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.

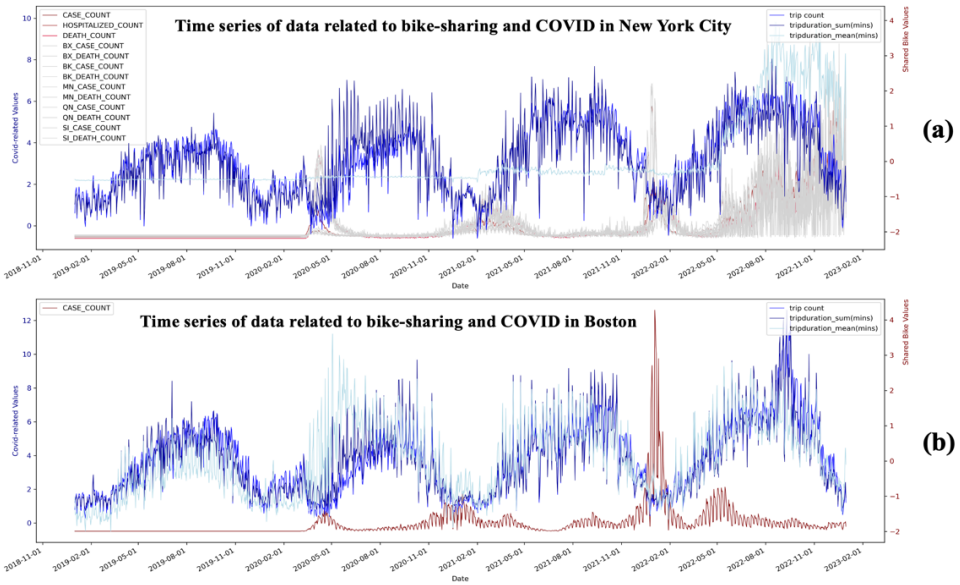

Figure 1. NYC and Boston datasets time series. (a for NYC, b for Boston City area)

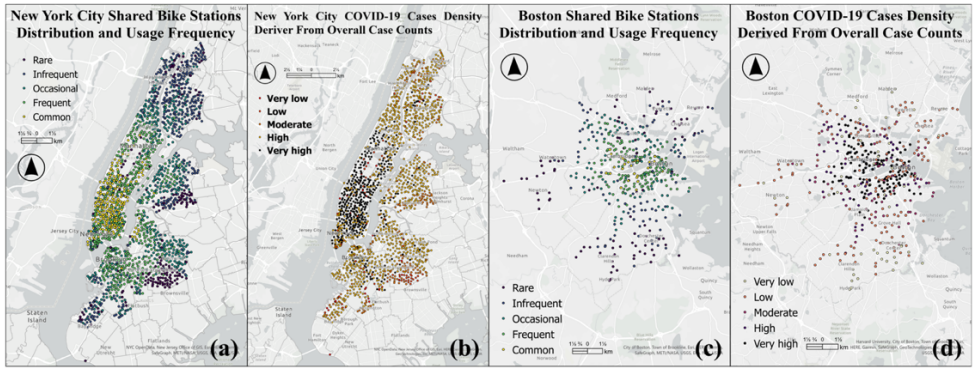

Figure 2. Distribution of NYC and Boston shared bike stations and corresponding usage frequency and COVID cases density. (a, b for NYC and c, d for Boston)

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.

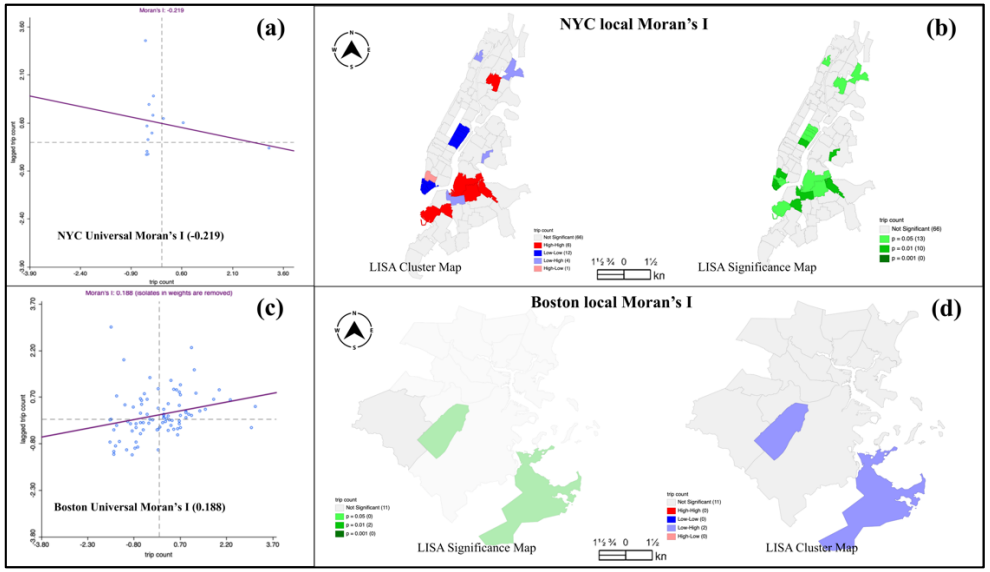

Figure 3. NYC and Boston spatial correlations (a is the global Moran's I scatter plot for NYC, b is the Local Indicators of Spatial Autocorrelation, (LISA) including cluster and significance maps for NYC; c is the global Moran's I scatter plot for Boston, d is the LISA cluster and significance maps for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

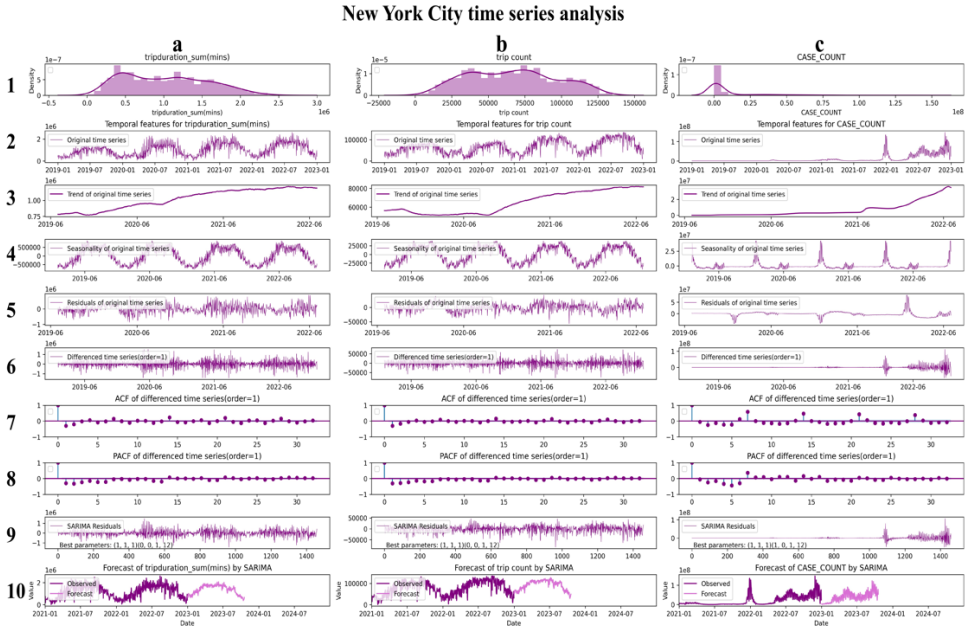

Figure 4. NYC temporal analysis (a, b, c for trip duration time (mins), trip counts and COVID cases respectively. Subplots are accessed by indexing in this report, e.g., the first subplot is referenced as *figure 4 (1a)),* the first line is referenced as *figure 4(1)*).

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.

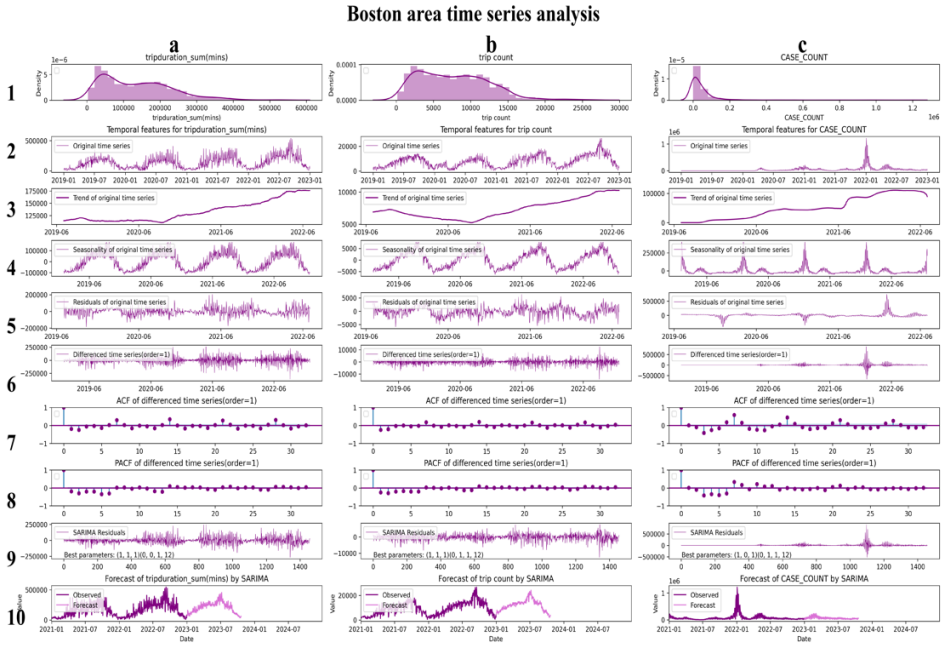

Figure 5. Boston area temporal analysis (a, b, c for trip duration time (mins), trip counts and COVID cases respectively. Subplots are accessed by indexing in this report, e.g., the first subplot is referenced as *figure 5(1a)),* the first line is referenced as *figure 5(1)*).

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.

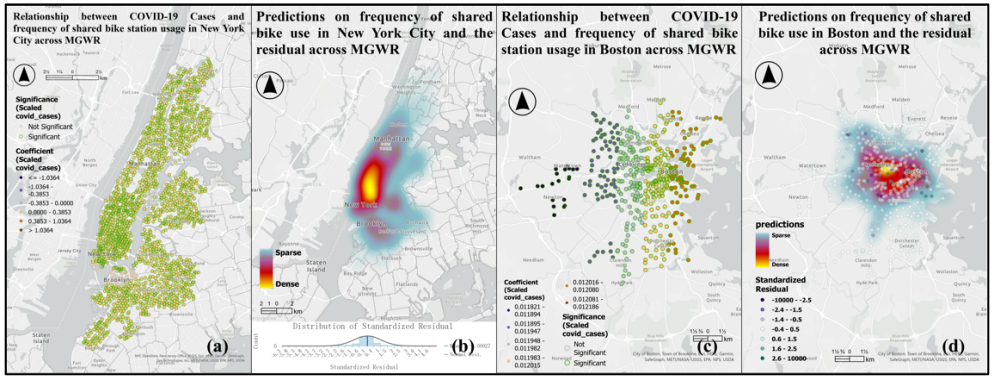

Figure 5. Relationship between shared bikes and COVID cases and predictions by MGWR in NYC and Boston

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]