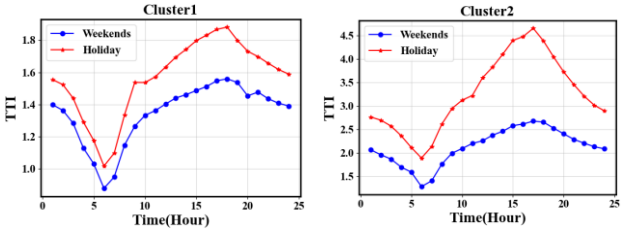



(c) (d)

**Figure 5:** Spatio-temporal patterns of roads in Queens

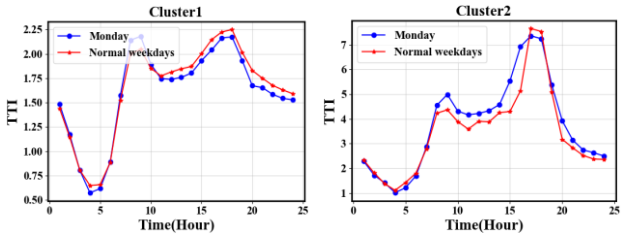



(a) (b)

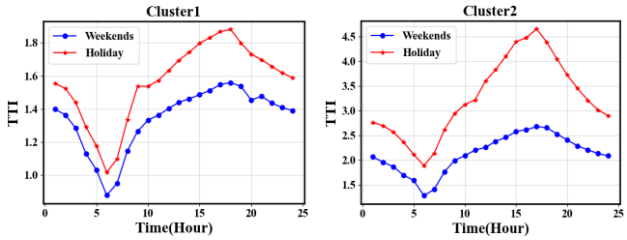



(c) (d)

**Figure 6:** Spatio-temporal patterns of roads in Bronx




(a) (b)




(c) (d)

**Figure 7:** Spatio-temporal patterns in Staten_Island

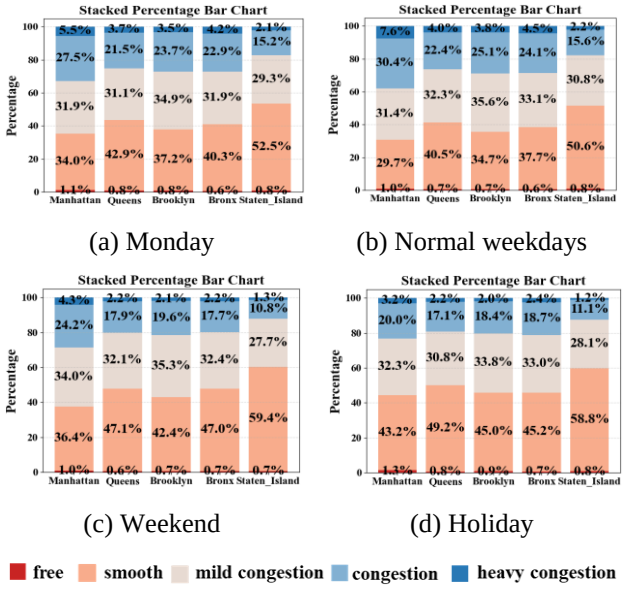



**Figure 8:** Proportion of TTI for the five boroughs

## 4.2 Exploration of potential factors

T-test and p-value are utilized to test the significance of regression coefficients. If the p-value is less than 0.05, it indicates that the corresponding regression coefficient is significant. The univariate linear regression is firstly performed between TTI and each factor, and the results (**Table 2** in **Appendix)** indicate that all 20 factors are highly significant with TTI. Then multiple linear regression based on least square method is used to explore the combined effects of multiple factors (**Table 3** in the **Appendix**), while the p-values for the four factors (i.e., number of residential facilities, transportation facilities and bus stops, area of parking lots) are bigger than 0.05. Therefore, the four insignificant factors are deleted and only the 16 significant factors are reserved in the final multiple linear regression model (**Table 2)**. It can be found that all variables related to distance to transit show significant negative correlation with the TTI, which is consistent with empirical conclusions. And the final regression model can be expressed as

$$\begin{aligned}\mathrm{TTI} = {} & 0.91 + 0.35X_1 + 1.19X_2 + 0.19X_4 - 0.11X_6 - 0.26X_7 \\ & + 0.19X_8 + 0.20X_{10} + 0.61X_{11} + 0.49X_{13} \\ & - 0.16X_{14} + 0.63X_{15} - 0.34X_{16} - 0.37X_{17} \\ & - 0.62X_{18} - 0.30X_{19} - 0.25X_{20}. \end{aligned} \quad (2)$$

**Table 2: Results of multiple regression model (OSL)**

| Ds | variable | coefficient | t-value | p-value |
|---|---|---|---|---|
| | Constant | 0.914 | 39.116 | *** |
| Diversity | $X_1$ | 0.348 | 4.764 | *** |
| | $X_2$ | 1.194 | 15.282 | *** |
| | $X_4$ | 0.189 | 4.507 | *** |
| | $X_6$ | -0.108 | -2.186 | *** |
| | $X_7$ | -0.258 | -3.067 | *** |
| | $X_8$ | 0.190 | 4.582 | *** |
| Density | $X_{10}$ | 0.198 | 3.236 | *(0.032) |
| | $X_{11}$ | 0.611 | 9.445 | *** |
| Design | $X_{13}$ | 0.490 | 1.035 | **(0.003) |
| | $X_{14}$ | 0.163 | -3.394 | *** |
| | $X_{15}$ | 0.627 | 15.115 | *** |
| | $X_{16}$ | -0.335 | -8.573 | *** |
| | $X_{17}$ | -0.356 | -2.433 | *** |
| Distance to transit | $X_{18}$ | -0.617 | -9.128 | *** |
| | $X_{19}$ | -0.259 | -7.187 | *** |
| | $X_{20}$ | -0.251 | -7.846 | *** |
| | $R^2_{adj}$=0.188 | | | |

Note:
***means that p-value is less than 0.001 (extremely significant).
**means that p-value is less than 0.05 (highly significant).
* means that p-value is less than 0.01 (significant).

Although formula (2) established a linear relationship between TTI and 16 significant factors, and the model passed the significance test, the R squared was very low with only 0.188. We consider it may be because linear models are not suitable for describing the potential relationship between TTI and variables. Besides, a noteworthy result is that some factors positively correlated with TTI in the univariate regression model, but show negative correlation in the multiple regression model, such as the number of educational and cultural facilities. This may be due to the multicollinearity between variables. To verify our hypothesis, we firstly resampled the data points according to an appropriate step size, replacing TTI within a step range with its mean. Then five non-linear function models (i.e., $y = \log_a x$; $y = a\ln x + b$; $y = ax^2 + bx + c$; $y = ae^x$; $y = ax^b$ ) were adopted to fit the underlying relationship, and the optimal model with the highest R squared was selected. Here we show several examples in **Fig. 9**. It can be found that the curves in **Fig.9(a)-(c)** are very similar, and the same as **Fig.9(d)-(f)**, implying that one factor may replace the other two. In other words, there may be multicollinearity between the factors with the same optimal model. To further prove our hypothesis, we perform linear regression on these factors, and the results were shown in **Table3**. The p-values are smaller than 0.001, and the R squared is relatively high. Thus, we can draw a conclusion that there was indeed multicollinearity between the factors with similar fitting model, although TTI and each factor display non-linear relationship.

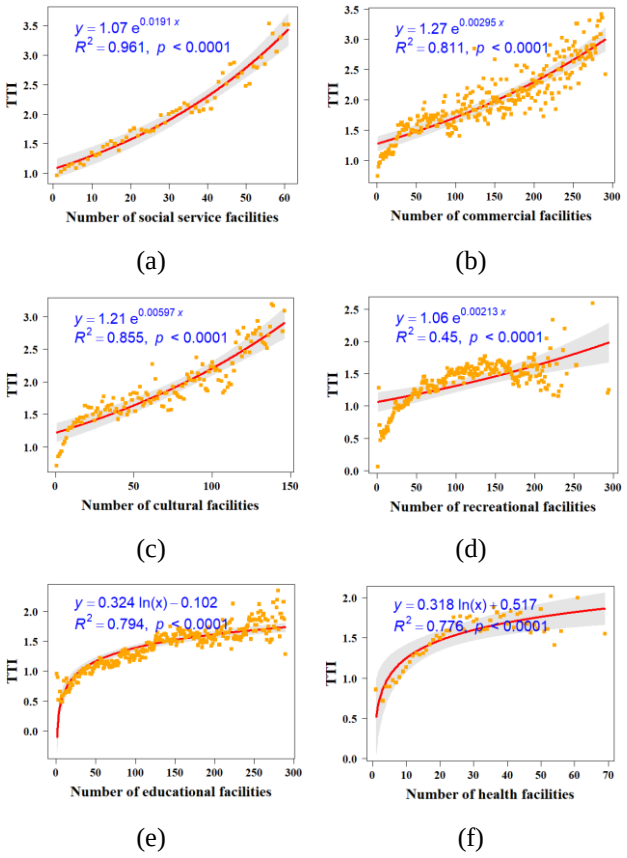



**Figure 9: Illustration of the optimal non-linear model**

**Table 3: Results of multiple regression model (OS)**

| Dependent variable | Independent variable | coefficient | R squared | p-value |
|---|---|---|---|---|
| $X_2$ | $X_1$ | 0.884 | 0.816 | *** |
| | $X_7$ | 0.849 | 0.865 | *** |
| $X_4$ | $X_6$ | 0.629 | 0.648 | *** |
| | $X_8$ | 0.810 | 0.512 | *** |

To further explore the potential factors of traffic congestion, we have to overcome the multicollinearity among the 16 significant factors. The Pearson correlation coefficients including significant test results are calculated between the 16 factors, as shown in **Fig.10**. The results are consistent with that in **Table 3**. Therefore, we replace $X_1$ and $X_7$ by $X_2$, and replace $X_6$ and $X_8$ by $X_4$. Finally, only 12 factors are reserved, and the Pearson correlation of them is shown in **Fig.11**. The reserved 12 factors pass the significant test of multiple linear regression model, and overcome the multicollinearity between variables. We regard the reserved 12 factors as the potential factors of traffic congestion, which may characterize the underlying mechanism of the spatio-temporal congestion patterns in NYC.

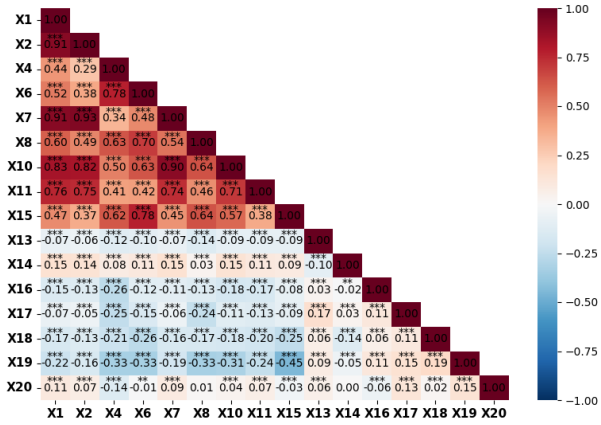



**Figure10: Pearson correlation analysis on 16 significant factors** (X1: the number of social service facilities; X2: commercial facilities; X4: recreational facilities; X6: educational facilities; X7: cultural facilities; X8: healthcare facilities; X10: population density; X11:building height; X13: road length; X14: road width; X15: number of street lights;X16: distance to the nearest bridge;X17: length of bike routes; X18: distance to the nearest bus stop; X19: distance to the nearest subway entrance; X20: distance to the nearest domestic airport.)

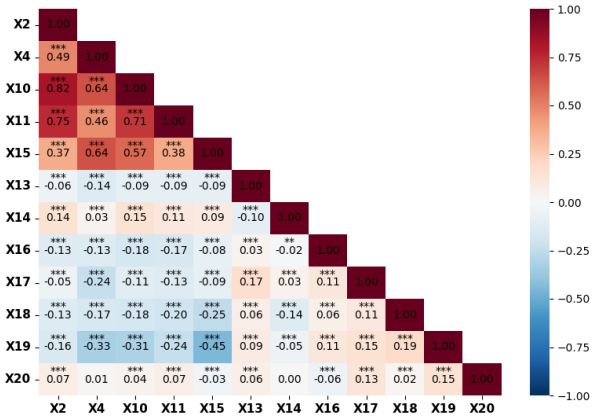



**Figure 11: Pearson correlation analysis on 12 significant factors**

# 5 Discussion and conclusions

Traffic congestion poses a significant challenge to urban sustainability, affecting travel efficiency, economic development, the environment, and residents' quality of life. Understanding its spatio-temporal patterns and influential factors is essential for effective traffic management policies and sustainable transportation. Existing work lack a systematic investigation for this issue. This study contributes to the literature by analyzing a total of 2,221,916 average

travel speed records in the NYC from Dec1, 2018 to Dec 31, 2018. The empirical findings yield three main conclusions:

i. Urban traffic congestion performance varies with different types of days (i.e., Mondays, normal weekdays, weekends and holidays) and different boroughs in the NYC, but the spatio-temporal patterns at borough level are consistent with that of city level. The morning rush hours on Mondays are more congested than other working days, while the evening rush hours display an opposite trend. Holidays are usually more congested than weekends, with the exception of Manhattan, while Manhattan presents the strongest traffic pressure among the five boroughs. It can be due to that Manhattan is the commercial center of NYC but limited by the old road network.

ii. The 20 factors selected in the experiment are found to be significantly correlated with traffic congestion, but only 16 factors are included in the multiple regression model. The model indicates that commercial facility, building height and street lights are relatively more notable than the other factors, which may provide inspiration for policy formulation to alleviate traffic congestion.

iii. Non-linear models and Pearson correlation analysis are utilized to explore the factors potentially shaping traffic congestion , addressing multicollinearity that suggests overlapping information among variables. As a result, only 12 key factors are retained.

The proposed framework for exploring the spatio-temporal patterns and the potential factors of traffic congestion can also provide a reference for other cities to understand the underlying mechanism of traffic congestion and formulate effective policies. Based on this study's empirical findings, NYC residents may avoid peak congestion by choosing optimal commuting times, while policymakers can implement targeted measures to alleviate congestion. The policy measures derived from the analysis results can be suggested as follows:

1) To address the issue of "Monday morning rush hour congestion", targeted policies could restrict private vehicle passage during specific periods or in certain areas, promote public transportation, and reduce the number of vehicles on the road. Employers could encourage flexible work hours, allowing employees to commute outside peak times, thereby dispersing traffic flow. Elderly residents might also consider staggered travel to avoid peak hours[11].

2) Given the high traffic pressure in Manhattan on weekends, public transportation could be optimized by increasing the efficiency, capacity, and frequency of buses, subways, or light rail services. On the other hand, urban land use and planning should be considered with traffic congestion issues. Taking into account the transportation and compactness of land use patterns, traffic congestion can be alleviated by rational diversion [12].

3) Targeting the explored potential factors of traffic congestions, it reminds traffic management department to pay more attention to commercial areas and building height. In some specific areas with severe congestion, it is possible to consider limiting the height of buildings. Lower building heights can reduce the likelihood of obstructing traffic views and causing shadow effects, improving the visibility and safety of road traffic[13].

This study also highlights areas for further research. Future work should explore extended temporal patterns of traffic congestion, assessing seasonal and periodic variations. Additionally, given the multifaceted nature of congestion, incorporating factors such as weather conditions, precise road structures, and remote sensing data would enhance understanding and provide a more comprehensive view of urban traffic dynamics.

# REFERENCES


[1] Chris Wright and Penina Roberg. 1998. The conceptual structure of traffic jams. Transport Policy 5, 1 (1998), 23–35.

[2] Jintao Ke, Hai Yang, and Zhengfei Zheng. 2020. On ride-pooling and traffic congestion. Transportation Research Part B: Methodological 142, (2020), 213–231.

[3] Xiaoxuan Wei, Yitian Ren, Liyin Shen, and Tianheng Shu. 2022. Exploring the spatiotemporal pattern of traffic congestion performance of large cities in China: A real-time data based investigation. Environmental Impact Assessment Review 95, (2022), 106808.

[4] Jinchao Song, Chunli Zhao, Shaopeng Zhong, Thomas Alexander Sick Nielsen, and Alexander V Prishchepov. 2019. Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques. Computers, Environment and Urban Systems 77, (2019), 101364.

[5] Reid Ewing and Robert Cervero. 2010. Travel and the built environment: A meta-analysis. Journal of the American planning association 76, 3 (2010), 265–294.

[6] Xiangfu Kong, Jiawen Yang, and Zhongyu Yang. 2015. Measuring traffic congestion with taxi GPS data and travel time index. In CICTP 2015. 3751–3762.

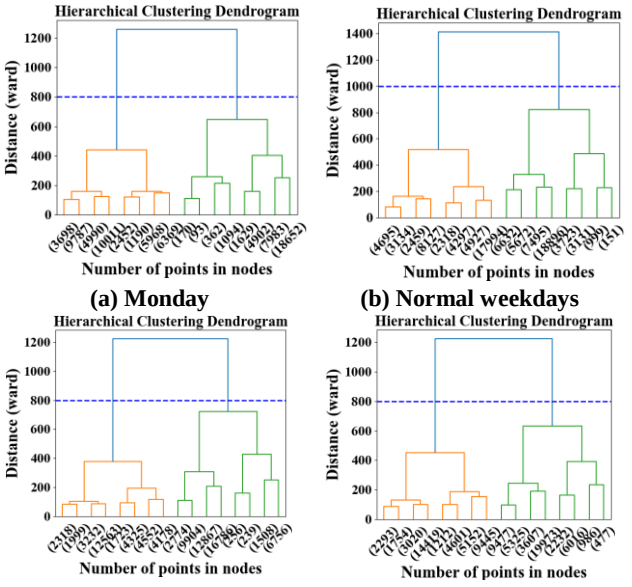
[7] Frank B Baker and Lawrence J Hubert. 1975. Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association 70, 349 (1975), 31–38.

[8] William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification 1, 1 (1984), 7–24.
[9] Paul Geladi and Bruce R Kowalski. 1986. Partial least-squares regression: a tutorial. Analytica chimica acta 185, (1986), 1–17.
[10] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. Noise reduction in speech processing (2009), 1–4.
[11] Rafael HM Pereira. 2019. Future accessibility impacts of transport policy scenarios: Equity and sensitivity to travel time thresholds for Bus Rapid Transit expansion in Rio de Janeiro. Journal of Transport Geography 74, (2019), 321–332.
[12] Jinhuan Wang, Yitian Ren, Liyin Shen, Zhi Liu, Ying Wu, and Fangchen Shi. 2020. A novel evaluation method for urban infrastructures carrying capacity. Cities 105, (2020), 102846.
[13] Chengri Ding. 2013. Building height restrictions, land development and economic costs. Land use policy 30, 1 (2013), 485–495.

# Appendix

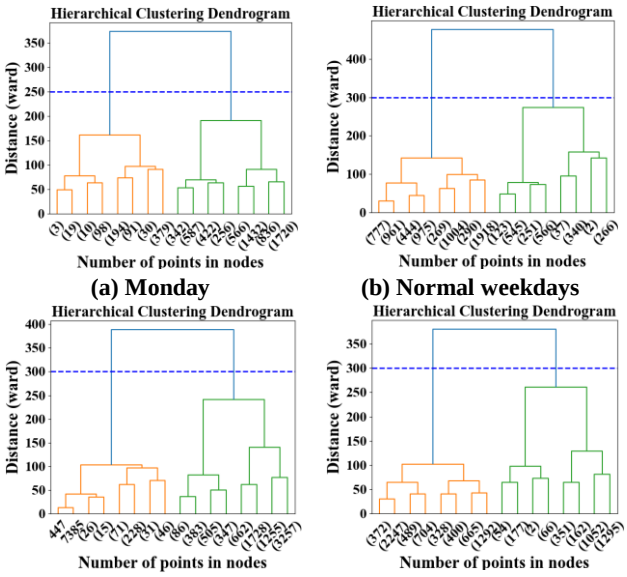
**Table 1: TTI level and its meaning**

| TTI | Traffic congestion performance |
|---|---|
| [0, 1.0) | Free flow |
| [1.0, 1.5) | Smooth |
| [1.5, 2.0) | Mild congestion |
| [2.0, 4.0) | Congestion |
| [4.0, +∞) | Heavy congestion |



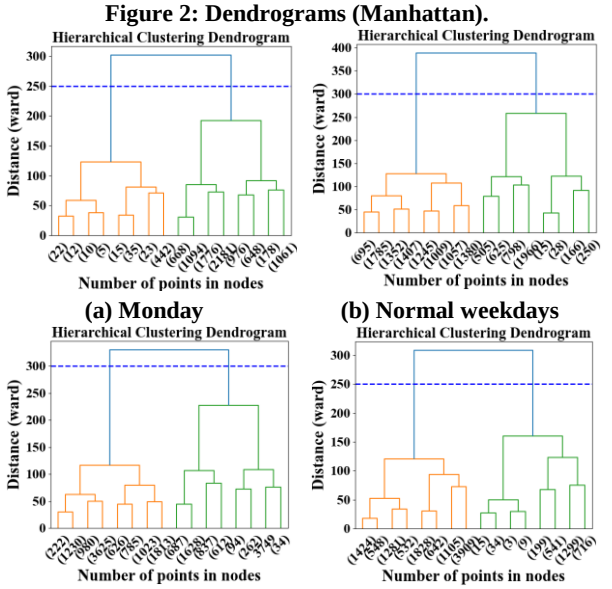

(a) Monday (b) Normal weekdays (c) Weekend (d) Holiday

**Figure 1: Dendrograms (NYC).**



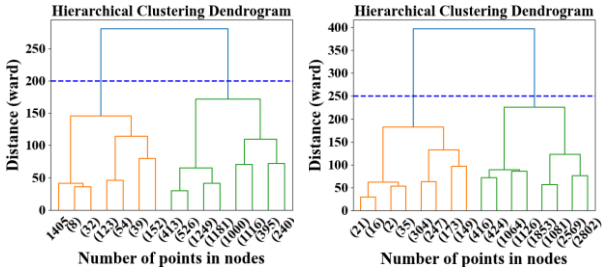

(a) Monday (b) Normal weekdays (a) Weekend (d) Holiday

**Figure 2: Dendrograms (Manhattan).**

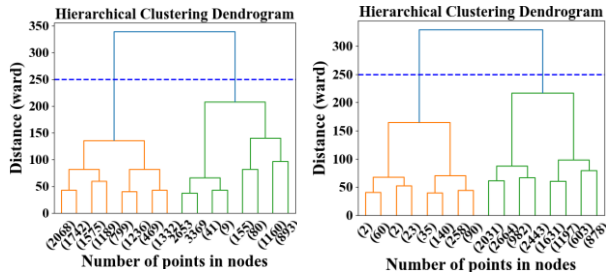



(a) Monday (b) Normal weekdays (c) Weekend (d) Holiday

**Figure 3: Dendrograms (Brooklyn).**

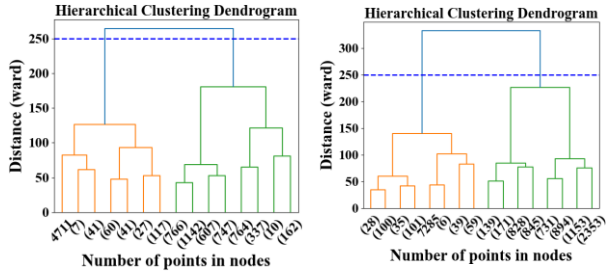



(a) Monday (b) Normal weekdays

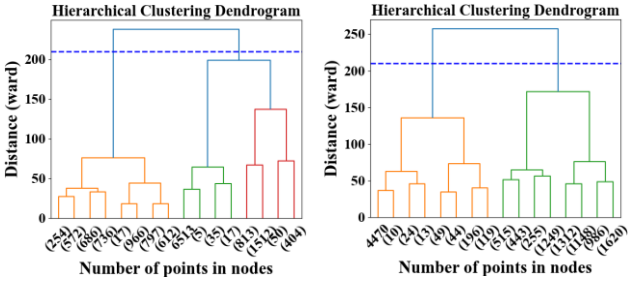



(c) Weekend (d) Holiday

**Figure 4: Dendrograms (Queens).**

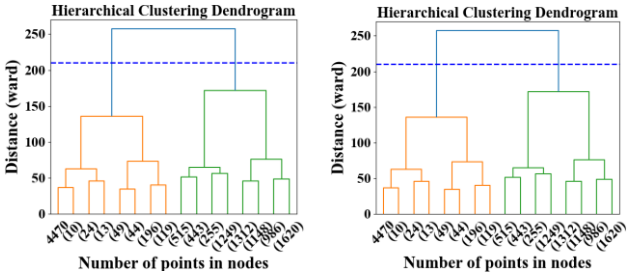



(a) Monday (b) Normal weekdays

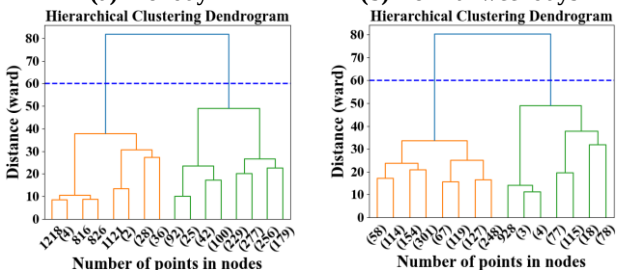



(c) Weekend (d) Holiday

**Figure 5: Dendrograms (Bronx).**




(a) Monday (b) Normal weekdays




(c) Weekend (d) Holiday

**Figure 6: Dendrograms (Staten_Island).**

**Table 2: Results of univariate regression model (20 factors)**

| variable | coefficient | R-squared | t-value | p-value |
|---|---|---|---|---|
| $X_1$ | 0.135 | 0.123 | 109.683 | *** |
| $X_2$ | 0.142 | 0.121 | 112.991 | *** |
| $X_3$ | 0.126 | 0.116 | 111.146 | *** |
| $X_4$ | 0.045 | 0.031 | 50.704 | *** |
| $X_5$ | 0.094 | 0.078 | 89.229 | *** |
| $X_6$ | 0.089 | 0.084 | 95.984 | *** |
| $X_7$ | 0.126 | 0.116 | 111.146 | *** |
| $X_8$ | 0.127 | 0.110 | 107.845 | *** |
| $X_9$ | 0.098 | 0.115 | 110.905 | *** |
| $X_{10}$ | 0.127 | 0.148 | 1128.418 | *** |
| $X_{11}$ | 0.210 | 0.109 | 107.531 | *** |
| $X_{12}$ | -0.119 | 0.008 | -27.998 | *** |
| $X_{13}$ | -0.067 | 0.002 | -9.847 | *** |
| $X_{14}$ | 0.043 | 0.004 | 14.246 | *** |
| $X_{15}$ | 0.112 | 0.112 | 109.036 | *** |
| $X_{16}$ | -0.057 | 0.013 | -34.641 | *** |
| $X_{17}$ | -0.126 | 0.007 | -21.977 | *** |
| $X_{18}$ | -0.117 | 0.026 | -50.528 | *** |
| $X_{19}$ | -1.240 | 0.057 | -75.967 | *** |
| $X_{20}$ | -0.160 | 0.116 | -40.601 | *** |

**Table 3: Results of multiple regression model (20 factors)**

| Ds | variable | coefficient | t-value | p-value |
|---|---|---|---|---|
| | Constant | 0.979 | 31.918 | *** |
| Diversity | $X_1$ | 0.332 | 4.292 | *** |
| | $X_2$ | 1.214 | 14.261 | *** |
| | $X_3$ | 0.039 | 1.444 | 0.149 |
| | $X_4$ | 0.218 | 4.590 | *** |
| | $X_5$ | 0.004 | 0.100 | 0.921 |
| | $X_6$ | -0.208 | -4.588 | *** |
| | $X_7$ | -0.217 | -2.421 | * |
| | $X_8$ | 0.169 | 3.504 | *** |
| | $X_9$ | -0.031 | 0.641 | 0.522 |
| Density | $X_{10}$ | 0.202 | 3.210 | ** |
| | $X_{11}$ | 0.565 | 8.570 | *** |
| | $X_{12}$ | -0.154 | -0.789 | 0.430 |
| Design | $X_{13}$ | 0.425 | 3.322 | ** |
| | $X_{14}$ | -0.170 | -3.398 | ** |
| | $X_{15}$ | 0.699 | 14.836 | *** |
| | $X_{16}$ | -0.300 | -7.098 | *** |
| | $X_{17}$ | -0.796 | -4.665 | *** |
| Distance to transit | $X_{28}$ | -0.726 | -9.367 | *** |
| | $X_{19}$ | -0.277 | -5.782 | *** |
| | $X_{20}$ | -0.297 | -7.350 | *** |
| $R^2_{adj}$=0.188 | | | | |