Exploring the Spatial and Temporal Relationship between Shared Bike Data and COVID Cases in New York City and Boston area during 2019-2022

1 INTRODUCTION

COVID-19 has caused significant changes in daily life, including transportation habits. Shared bikes have become a more popular alternative to public transit, making it essential to investigate their relationship with COVID-19 cases, especially in urban areas. Prior studies have employed spatial and temporal analysis methods to explore this relationship. Spatial methods investigate the distribution of shared bike usage and COVID-19 cases, while temporal methods examine their changes over time. Combining these methods offers a more complete understanding of the relationship.

Temporal analysis methods have been used to investigate the relationship between shared-bike usage and COVID-19 cases. For example, Padmanabhan et al. (2021) conducted a time-series analysis in US cities to understand the impacts of COVID-19 on biking, while Mehdizadeh Dastjerdi and Morency (2022) used the Autoregressive integrated moving average (ARIMA) model to predict pickup demand in Montreal. Spatial analysis methods have also been used to map the distribution of shared bike stations and COVID-19 cases. Combining these methods can provide a more complete understanding of the relationship, as demonstrated by Hu et al. (2021) in their spatio-temporal analysis of bike-sharing usage across the pandemic in Boston. Such analysis can inform public health policies related to shared bike usage.

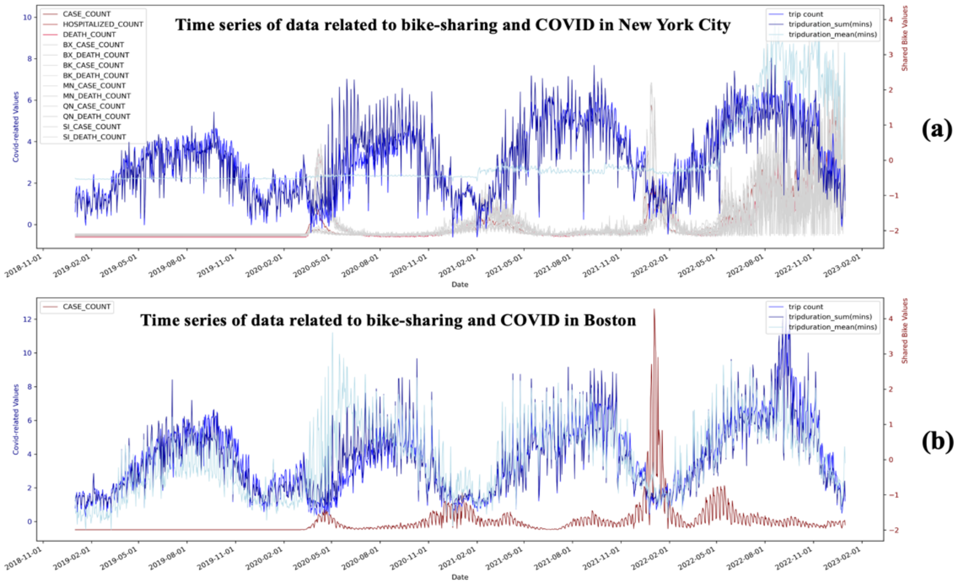

Figure 1. NYC and Boston datasets time series.
(a for NYC, b for Boston City area)

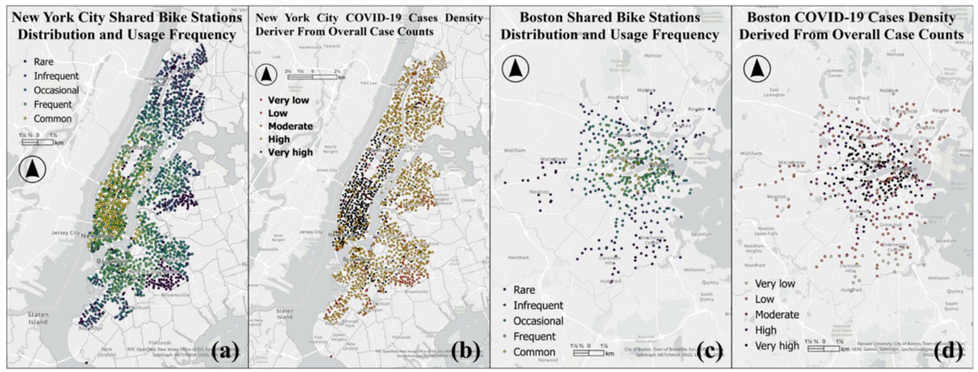

Figure 2. Distribution of NYC and Boston shared bike stations and corresponding usage frequency and COVID cases density. (a, b for NYC and c, d for Boston)

Overall, the use of spatial and temporal analysis methods has undoubtedly contributed to a deeper understanding of the relationship between shared-bike usage and COVID-19 cases. While these methods have their limitations, they have enabled researchers to identify areas at higher risk for COVID-19 transmission and inform public health policies related to shared bike usage. Moving forward, it will be important to continue to refine and develop these methods to ensure that they remain effective tools for studying the impacts of COVID-19 on transportation and other aspects of daily life.

This project aims to explore the spatial and temporal relationship between shared-bike data and COVID-19 cases in NYC and Boston during 2019-2022. By employing a combination of spatial and temporal analysis methods, including GIS and time-series models, a better understanding of the relationship would be gained between these variables and inform policy decisions related to transportation and public health in urban areas.

The datasets used for this project were all obtained from open sources. For example, the shared bike datasets were accessed from CityBike and Bluebikes including multiple records per day, and the COVID-related datasets were given by NYC Open Data and the Boston Government. The study areas are NYC and Boston City area with the daily temporal resolution and zip code spatial resolution. The datasets used contain serval variables such as trip count, trip duration time, trip ID, station information (geo-stamped), user gender, user age group and membership kinds etc. The temporal features could be observed in figure 1, and the spatial distribution is shown in figure 2.

2 Exploratory spatio-temporal data analysis

Exploratory spatio-temporal data analysis (ESTDA) is an essential way for investigating the spatial and temporal characteristics of datasets. In the context of this project on exploring the relationship between shared-bike data and COVID cases in New York City (NYC) and Boston from 2019 to 2022, ESTDA can provide insights into the underlying patterns and trends of the data.

To begin with, global and local Moran's I statistics could explore the spatial autocorrelation of the data and calculated as shown in figure 3(a) and 3(b) using the spatial matrix generated by the K-Nearest Neighbors (KNN) algorithm which the distance and adjacent was considered by. Global Moran's I measures the overall spatial clustering of the data, while local Moran's I identifies specific locations where the data is clustered or dispersed as shown in figures 3(b) and 3(d) from both cluster and significance map. By visualizing these spatial patterns, we can gain insights into the spatial relationships between shared bike data and COVID cases in different parts of NYC and Boston.

The global Moran I value was -0.219 for New York City and 0.188 for Boston. the LISA clustering and significance maps show that most portions of New York City (74%) and Boston (78%) were insignificant, with the low-low and high-high portions representing 20.22% of the New York City study area.

Based on the global Moran's I and LISA cluster analysis, it appears that the spatial patterns of shared bike trip count in both New York City and Boston area are not significantly clustered. However, the presence of some high-high and low-low clusters in New York City suggests that there may be underlying factors that contribute to spatial variation in trip counts.

Furthermore, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to examine the temporal autocorrelation of the data. The ACF and PACF plots can help us identify the statistically significant lag periods, indicating the presence of temporal patterns in the data. The trip counts, trip duration time, and COVID cases variables exhibit cyclical patterns that suggest a degree of seasonality according to figure 1 and the results from ACF and PACF were conducted further for each variable that is not plotted here due to space constraints.

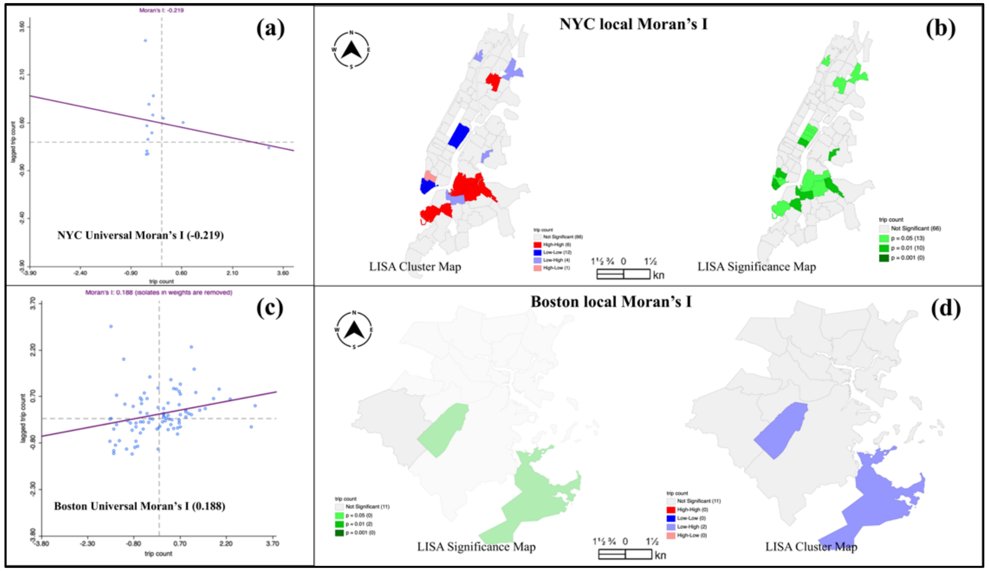

Figure 3. NYC and Boston spatial correlations (a is the global Moran's I scatter plot for NYC, b is the Local Indicators of Spatial Autocorrelation, (LISA) including cluster and significance maps for NYC; c is the global Moran's I scatter plot for Boston, d is the LISA cluster and significance maps for Boston).

Basic temporal characteristics could be observed from the histogram and time series in figure 4(1) for NYC and figure 5(1) for Boston. The ACF and PACF calculated for the trip count variable indicate a significant autocorrelation at lag 7, which suggests a weekly seasonality in trip counts.

The PACF plot also shows significant spikes at lags 1 and 2, which suggest a first- and second-order autoregressive process in the data. For the trip duration time variable, the ACF and PACF plots show significant autocorrelation at lag 1 and some evidence of a seasonal component at lags 5 and 6, which suggests a weekly seasonality in trip duration times. Finally, for the COVID cases variable, there is no fixed cyclic pattern presented. Overall, these results suggest that our dataset exhibits cyclicality and seasonality. It is clear from these results that further exploration of the temporal characteristics of the dataset is necessary.

By using ACF and PACF to explore the temporal autocorrelation and global and local Moran's I to explore the spatial autocorrelation, a deeper understanding of the dataset and further exploration is necessary for spatio-temporal relationships between shared bike data and COVID cases in NYC and Boston based-on insights above.

3 Methodology and results

In this study, a combination of time-series and spatial analysis methods was employed to explore the relationship between shared-bike data and COVID cases in NYC and Boston during 2019-2022. Specifically, this project utilized Seasonal ARIMA (SARIMA) for time-series analysis based on ARIMA, Multiscale Geographically Weighted Regression (MGWR) for spatial analysis and Mixed Geographically and Temporally Weighted Regression (MGTWR) for spatio-temporal exploration.

For temporal analysis, further insights are essential for the modelling besides the features obtained from previous ESTDA. The results for each step are as follows.

1. **Data preparation:** aggregating data by day and performing different operations on variables, such as sum, mean, count, etc. Then, a data frame with the shape of 1460*17 and 1459*5 was taken as input for NYC and Boston respectively.
2. **Decomposition:** to get the trend, seasonality and residuals for trip duration time (mins), trip counts and COVID cases respectively in both NYC and Boston as shown in lines (3) to (5) of figure 4 and 5. There are significantly increasing trends and 12 months cycle for both shared bike and COVID variables.
3. **Augmented Dickey–Fuller (ADF) test for the original time series:** the ADF test (Mushtaq, 2011) suggests that the original time series may not be stationary, as the p-value is greater than the significance level of 0.05 and the ADF statistic is between the 5% and 1% critical values.
4. **Differencing:** 6 variables for 2 cities using first-order differencing.
5. **ADF test for the differenced time series:** ADF test results show that the time series data is stationary after first-order differencing.
6. **ACF & PACF for differenced time series:** determining the parameters of the ARIMA model through the variation of ACF and PACF.

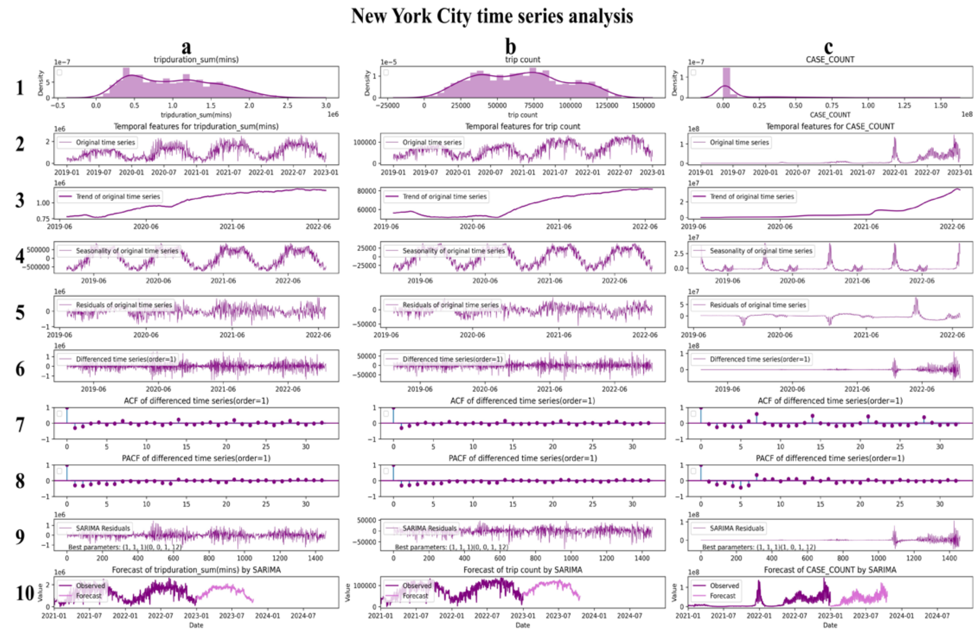

Figure 4. NYC temporal analysis (a, b, c for trip duration time (mins), trip counts and COVID cases respectively. Subplots are accessed by indexing in this report, e.g., the first subplot is referenced as *figure 4 (1a)),* the first line is referenced as *figure 4(1)*).

7. **ARIMA model fitting:** observing the performance of the ARIMA model by running it in the background to determine if the parameters are appropriate best parameters selection for the ARIMA model. Also preparing for the SARIMA model.

8. **Best parameters selection for SARMA:** best parameters selected based on BIC due to huge data volumes according to Zhao, Jin and Shi (2015).
9. **SARIMA model fitting:** fitting SARIMA models with optimal parameters.
10. **Cycle forecasting:** forecasting for the following 12 months by an optimised algorithm of day-by-day forecasting, rather than forecasting all data at once.

Based on the results provided in table 1, we can observe that for all the time series variables in both NYC and Boston, the ADF test was conducted to check for stationarity. The p-values for all variables are greater than 0.05, indicating that the null hypothesis of non-stationarity cannot be rejected at a 5% significance level. However, differencing was applied to the time series variables to achieve stationarity for SARIMA modelling.

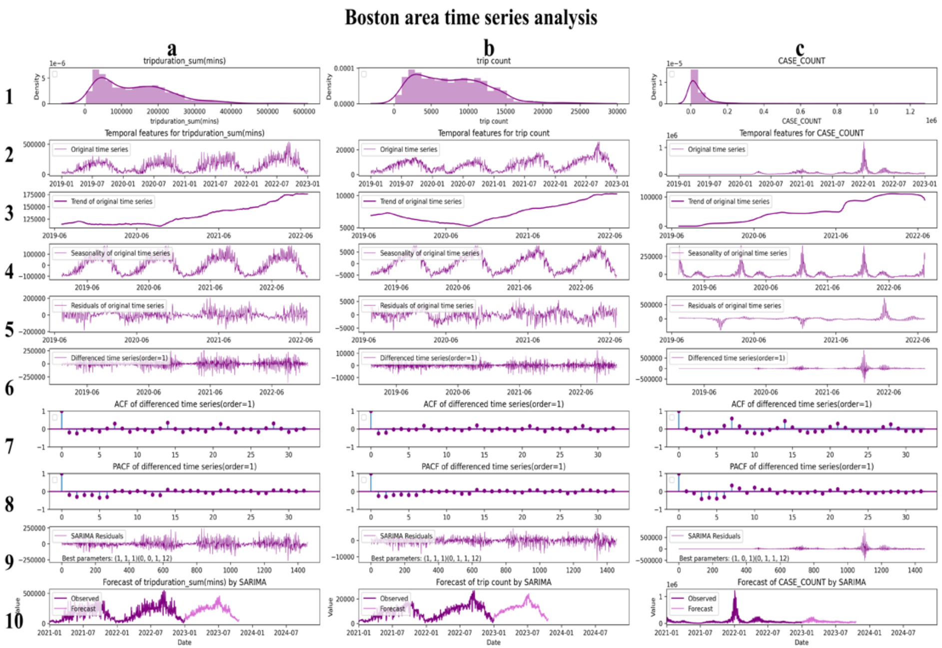

Figure 5. Boston area temporal analysis (a, b, c for trip duration time (mins), trip counts and COVID cases respectively. Subplots are accessed by indexing in this report, e.g., the first subplot is referenced as *figure 5(1a)),* the first line is referenced as *figure 5(1)*).

For the NYC trip duration time and Boston trip duration time, the best SARIMA models were (1,1,1)(1,0,1,12) and (1,1,1)(0,0,1,12) respectively. In both cases, the autoregressive coefficient was positive, indicating that the current value of the variable is positively influenced by its previous values. The autoregressive seasonal coefficient was negative in both cases, suggesting that the seasonal component harms the current value of the variable.

For NYC trip counts and Boston trip counts, the best SARIMA models were (1,1,1)(0,0,1,12) and (1,1,1)(0,1,1,12) respectively. The autoregressive coefficient was positive in both cases, indicating that the current value of the variable is positively influenced by its previous values. The autoregressive seasonal coefficient was negative for NYC trip counts, and very negative for Boston trip counts, suggesting that the seasonal component has a significant negative impact on the current value of the variable in Boston.

For NYC COVID cases and Boston COVID cases, the best SARIMA models were (1,1,1)(1,0,1,12) and (1,0,1)(0,1,1,12) respectively. The autoregressive coefficient was positive for NYC COVID cases, and relatively high for Boston COVID cases, indicating that the current value of the variable is positively influenced by its previous values. The autoregressive seasonal coefficient was positive for NYC COVID cases and very negative for Boston COVID cases, suggesting that the seasonal component has a significant impact on the current value of the variable in Boston.

Finally, the Ljung-Box probability test (Hassani and Yeganegi, 2019) was conducted to check for the presence of residual autocorrelation, and the heteroskedasticity test was performed to check for the presence of non-constant variance. For all variables, the Ljung-Box probability test was not significant at a 5% significance level, indicating that there is no evidence of residual autocorrelation. Additionally, the heteroskedasticity test (Davidson, Mackinnon and Davidson, 1985) was not significant at a 5% significance level, suggesting that there is no evidence of non-constant variance in the residuals.

Table 1. Results of temporal modelling analysis of NYC and Boston on a shared bike and COVID cases

Time series	The P-Value of ADF Test		SARIMA Model summary				
	Original time series	Differenced time series	Best SARIMA parameters	Autoregressive Coefficient	autoregressive seasonal coefficient	Ljung-Box Probability	Heteroskedasticity Probability
NYC trip duration time (mins)	0.184	0	(1, 1, 1) (1, 0, 1, 12)	0.2554	-0.076	0.3	0
NYC trip counts	0.222	0	(1, 1, 1) (0, 0, 1, 12)	0.2957	-0.0596	0.5	0
NYC COVID cases	0.208	0	(1, 1, 1) (1, 0, 1, 12)	0.4825	0.5942	0	0
Boston trip duration time (mins)	0.229	0	(1, 1, 1) (0, 0, 1, 12)	0.3393	-0.1402	0.03	0
Boston trip counts	0.120	0	(1, 1, 1) (0, 1, 1, 12)	0.3298	-1.0013	0.18	0
Boston COVID cases	0.000	0	(1, 0, 1) (0, 1, 1, 12)	0.8147	-0.9792	0.98	0

Overall, the SARIMA models were suitable for modelling these time series, but further analysis and validation may be necessary.

For spatial analysis, the MGWR models were used for NYC and Boston. The results of each step are as follows.

1. **Data preparation**: aggregating data by shared bike stations, performing different calculations on variables, such as sum, mean, count, etc. Scaled and specify the dependent variable as trip counts, explanatory variables as start station id, covid cases, trip duration sum (mins), trip duration mean (mins), and user type count.
2. **Spatial weighed matrix construction:** spatial weight matrix was constructed based on the distance between the observations. This matrix will be used to weigh the observations in the regression model.
3. **Parameters selection:** using distance with the golden research method to automatically determine the number of neighbours to include in the local regression estimation for each station. The weights were based on the distance between observations in kilometres and the gaussian function was used as the local weighting scheme.
4. **Model Fitting:** using the settled parameters to fit MGWR models for NYC and Boston.
5. **Results visualization:** focusing on the relationship between COVID cases and trip count based on the topic of our study**.**

Table 2. Results of MGWR for NYC and Boston on the shared bike and COVID cases

Model	Adjusted R2	Time Cost (mins)	Optimal Bandwidth (km)	Importance of COVID Cases
NYC MGWR	0.9993	51.92	1.96	96.70%
Boston MGWR	0.9991	0.26	3.67	100%

For spatio-temporal analysis, the MGTWR was conducted, and the results of each step are as follows.

1. **Data preparation:** aggregating data by 'date' and 'station' simultaneously, using different calculation methods such as sum, mean, count, etc. Also, processing the dataset according to the input requirements of the MGTWR model, e.g., converting date to integer timestamp, etc.
2. **Bandwidth selection:** selecting the best parameters employing a defined parameter search method.
3. **MGTWR fitting:** fitting the MGTWR model using optimal parameters. This step was not possible due to a lack of memory.

Combining Figure 5 and Table 2, there is a spatial pattern of decreasing frequency of shared bike usage in both NYC and Boston, fading from the city centre to the surroundings. It can also be seen that the COVID cases variable has a strong relationship and correlation with the usage frequency of shared bikes, which can explain the spatial distribution of the usage pattern of shared bikes very well.

Overall, our methodology and results provide a comprehensive analysis of the spatio-temporal relationships between shared bike data and COVID cases in NYC and Boston. The combination of SARIMA, MGWR, and MGTWR models allowed us to explore both the temporal and spatial dimensions of the data, providing valuable insights into the underlying patterns and trends of the data.

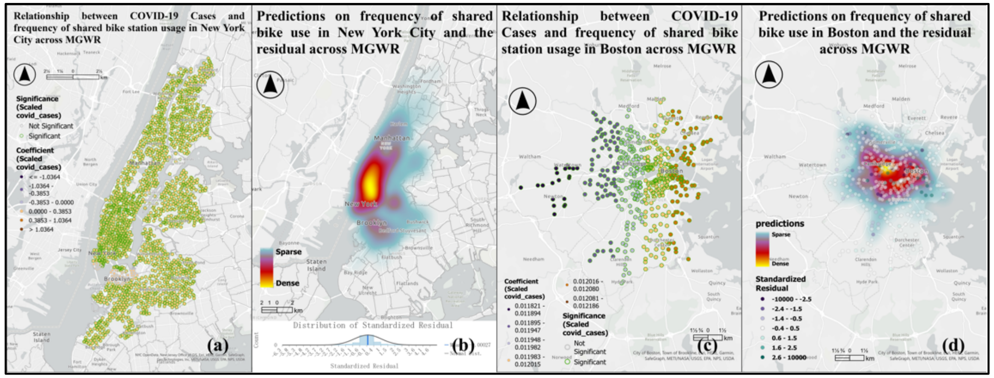

Figure 5. Relationship between shared bikes and COVID cases and predictions by MGWR in NYC and Boston

4 Discussion and conclusions

In conclusion, this study utilized a combination of time-series and spatial analysis methods to explore the relationship between shared-bike data and COVID cases in NYC and Boston. The results of the study suggest that there is a significant relationship between these variables and that the trends and seasonality components play a crucial role in the variation of these variables over time.

The SARIMA models were used to forecast the future values of the time-series variables, and it was observed that the seasonal component had a significant impact on the current value of the variable in some cases. The spatio-temporal analysis using MGWR and MGTWR provided valuable insights into the spatial patterns of shared-bike usage and COVID cases.

These findings have important implications for policymakers and city planners, as they can use this information to allocate resources and implement targeted interventions to mitigate the spread of COVID-19 and promote the use of shared bikes in areas where it is most needed.

There are some limitations to this study. For example, the dimensionality of the data was not selected sufficiently, which may cause potential multicollinearity problems and overfitting. When fitting the MGTWR, there is still insufficient memory to perform after aggregation either by week or month due to the sheer volume of data. For this issue, the algorithm developer Sun (n.d.) emailed the author of this report a response that its schematic design was not friendly to large data sets.

Future research could explore other factors that may influence the relationship between shared-bike usage and COVID cases, such as weather conditions, events, and transportation infrastructure.

5 Code availability

This project is based on the implementation of Python 3.8 and the corresponding version of the dependency libraries. The specific code and resources can be accessed via GitHub.

The raw data is available via links in the references and there are links to them in the GitHub code block also. Therefore, this report would not upload the raw data (over 23GB) to Moodle and GitHub, but all related outputs are given in this report.

References

Padmanabhan, V., Penmetsa, P., Li, X., Dhondia, F., Dhondia, S. and Parrish, A. (2021). COVID-19 effects on shared-biking in New York, Boston, and Chicago. Transportation Research Interdisciplinary Perspectives, 9, p.100282. doi:https://doi.org/10.1016/j.trip.2020.100282.

Mehdizadeh Dastjerdi, A. and Morency, C. (2022). Bike-Sharing Demand Prediction at Community Level under COVID-19 Using Deep Learning. Sensors, 22(3), p.1060. doi:https://doi.org/10.3390/s22031060.

Li, X., Xu, Y., Zhang, X., Shi, W., Yue, Y. and Li, Q. (2023). Improving short-term bike sharing demand forecast through an irregular convolutional neural network. Transportation Research Part C: Emerging Technologies, 147, p.103984.

Xin, R., Ding, L., Ai, B., Yang, M., Zhu, R., Cao, B. and Meng, L. (2023). Geospatial Network Analysis and Origin-Destination Clustering of Bike-Sharing Activities during the COVID-19 Pandemic. ISPRS International Journal of Geo-Information, 12(1), p.23. doi:https://doi.org/10.3390/ijgi12010023.

Hu, S., Xiong, C., Liu, Z. and Zhang, L. (2021). Examining spatiotemporal changing patterns of bike-sharing usage during COVID-19 pandemic. Journal of Transport Geography, 91, p.102997. doi:https://doi.org/10.1016/j.jtrangeo.2021.102997.

citibikenyc.com. (n.d.). Citi Bike System Data | Citi Bike NYC. [online] Available at: https://citibikenyc.com/system-data.

Blue Bikes Boston. (n.d.). Bluebikes System Data. [online] Available at: https://www.bluebikes.com/system-data.

NYC Open Data. (n.d.). NYC Open Data. [online] Available at: https://data.cityofnewyork.us/browse?q=covid&sortBy=relevance [Accessed 31 Mar. 2023].

Boston.gov. (2022). COVID-19 in Boston. [online] Available at: https://www.boston.gov/government/cabinets/boston-public-health-commission/covid-19-boston

Mushtaq, R. (2011). Augmented Dickey Fuller Test. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1911068.

Hassani, H. and Yeganegi, M.R. (2019). Sum of squared ACF and the Ljung–Box statistics. Physica A: Statistical Mechanics and its Applications, 520, pp.81–86. doi:https://doi.org/10.1016/j.physa.2018.12.028.

Davidson, Mackinnon and Davidson (1985). Heteroskedasticity-Robust Tests in Regressions Directions. Annales de l'inséé, (59/60), p.183. doi:https://doi.org/10.2307/20076563.

Zhao, J., Jin, L. and Shi, L. (2015). Mixture model selection via hierarchical BIC. Computational Statistics & Data Analysis, [online] 88, pp.139–153. doi:https://doi.org/10.1016/j.csda.2015.01.019.

Sun, K. (n.d.). mgtwr. [online] PyPI. Available at: https://pypi.org/project/mgtwr/ [Accessed 31 Mar. 2023]

The results of models

Temporal analysis for NYC

```
Merge process is running...

========================df_temporal========================
(1460, 17)
            start station id  trip count  ...  SI_CASE_COUNT  SI_DEATH_COUNT
date                                      ...                               
2019-01-01               750       21932  ...            0.0             0.0
2019-01-02               756       37773  ...            0.0             0.0
2019-01-03               758       41644  ...            0.0             0.0
2019-01-04               757       43893  ...            0.0             0.0
2019-01-05               744       17416  ...            0.0             0.0

[5 rows x 17 columns]
start station id              750.0000
trip count                  21932.0000
tripduration_sum(mins)     354043.4998
tripduration_mean(mins)     11623.3758
CASE_COUNT                      0.0000
HOSPITALIZED_COUNT              0.0000
DEATH_COUNT                     0.0000
BX_CASE_COUNT                   0.0000
BX_DEATH_COUNT                  0.0000
BK_CASE_COUNT                   0.0000
BK_DEATH_COUNT                  0.0000
MN_CASE_COUNT                   0.0000
MN_DEATH_COUNT                  0.0000
QN_CASE_COUNT                   0.0000
QN_DEATH_COUNT                  0.0000
SI_CASE_COUNT                   0.0000
SI_DEATH_COUNT                  0.0000
Name: 2019-01-01 00:00:00, dtype: float64
       start station id     trip count  ...  SI_CASE_COUNT  SI_DEATH_COUNT
count       1460.000000    1460.000000  ...   1.460000e+03     1460.000000
mean        4533.413014   66366.094521  ...   6.953576e+05     5809.246575
```

```
std         7542.290548   30285.766864  ...   1.330818e+06    13176.758662
min          143.000000     177.000000  ...   0.000000e+00        0.000000
25%          854.000000   40870.750000  ...   0.000000e+00        0.000000
50%         1129.000000   66511.500000  ...   1.031350e+05        0.000000
75%         2840.250000   87609.500000  ...   5.085470e+05     5385.500000
max        35012.000000  134892.000000  ...   9.657120e+06   138138.000000

[8 rows x 17 columns]
Index(['start station id', 'trip count', 'tripduration_sum(mins)',
       'tripduration_mean(mins)', 'CASE_COUNT', 'HOSPITALIZED_COUNT',
       'DEATH_COUNT', 'BX_CASE_COUNT', 'BX_DEATH_COUNT', 'BK_CASE_COUNT',
       'BK_DEATH_COUNT', 'MN_CASE_COUNT', 'MN_DEATH_COUNT', 'QN_CASE_COUNT',
       'QN_DEATH_COUNT', 'SI_CASE_COUNT', 'SI_DEATH_COUNT'],
      dtype='object')
------Time series analysis for tripduration_sum(mins)-------

The ADF test for original time series:
p-value: 0.184132
ADF Statistic: -2.263138
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

The ADF test for differenced time series (Difference order 1):
p-value: 0.000000
ADF Statistic: -11.174199
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

------------------------SARIMA Model------------------------
Best SARIMA parameters: (1, 1, 1) (0, 0, 1, 12)
Summary of the SARIMA model for tripduration_sum(mins):
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  0   No. Observations:                 1459
Model:             SARIMAX(1, 1, 1)x(0, 0, 1, 12)   Log Likelihood              -20270.718
```

```
Date:                          Thu, 23 Mar 2023   AIC                          40549.436
Time:                                  22:08:59   BIC                          40570.536
Sample:                                       0   HQIC                         40557.312
                                         - 1459
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2554      0.030      8.440      0.000       0.196       0.315
ma.L1         -0.8788      0.017    -51.237      0.000      -0.912      -0.845
ma.S.L12      -0.0760      0.027     -2.795      0.005      -0.129      -0.023
sigma2      1.011e+11   6.07e-14   1.67e+24      0.000    1.01e+11    1.01e+11
===================================================================================
Ljung-Box (L1) (Q):                   1.09   Jarque-Bera (JB):               317.47
Prob(Q):                              0.30   Prob(JB):                         0.00
Heteroskedasticity (H):               1.60   Skew:                            -0.05
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.30
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.43e+39. Standard errors may be unstable.
                    value
2023-01-01  688894.336198
2023-01-02  810223.659315
2023-01-03  846431.815355
2023-01-04  674991.822248
2023-01-05  994937.836550
...                   ...
2023-10-15  395095.566773
2023-10-16  438703.405225
2023-10-17  461302.680828
2023-10-18  546094.523498
2023-10-19  657553.691776

[292 rows x 1 columns]
------------Time series analysis for trip count-------------

The ADF test for original time series:
```

```
p-value: 0.222526
ADF Statistic: -2.156196
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

The ADF test for differenced time series (Difference order 1):
p-value: 0.000000
ADF Statistic: -10.173839
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

------------------------SARIMA Model------------------------
Best SARIMA parameters: (1, 1, 1) (0, 0, 1, 12)
Summary of the SARIMA model for trip count:
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 0   No. Observations:                 1459
Model:             SARIMAX(1, 1, 1)x(0, 0, 1, 12)   Log Likelihood              -15917.714
Date:                            Thu, 23 Mar 2023   AIC                          31843.428
Time:                                    22:31:39   BIC                          31864.529
Sample:                                         0   HQIC                         31851.304
                                           - 1459
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2957      0.030      9.996      0.000       0.238       0.354
ma.L1         -0.8670      0.018    -47.929      0.000      -0.903      -0.832
ma.S.L12      -0.0596      0.030     -2.019      0.043      -0.117      -0.002
sigma2      2.375e+08   4.28e-12   5.55e+19      0.000    2.38e+08    2.38e+08
===================================================================================
Ljung-Box (L1) (Q):                   0.50   Jarque-Bera (JB):               432.22
Prob(Q):                              0.48   Prob(JB):                         0.00
Heteroskedasticity (H):               1.52   Skew:                            -0.81
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.13
```

```
==================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.55e+36. Standard errors may be unstable.
                   value
2023-01-01  53668.536948
2023-01-02  61473.525666
2023-01-03  64667.756700
2023-01-04  54094.835324
2023-01-05  69886.269986
...                  ...
2023-10-15  32561.860393
2023-10-16  37097.585422
2023-10-17  39114.824364
2023-10-18  44383.072306
2023-10-19  49295.189496

[292 rows x 1 columns]
------------Time series analysis for CASE_COUNT-------------

The ADF test for original time series:
p-value: 0.208837
ADF Statistic: -2.192887
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

The ADF test for differenced time series (Difference order 1):
p-value: 0.000000
ADF Statistic: -9.219584
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

------------------------SARIMA Model------------------------
Best SARIMA parameters: (1, 1, 1) (1, 0, 1, 12)
```

```
Summary of the SARIMA model for CASE_COUNT:
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  0   No. Observations:                 1459
Model:             SARIMAX(1, 1, 1)x(1, 0, 1, 12)   Log Likelihood              -25061.601
Date:                            Thu, 23 Mar 2023   AIC                          50133.201
Time:                                    22:34:08   BIC                          50159.577
Sample:                                         0   HQIC                         50143.046
                                           - 1459
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4825      0.015     31.414      0.000       0.452       0.513
ma.L1         -0.8375      0.010    -85.550      0.000      -0.857      -0.818
ar.S.L12       0.5942      0.051     11.597      0.000       0.494       0.695
ma.S.L12      -0.8018      0.048    -16.578      0.000      -0.897      -0.707
sigma2      7.996e+13   1.65e-15   4.86e+28      0.000        8e+13       8e+13
===================================================================================
Ljung-Box (L1) (Q):                  14.61   Jarque-Bera (JB):             85261.85
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):            2896.08   Skew:                             2.83
Prob(H) (two-sided):                  0.00   Kurtosis:                        40.22
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.49e+43. Standard errors may be unstable.
                   value
2023-01-01  2.860755e+06
2023-01-02  2.488141e+06
2023-01-03  2.851294e+06
2023-01-04  1.887989e+06
2023-01-05  2.654590e+06
...                  ...
2023-10-15  4.592846e+07
2023-10-16  5.994604e+07
2023-10-17  5.225294e+07
2023-10-18  7.900210e+07
```

```
2023-10-19  5.371038e+07

[292 rows x 1 columns]
```

Temporal analysis for Boston

```
Merge process is running...

========================df_temporal=========================
(1459, 5)
            start station id  trip count  ...  tripduration_mean(mins)  CASE_COUNT
date                                      ...
2019-01-01               188        1294  ...                3876.6376         0.0
2019-01-02               196        2629  ...                2740.0147         0.0
2019-01-03               202        2999  ...                2904.4085         0.0
2019-01-04               196        3392  ...                2752.2366         0.0
2019-01-05               165         781  ...                2083.3450         0.0

[5 rows x 5 columns]
start station id             188.0000
trip count                  1294.0000
tripduration_sum(mins)     26500.1000
tripduration_mean(mins)     3876.6376
CASE_COUNT                     0.0000
Name: 2019-01-01 00:00:00, dtype: float64
       start station id    trip count  ...  tripduration_mean(mins)    CASE_COUNT
count       1459.000000   1459.000000  ...              1459.000000  1.459000e+03
mean         307.660041   7718.071282  ...              6091.211575  4.510260e+04
std           62.534563   4662.247940  ...              2184.341658  9.273940e+04
min           85.000000    154.000000  ...              1740.462000  0.000000e+00
25%          258.000000   3701.000000  ...              4295.189750  0.000000e+00
50%          311.000000   7261.000000  ...              5941.680200  2.290800e+04
75%          355.000000  11089.000000  ...              7670.961700  5.482100e+04
max          428.000000  26677.000000  ...             13954.989800  1.214487e+06

[8 rows x 5 columns]
Index(['start station id', 'trip count', 'tripduration_sum(mins)',
       'tripduration_mean(mins)', 'CASE_COUNT'],
      dtype='object')
------Time series analysis for tripduration_sum(mins)-------
```

```
The ADF test for original time series:
p-value: 0.229129
ADF Statistic: -2.138988
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

The ADF test for differenced time series (Difference order 1):
p-value: 0.000000
ADF Statistic: -9.513302
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

------------------------SARIMA Model------------------------
Best SARIMA parameters: (1, 1, 1) (0, 0, 1, 12)
Summary of the SARIMA model for tripduration_sum(mins):
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 0   No. Observations:                  1458
Model:             SARIMAX(1, 1, 1)x(0, 0, 1, 12)   Log Likelihood              -17694.892
Date:                            Thu, 23 Mar 2023   AIC                          35397.784
Time:                                    22:46:10   BIC                          35418.882
Sample:                                        0   HQIC                         35405.659
                                          - 1458
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3393      0.033     10.193      0.000       0.274       0.404
ma.L1         -0.8843      0.018    -49.842      0.000      -0.919      -0.850
ma.S.L12      -0.1402      0.026     -5.462      0.000      -0.190      -0.090
sigma2      3.036e+09   1.98e-12   1.53e+21      0.000    3.04e+09    3.04e+09
===================================================================================
Ljung-Box (L1) (Q):                   4.78   Jarque-Bera (JB):               530.15
Prob(Q):                              0.03   Prob(JB):                         0.00
```

```
Heteroskedasticity (H):               2.26   Skew:                             0.41
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.85
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.19e+36. Standard errors may be unstable.
                    value
2023-01-01   68900.012384
2023-01-02   82218.811509
2023-01-03   80638.589496
2023-01-04   70086.334253
2023-01-05  158809.608069
...                   ...
2023-10-15   36803.143176
2023-10-16   45505.915022
2023-10-17   38601.978232
2023-10-18   50413.433741
2023-10-19   61657.064074

[292 rows x 1 columns]
------------Time series analysis for trip count-------------

The ADF test for original time series:
p-value: 0.120097
ADF Statistic: -2.481273
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

The ADF test for differenced time series (Difference order 1):
p-value: 0.000000
ADF Statistic: -8.471186
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568
```

```
------------------------SARIMA Model------------------------
Best SARIMA parameters: (1, 1, 1) (0, 1, 1, 12)
Summary of the SARIMA model for trip count:
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  0   No. Observations:                 1458
Model:             SARIMAX(1, 1, 1)x(0, 1, 1, 12)   Log Likelihood              -12833.318
Date:                            Thu, 23 Mar 2023   AIC                          25674.636
Time:                                    23:00:17   BIC                          25695.701
Sample:                                         0   HQIC                         25682.502
                                           - 1458                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3298      0.027     12.074      0.000       0.276       0.383
ma.L1         -0.8404      0.017    -48.953      0.000      -0.874      -0.807
ma.S.L12      -1.0013      0.023    -43.735      0.000      -1.046      -0.956
sigma2      3.515e+06   5.88e-09   5.98e+14      0.000    3.52e+06    3.52e+06
===================================================================================
Ljung-Box (L1) (Q):                   1.84   Jarque-Bera (JB):               650.49
Prob(Q):                              0.18   Prob(JB):                         0.00
Heteroskedasticity (H):               1.70   Skew:                            -0.68
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.01
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.02e+28. Standard errors may be unstable.
                  value
2023-01-01  5008.699552
2023-01-02  5616.997192
2023-01-03  5969.521223
2023-01-04  5656.188384
2023-01-05  7891.032908
...                 ...
2023-10-15  2039.647794
2023-10-16  2436.075279
2023-10-17  2525.499776
```

```
2023-10-18  3087.330308
2023-10-19  3738.956828

[292 rows x 1 columns]
------------Time series analysis for CASE_COUNT-------------

The ADF test for original time series:
p-value: 0.000001
ADF Statistic: -5.757126
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

The ADF test for differenced time series (Difference order 1):
p-value: 0.000000
ADF Statistic: -7.980677
Critical Values:
   1%: -3.435
   5%: -2.864
   10%: -2.568

------------------------SARIMA Model------------------------
Best SARIMA parameters: (1, 0, 1) (0, 1, 1, 12)
Summary of the SARIMA model for CASE_COUNT:
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  0   No. Observations:                 1458
Model:             SARIMAX(1, 0, 1)x(0, 1, 1, 12)   Log Likelihood              -17600.583
Date:                            Thu, 23 Mar 2023   AIC                          35209.166
Time:                                    23:05:45   BIC                          35230.234
Sample:                                         0   HQIC                         35217.033
                                           - 1458
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8147      0.009     93.517      0.000       0.798       0.832
ma.L1          0.1059      0.014      7.310      0.000       0.078       0.134
```

```
ma.S.L12      -0.9792      0.009   -114.098      0.000      -0.996      -0.962
sigma2      4.264e+09   2.42e-12   1.76e+21      0.000    4.26e+09    4.26e+09
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):          1061658.36
Prob(Q):                              0.98   Prob(JB):                        0.00
Heteroskedasticity (H):             124.37   Skew:                            6.66
Prob(H) (two-sided):                  0.00   Kurtosis:                      135.72
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number  9e+35. Standard errors may be unstable.
                   value
2023-01-01  49859.270164
2023-01-02  40171.939189
2023-01-03  53751.444243
2023-01-04  27949.994270
2023-01-05  21663.730137
...                  ...
2023-10-15  31786.404590
2023-10-16  50265.392538
2023-10-17  48991.601090
2023-10-18  55637.027006
2023-10-19  52011.661927

[292 rows x 1 columns]
```

Spatial analysis for NYC

```
Multiscale Geographically Weighted Regression (MGWR)
=====================
Parameters

Input Features     df_GWR_XYTableToPoint
Dependent Variable     trip_count
Model Type     CONTINUOUS
Explanatory Variables
start_station_id;covid_cases;tripduration_sum_mins_;tripduration_mean_mins_;usertype_member_count
Output Features     D:\Desktop\nyc_MGWR\nyc_mgwr.shp
Neighborhood Type     DISTANCE_BAND
```

```
Neighborhood Selection Method     GOLDEN_SEARCH
Minimum Number of Neighbors
Maximum Number of Neighbors
Distance Unit     KILOMETERS
Minimum Search Distance
Maximum Search Distance
Number of Neighbors Increment
Search Distance Increment
Number of Increments
Number of Neighbors
Distance Band
Number of Neighbors for Golden Search
Number of Neighbors for Manual Intervals
User Defined Number of Neighbors
Search Distance for Golden Search     start_station_id # #;covid_cases # #;tripduration_sum_mins_ #
#;tripduration_mean_mins_ # #;usertype_member_count # #
Search Distance for Manual Intervals
User Defined Search Distance
Prediction Locations     df_GWR_XYTableToPoint
Explanatory Variables to Match     start_station_id 'start station id';covid_cases
covid_cases;tripduration_sum_mins_ tripduration_sum(mins);tripduration_mean_mins_
tripduration_mean(mins);usertype_member_count usertype_member_count
Output Predicted Features     D:\Desktop\nyc_MGWR\nyc_predictions.shp
Robust Prediction     ROBUST
Local Weighting Scheme     GAUSSIAN
Output Neighborhood Table     D:\Desktop\bst_MGWR\neighborhood table.dbf
Coefficient Raster Workspace
Scale Data     SCALE_DATA
Coefficient Raster Layers
Output Layer Group     nyc_MGWR_Results
Elapsed Time: 51 minutes 55 seconds
=====================
Summary Statistics for Coefficients Estimates
Explanatory Variables  Mean   Standard Deviation Minimum    Median Maximum
Intercept  0.0070 0.0217 -0.0473    0.0064 0.0609
start station id   0.0076 0.0057 -0.0111    0.0074 0.0253
covid_cases    0.0232 0.0082 0.0058 0.0208 0.0376
tripduration_sum(mins) 0.3446 0.0191 0.2608 0.3465 0.3821
tripduration_mean(mins)    -0.0189    0.0165 -0.0477    -0.0154    0.0080
```

```
usertype_member_count  0.6704 0.0195 0.6216 0.6728 0.7152
=====================
Model Diagnostics
Statistic  GWR    MGWR
R-Squared  0.9993 0.9993
Adjusted R-Squared 0.9993 0.9993
AICc   -14061.4512    -13964.7686
Sigma-Squared  0.0007 0.0007
Sigma-Squared MLE  0.0007 0.0007
Effective Degrees of Freedom   3079.7120  3157.4702
Optimal GWR Bandwidth: 1.96 kilometers (Distance).
=====================
Summary of Explanatory Variables and Neighborhoods
Explanatory Variables  Bandwidth (% of Extent)a   Significant (% of Features)b
Intercept  1.96 (6.35)    1701 (52.89)
start station id   1.96 (6.35)    1392 (43.28)
covid_cases    3.08 (9.98)    3110 (96.70)
tripduration_sum(mins) 1.96 (6.35)    3216 (100.00)
tripduration_mean(mins)    2.67 (8.64)    1858 (57.77)
usertype_member_count  1.96 (6.35)    3216 (100.00)
Distance Unit: kilometers
a: This number in the parenthesis ranges from 0 to 100%, and can be interpreted as a local, regional, global scale
based on the geographical context from low to high.
b: In the parentheses, the percentage of features that have significant coefficients of an explanatory variable.
=====================
Optimal Bandwidths Search History
Iterations Intercept  start station id   covid_cases    tripduration_sum(mins) tripduration_mean(mins)
usertype_member_count  AICc
0  1.96   1.96   1.96   1.96   1.96   1.96   -14061.4512
1  1.96   1.96   1.96   1.96   1.97   1.96   -13013.2744
2  1.96   1.96   2.19   1.96   1.97   1.97   -13627.3486
3  1.96   1.96   2.66   1.96   1.96   1.96   -13725.5510
4  1.96   1.96   2.66   1.96   2.36   1.97   -13787.3912
5  1.96   1.96   2.94   1.96   2.48   1.97   -13817.5688
6  1.96   1.96   2.96   1.96   2.66   1.96   -13827.1561
7  1.96   1.96   2.96   1.96   2.66   1.96   -13835.3348
8  1.96   1.96   2.96   1.96   2.66   1.96   -13844.3766
9  1.96   1.96   3.1    1.96   2.66   1.96   -13858.7565
10 1.96   1.96   3.08   1.96   2.66   1.96   -13870.7500
```

```
11 1.96   1.96   3.08   1.96   2.66   1.96   -13882.0303
12 1.96   1.96   3.08   1.96   2.66   1.96   -13893.2153
13 1.96   1.96   3.08   1.96   2.66   1.96   -13903.6576
14 1.96   1.96   3.08   1.96   2.66   1.96   -13913.1190
15 1.96   1.96   3.08   1.96   2.66   1.96   -13922.1445
16 1.96   1.96   3.08   1.96   2.66   1.96   -13929.5200
17 1.96   1.96   3.08   1.96   2.66   1.96   -13935.9351
18 1.96   1.96   3.08   1.96   2.66   1.96   -13941.4922
19 1.96   1.96   3.08   1.96   2.66   1.96   -13946.2980
20 1.96   1.96   3.08   1.96   2.66   1.96   -13950.4589
21 1.96   1.96   3.08   1.96   2.66   1.96   -13954.0753
22 1.96   1.96   3.08   1.96   2.66   1.96   -13957.2382
23 1.96   1.96   3.08   1.96   2.67   1.96   -13960.0467
24 1.96   1.96   3.08   1.96   2.67   1.96   -13962.5424
25 1.96   1.96   3.08   1.96   2.67   1.96   -13964.7686
Distance Unit: kilometers
=====================
Bandwidth Statistics Summary
Explanatory Variables  Optimal Distance Bandwidth Effective Number of Parameters Adjusted Value of Alpha    Adjusted
Critical Value of Pseudo-t Statistics
Intercept  1.96   13.38  0.0037 2.9018
start station id   1.96   15.19  0.0033 2.9414
covid_cases    3.08   4.33   0.0116 2.5270
tripduration_sum(mins) 1.96   9.83   0.0051 2.8034
tripduration_mean(mins)    2.67   6.07   0.0082 2.6437
usertype_member_count  1.96   9.73   0.0051 2.8003
Distance Unit: kilometers
```

Spatial analysis for Boston

```
Multiscale Geographically Weighted Regression (MGWR)
=====================
Parameters

Input Features     df_bst_GWR_XYTableToPoint1
Dependent Variable     trip_count
Model Type     CONTINUOUS
Explanatory Variables
```

```
start_station_id;covid_cases;tripduration_sum_mins_;tripduration_mean_mins_;usertype_member_count
Output Features     D:\Desktop\bst_MGWR\BST_MGWR.shp
Neighborhood Type     DISTANCE_BAND
Neighborhood Selection Method     GOLDEN_SEARCH
Minimum Number of Neighbors
Maximum Number of Neighbors
Distance Unit     KILOMETERS
Minimum Search Distance
Maximum Search Distance
Number of Neighbors Increment
Search Distance Increment
Number of Increments
Number of Neighbors
Distance Band
Number of Neighbors for Golden Search
Number of Neighbors for Manual Intervals
User Defined Number of Neighbors
Search Distance for Golden Search     start_station_id # #;covid_cases # #;tripduration_sum_mins_ #
#;tripduration_mean_mins_ # #;usertype_member_count # #
Search Distance for Manual Intervals
User Defined Search Distance
Prediction Locations     df_bst_GWR_XYTableToPoint1
Explanatory Variables to Match     start_station_id 'start station id';covid_cases
covid_cases;tripduration_sum_mins_ tripduration_sum(mins);tripduration_mean_mins_
tripduration_mean(mins);usertype_member_count usertype_member_count
Output Predicted Features     D:\Desktop\bst_MGWR\predictions.shp
Robust Prediction     ROBUST
Local Weighting Scheme     GAUSSIAN
Output Neighborhood Table     D:\Desktop\bst_MGWR\neighborhood table.dbf
Coefficient Raster Workspace
Scale Data     SCALE_DATA
Coefficient Raster Layers
Output Layer Group     BST_MGWR_Results
=====================

Summary Statistics for Coefficients Estimates
Explanatory Variables  Mean   Standard Deviation Minimum    Median Maximum
Intercept  0.0000 0.0002 -0.0004    0.0000 0.0017
start station id   -0.0076    0.0000 -0.0078    -0.0076    -0.0074
```

covid_cases   0.0120 0.0001 0.0118 0.0120 0.0122
tripduration_sum(mins) 0.3669 0.0000 0.3668 0.3669 0.3669
tripduration_mean(mins)   -0.0263   0.0000 -0.0265   -0.0263   -0.0261
usertype_member_count 0.6503 0.0000 0.6503 0.6503 0.6503
=====================
Model Diagnostics
Statistic GWR   MGWR
R-Squared 0.9983 0.9983
Adjusted R-Squared 0.9982 0.9982
AICc   -1645.9879 -1646.1656
Sigma-Squared 0.0018 0.0018
Sigma-Squared MLE 0.0017 0.0017
Effective Degrees of Freedom   466.6068   466.8056
Optimal GWR Bandwidth: 32.15 kilometers (Distance).
=====================
Summary of Explanatory Variables and Neighborhoods
Explanatory Variables Bandwidth (% of Extent)a   Significant (% of Features)b
Intercept 29.40 (57.63) 0 (0.00)
start station id   39.35 (77.15) 473 (100.00)
covid_cases   39.35 (77.15) 473 (100.00)
tripduration_sum(mins) 51.01 (100.00) 473 (100.00)
tripduration_mean(mins)   39.35 (77.15) 473 (100.00)
usertype_member_count 51.01 (100.00) 473 (100.00)
Distance Unit: kilometers
a: This number in the parenthesis ranges from 0 to 100%, and can be interpreted as a local, regional, global scale based on the geographical context from low to high.
b: In the parentheses, the percentage of features that have significant coefficients of an explanatory variable.
=====================
Optimal Bandwidths Search History
Iterations Intercept start station id   covid_cases   tripduration_sum(mins) tripduration_mean(mins) usertype_member_count AICc
0 32.15 32.15 32.15 32.15 32.15 32.15 -1645.9879
1 32.15 32.15 39.35 51.01 39.35 51.01 -1646.1112
2 29.40 39.35 39.35 51.01 39.35 51.01 -1646.1656
Distance Unit: kilometers
=====================
Bandwidth Statistics Summary
Explanatory Variables Optimal Distance Bandwidth Effective Number of Parameters Adjusted Value of Alpha   Adjusted Critical Value of Pseudo-t Statistics

```
Intercept  29.40  1.06   0.0472 1.9896
start station id   39.35  1.03   0.0484 1.9788
covid_cases    39.35  1.04   0.0480 1.9828
tripduration_sum(mins) 51.01  1.01   0.0496 1.9689
tripduration_mean(mins)    39.35  1.04   0.0479 1.9838
usertype_member_count  51.01  1.01   0.0496 1.9682
Distance Unit: kilometers
Elapsed Time: 15.65 seconds
```