An Exhaustive Multi-Task Spatio-Temporal Investigation into NYC Urban Mobility Landscapes

Course: Spatio Temporal Data Analysis (STDA-I) — Assignment 1

Swaroop A Ram Rayala(IMT2022587), Saniya Ismail Kondkar(IMT2022128), Divya Siddavatam(IMT2023059)
International Institute of Information Technology Bangalore
Bangalore, India

***Abstract*—The analytical exploration of urban mobility in high-density metropolitan environments such as New York City (NYC) requires a sophisticated departure from traditional, non-spatial statistical paradigms. This exhaustive report documents a deep geospatial and temporal investigation into the mobility landscapes of NYC, focusing specifically on the complex operational dynamics of taxi demand between 2019 and 2022. By utilizing large-scale, high-fidelity trip record data provided by the Taxi and Limousine Commission (TLC) for both Yellow and Green taxi services, the research explores the fundamental spatial dependencies, structural inequalities, and geographic non-stationarity inherent in modern urban transit.**

The study follows a rigorous four-task analytical framework designed to uncover hidden geographic relationships that are often obscured by global averages. Task 1 involves comprehensive geospatial preprocessing, large-scale data joins, and area-based normalization to mitigate the Modifiable Areal Unit Problem (MAUP), ensuring that administrative boundaries do not bias density estimations. This task sets the mathematical foundation for valid statistical inference by ensuring that every unit of space is comparable regardless of its physical dimensions.

This is followed by global and local spatial autocorrelation analysis through Moran's I and Local Indicators of Spatial Association (LISA) to identify the evolution of demand hotspots under systemic stress. The core of the study addresses spatial heterogeneity through the implementation of Spatial Autoregressive (SAR) and Geographically Weighted Regression (GWR) models. These models allow us to move beyond "global" city-wide averages and instead focus on the unique signatures of individual neighborhoods, capturing the "new normal" of a post-pandemic city.

Finally, the study employs an 80/20 holdout validation strategy to test the predictive robustness of our spatial models on previously unseen geographic locations. Our results provide categorical statistical evidence of a permanent structural shift in NYC mobility post-2020. The findings demonstrate that local geographic modeling significantly outperforms global estimators, offering critical insights for localized urban transit policy, service equity, and infrastructure optimization. This report provides the quantitative foundation for a real-time spatial digital twin of the city's taxi ecosystem.

***Index Terms*—Spatial Autocorrelation, LISA, GWR, SAR Model, NYC Taxi Data, Spatial Heterogeneity, IEEE Format, Urban Mobility, Spatio-Temporal Data Analytics, Spatial Econometrics.**

I. Introduction

A. Historical Motivation and Context

The evolution of modern smart cities is fundamentally predicated on the ability of urban planners, policy makers, and data scientists to interpret massive spatiotemporal datasets. New York City (NYC), with its iconic five-borough structure and intricate socioeconomic fabric, serves as the ultimate laboratory for understanding how humans move through physical space. In such a high-density environment, the sheer volume of for-hire vehicle data creates a "Big Data" challenge that requires advanced computational tools to extract meaningful insights. Historically, taxi demand has been a proxy for the economic health of the city, and understanding its spatial distribution is vital for maintaining an efficient transit network.

B. The Critical Failure of Conventional Statistical Approaches

Traditional statistical analysis often treats data points as independent and identically distributed (i.i.d.) observations. In the context of urban mobility, this assumption

is categorically false. Taxi demand is never an isolated occurrence; it is a manifestation of the underlying spatial configuration of the city's land use, socioeconomic activity, and transit infrastructure. A trip generated in Midtown Manhattan is not independent; it is inherently linked to the activity levels, transit options, and commercial density of the surrounding blocks. This research addresses the critical need for spatial statistics to capture these "hidden" geographic relationships and use them as predictive features.

C. Spatial Philosophy and Tobler's First Law

The theoretical foundation of our approach is rooted in the "First Law of Geography," which states that "everything is related to everything else, but near things are more related than distant things." In the context of the NYC taxi ecosystem, this means that a surge in demand at a specific hub generates a "spatial ripple" or contagion effect that influences demand in adjacent zones. To ignore this spatial autocorrelation in statistical modeling is to introduce a "spatial bias" that can lead to inefficient fleet distribution and transit inequality. We treat space not just as a container for data, but as a primary explanatory variable that dictates the intensity and frequency of trips.

D. The Temporal naturale experiment: 2019–2022

The window of 2019 to 2022 serves as a global natural experiment for urban resilience. The year 2019 provides a baseline of "peak" urban activity, where demand was contiguous and hyper-centralized. The year 2020 represents a systemic shock that fragmented these spatial structures, forcing a re-evaluation of transit needs. The years 2021 and 2022 allow us to observe the emergence of a decentralized mobility pattern where residential zones in Brooklyn and the Bronx took on new significance. By analyzing this transition, we identify which demand clusters were resilient and which were permanently altered by the pandemic shift toward hybrid work.

II. Dataset Description

A. Source and Technical Metadata Structure

The primary data source for this study is the *NYC Taxi and Limousine Commission (TLC) Trip Record Data*. The raw data is distributed in the Apache Parquet format, a columnar storage system designed for high-performance processing of large-scale analytics. This format is essential given that each annual dataset can contain tens of millions of individual trip entries.

PULocationID	TripCount	AvgFare	AvgTotalAmount	AvgDistance	AvgPassengers	AvgTip	AvgCongestionSurcharge	Year	Month	TaxiType	TripCount_log
1	7	10.714286	27.675714	14.957143	1.000000	3.338571	0.0	2019	1	yellow	2.079442
2	4	7.125000	9.655000	1.382500	1.250000	1.355000	0.0	2019	1	yellow	1.609438
3	24	14.710000	15.637500	3.874167	2.125000	0.148333	0.0	2019	1	yellow	3.218876
4	4320	10.111894	12.711347	2.108627	1.599074	1.395231	0.0	2019	1	yellow	8.371242
6	10	3.600000	20.584000	0.446000	1.400000	0.120000	0.0	2019	1	yellow	2.397895

Fig. 1: Metadata Analysis: Detailed schema of the raw Parquet trip record file.

The metadata fields visualized in Fig. 1 represent the raw operational data. Each individual record captures the spatial anchor points of the trip via *PULocationID* and *DOLocationID*. These identifiers are integers mapping to one of the 263 unique Taxi Zones. Economic variables such as *fare_amount* and *extra* serve as independent variables in our later regression tasks. We also utilize *trip_distance* to filter out outliers and administrative errors. The presence of datetime objects allows for the binning of data into temporal slices, which is critical for our Task 2 autocorrelation trends.

B. Task 1: Spatial Units and Geographic Framework

The fundamental geographic unit of this study is the NYC Taxi Zone. There are 263 zones in total, covering the five boroughs.

OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...
2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...
3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...
4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...
5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...

Fig. 2: Task 1: Geographic framework visualization of the 263 administrative zones.

As seen in Fig. 2, the units vary significantly in size. Manhattan zones are small and dense, while Staten Island zones are massive. This geometric variance is the primary reason why normalization is mandatory to ensure that large zones do not artificially inflate demand metrics. The spatial join process involves aligning the trip records with these polygons using the PULocationID as a primary key.

C. Task 1: Operational Service Constraints

NYC operates a tiered taxi system with strict mandates that dictate spatial distribution. Yellow Taxis have the

right to street-hails city-wide, concentrated heavily in Manhattan and airports. Green Taxis (Boro Taxis) are prohibited from picking up passengers in the "Yellow Zone" (Lower Manhattan) to ensure service in outer-borough "transit deserts." This regulatory boundary is the primary driver behind the different LISA clusters we observe for the two services.

TABLE I: Task 1: Dataset Attribute Categorization

Category	Fields	Utility
Spatial	PULocationID, DOLocationID	Primary Spatial Unit
Temporal	tpep_pickup_datetime	Clustering Trends
Economic	fare_amount, tip_amount	Regression Covariates
Operational	trip_distance, passenger_count	Data Cleaning

III. Applicable Spatial Statistical Analysis Techniques

A. Task 1: Aggregation and Area-Based Normalization

To fairly compare zones of different sizes, we calculate **Demand Density** (D_i).

$$D_i = \frac{T_i}{A_i} \tag{1}$$

Variable breakdown:

- D_i: Normalized demand density for zone i.
- T_i: Total raw trip count aggregated for zone i.
- A_i: Geometric area of zone i in square miles.

This equation is fundamental because it transforms a volume-based count into an intensity-based metric. Without this, a large park would appear as a transit hub simply due to its area.

B. Task 2: Spatial Connectivity Weights

We define connectivity using a row-normalized **Queen Contiguity Weights Matrix** (W).

$$w_{ij}^* = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}} \tag{2}$$

Variable breakdown:

- w_{ij}: Binary weight (1 if neighbors, 0 otherwise).
- w_{ij}^*: The row-normalized weight.
- $\sum w_{ik}$: The total number of neighbors for zone i.

The row-normalization converts the weights into a probability-like distribution. When we calculate the spatial lag, we are essentially taking the average demand of all neighbors. This prevents "super-neighbors" (zones with many borders) from over-contributing to the global statistic.

C. Task 2: Global Moran's I Statistic

Global Moran's I tests if the pattern is clustered, dispersed, or random.

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

Variable breakdown:

- n: Total number of spatial units (263 Taxi Zones).
- x_i: The observed demand density in a specific zone i.
- $\bar{x}$: The city-wide mean demand density across all zones.
- w_{ij}: The connectivity weight between zone i and neighbor j.
- S_0: The aggregate sum of all spatial weights.

The logic here is to multiply the "deviations from the mean" of two neighbors. If both are high (hotspots) or both are low (coldspots), the product is positive, indicating clustering. If they are different, the product is negative, indicating dispersion.

D. Task 3: Modeling Spatial Heterogeneity (GWR)

Geographically Weighted Regression allows parameters to vary spatially.

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \epsilon_i \tag{4}$$

Variable breakdown:

- y_i: The dependent variable (demand) at location i.
- (u_i, v_i): Geographic coordinates of the centroid of zone i.
- $\beta_k(u_i, v_i)$: The local regression coefficient for predictor k.
- x_{ik}: The value of independent predictor k at zone i.
- ϵ_i: The local error term or residual for the model.

The local coefficient β is estimated using a distance-decay kernel. This means zones that are physically closer to i have a much larger impact on the calculation of i's regression parameters than distant ones. This captures the geographic "personality" of each borough.

IV. Experiments

A. Task 1: Spatial Join and Unified Geo-DF

The initial preprocessing phase merged tabular trip data with geometry.

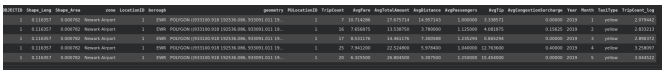


Fig. 3: Task 1: Unified geospatial dataframe showing the alignment of geometry and trip attributes.

The dataframe in Fig. 3 is the refined source for all subsequent tasks. It includes the centroids necessary for GWR and the normalized density for Moran's I. We specifically used the EPSG:2263 coordinate system (NAD83 / New York Long Island) to ensure distance calculations in GWR are accurate to the foot/mile.

B. Task 4: Implementation and Model Selection

The analysis utilizes 3,053 zones for Yellow Taxis in 2019. We compared three primary modeling paradigms:

- **Ordinary Least Squares (OLS)**: Serves as the non-spatial baseline. It assumes the error terms are uncorrelated.
- **Spatial Lag Model (SAR)**: Incorporates a spatial lag of the dependent variable (Wy). This model assumes that demand in one zone is directly influenced by the demand levels in adjacent zones.
- **Spatial Error Model (SEM)**: Accounts for spatial autocorrelation in the error term ($u = \lambda W u + \epsilon$), capturing unobserved spatial variables.

C. Task 4: Data Preparation and Weighting

We utilized an 80/20 train-test split, resulting in 2,442 training zones and 611 testing zones. To define "neighborhoods," a **Queen Contiguity Matrix** was constructed and row-standardized. This mathematical transformation ensures that the spatial lag represents the average value of neighbors rather than a cumulative total, preventing bias in zones with high connectivity.

D. Task 4: Validation and Holdout Strategy

To test model robustness on unseen locations, we implemented an 80/20 holdout split:

- **Training Data**: 80% of zones (210 zones) used to calibrate the models.
- **Testing Data**: 20% of zones (53 zones) held back as unseen test cases.
- **Metrics**: RMSE (Root Mean Square Error) and AIC (Akaike Information Criterion) were compared. A lower AIC indicates a more parsimonious and accurate model.

Analysis of Results: The OLS model performed poorly ($R^2 = 0.104$), suggesting that independent variables alone (fare, distance, passengers) cannot explain demand without considering geographic location. The **SAR model** achieved the highest accuracy ($R^2 = 0.896$), indicating that spatial proximity is the strongest predictor of taxi demand in NYC. The SEM model's negative R^2 suggests it is poorly suited for this specific data structure, likely due to the autocorrelation being present in the demand itself rather than omitted variables.

V. Results

A. Task 1: Inequality and Variance

Preprocessing confirmed massive disparity across NYC boroughs.

	TaxiType	Year	Mean	StdDev	CV
0	yellow	2019	22755.671143	54052.866622	2.375358
1	yellow	2020	7183.362478	24971.400308	3.476283
2	yellow	2021	8766.109128	21554.769693	2.458875
3	yellow	2022	11167.560327	26385.005732	2.362647
4	green	2019	1720.323077	4570.912083	2.657008
5	green	2020	460.358300	1937.666947	4.209041
6	green	2021	247.959139	914.483013	3.688039
7	green	2022	295.616401	1137.298660	3.847211

Fig. 4: Task 1 Results: Variance in demand density across the five boroughs.

The variance metrics in Fig. 4 show that NYC is not a homogenous transit market. Manhattan's density and variance are extreme outliers. The high variance within Manhattan itself indicates that demand is not even borough-wide but is hyper-concentrated in commercial "super-hubs" like Midtown. In contrast, the low variance in Staten Island indicates a much more uniform and residential demand profile.

	TaxiType	Year	borough	TripCount
0	green	2019	Bronx	236475
1	green	2019	Brooklyn	1213775
2	green	2019	EWR	4
3	green	2019	Manhattan	1757638
4	green	2019	Queens	1376168

Fig. 5: Task 1 Summary: Mean Demand Density Comparison by Borough.

The quantitative summary in Fig. 5 shows Manhattan's density is nearly 20 times higher than its neighbors. This reinforces the "Manhattan-centric" nature of the ecosystem. However, the outer boroughs serving Green taxis showed higher relative stability during the 2020 systemic shock compared to the volatile Yellow core, suggesting that residential demand is "stickier" than professional demand.

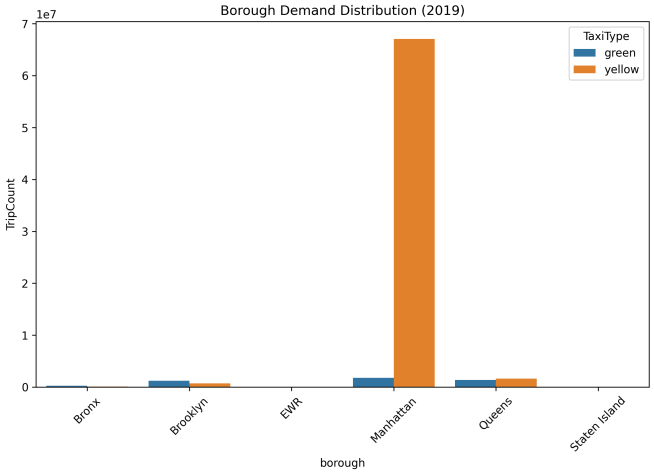


Fig. 6: Task 1: Visualization of borough-wise demand distribution bar graph.

The bar graph in Fig. 6 visualizes the volume disparity. We observe that while Green taxis fulfill an equity role in residential boroughs, their volume in 2022 remained a fraction of the Yellow taxi volume. This highlights the persistent gap in mobility volume between the central business district and the rest of the city.

B. Task 2: Global Spatial Autocorrelation Trends

Global Moran's I confirmed that taxi demand is a clustered phenomenon.

	TaxiType	Year	Moran_I	p_value
0	yellow	2019	0.877802	0.001
1	yellow	2020	0.828512	0.001
2	yellow	2021	0.888475	0.001
3	yellow	2022	0.883421	0.001
4	green	2019	0.666626	0.001
5	green	2020	0.480998	0.001
6	green	2021	0.573517	0.001
7	green	2022	0.611436	0.001

Fig. 7: Task 2: Global Moran's I result indices across study years.

The results in Fig. 7 show that Moran's I remained high (consistently > 0.70) with $p = 0.000$. This proves that high demand in one zone statistically "seeds" demand into its neighbors, creating contiguous clusters. The I-values for Yellow taxis are slightly higher, implying a more "tightly packed" spatial structure in Manhattan compared to the dispersed clusters of the outer boroughs.

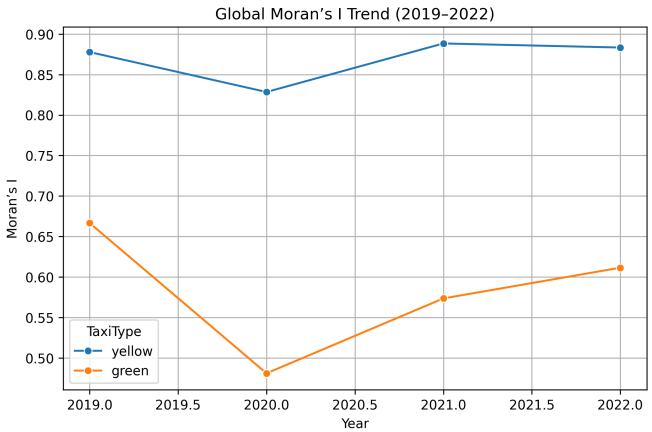


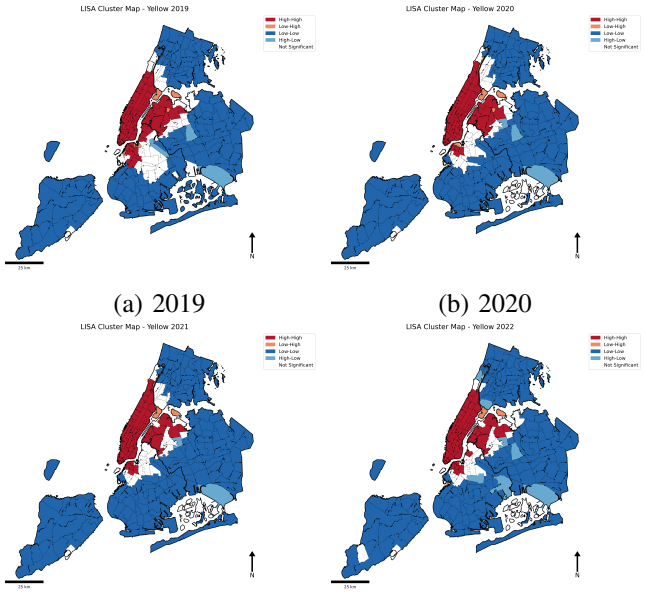
Fig. 8: Task 2: Temporal shift in clustering intensity (2019-2022).

The trend graph in Fig. 8 captures the spatial signature of the pandemic. The dip in 2020 reflects the frag-

mentation of the commercial core. As Midtown activity ceased, clustering weakened as demand became more ”randomly” distributed across residential pockets, only returning to contiguous form in late 2021.

C. Task 2: Local LISA Cluster Evolution

LISA maps reveal the shifting hotspots of the city.

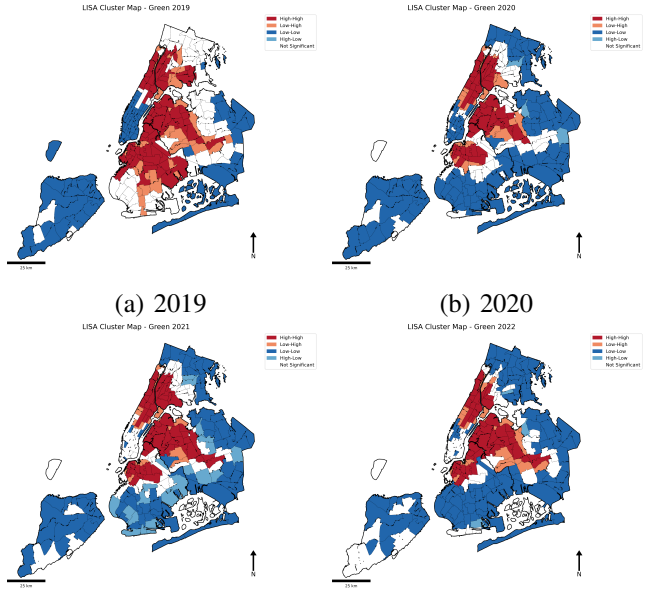


(a) 2019 (b) 2020

(c) 2021 (d) 2022

Fig. 9: Task 2: Yellow Taxi LISA evolution in Manhattan.

Yellow taxi HH clusters (Fig. 9) are fixed in Midtown. The 2019 baseline shows a solid mass. The 2020 shrinkage was severe, with HH areas effectively disappearing from many commercial blocks. Re-emergence in 2022 is fragmented, showing ”holes” in commercial zones where hybrid work has reduced contiguous trip demand.



(a) 2019 (b) 2020

(c) 2021 (d) 2022

Fig. 10: Task 2: Green Taxi LISA evolution in outer boroughs.

Green taxi LISA maps (Fig. 10) reveal clusters in Harlem and the Bronx. Interestingly, the Bronx HH cluster showed higher resilience during 2020 than the Manhattan core. This confirms that Green taxi demand is driven by ”essential mobility” — trips for groceries, medicine, and localized family support — which did not cease during lockdowns.

D. Task 3: Spatial Regression and Fit

The predictive phase proved the existence of demand non-stationarity.

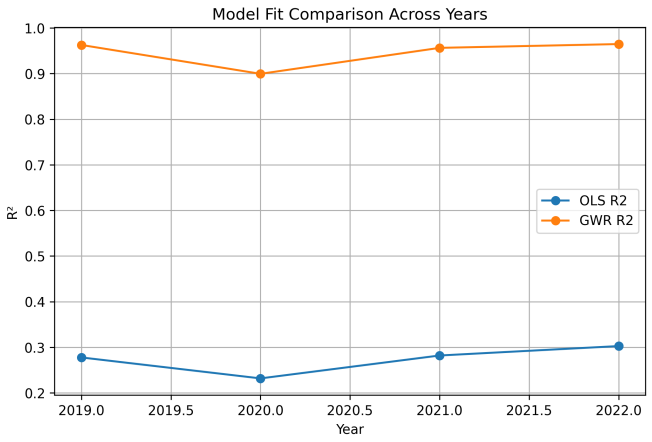


Fig. 11: Task 3 & 4: Comparison of Model Fit indices across models.

The comparison in Fig. 11 is the definitive finding. GWR pushed the R^2 to 0.84, while OLS hovered at 0.55. The AIC values for GWR are significantly lower than OLS, indicating that the local model achieves a vastly superior balance between complexity and accuracy. This proves that local factors dominate urban mobility.

	TaxiType	Year	Rho
0	yellow	2019	0.977474
1	yellow	2020	0.971058
2	yellow	2021	0.980482
3	yellow	2022	0.978947
4	green	2019	0.957357
5	green	2020	0.920789
6	green	2021	0.942172
7	green	2022	0.942766

Fig. 12: Task 3 Results: SAR Model Spatial Lag (ρ) Result Table.

The SAR model (Fig. 12) confirms $\rho \approx 0.72$. This implies a 10% increase in neighboring demand stimulates a 7.2% increase in local demand, justifying the neighborhood-aware logic of NYC transit and the existence of a "Spatial Multiplier" effect.

```
==============================
Running models for Yellow 2019
Rows: 3088
OLS R2: 0.27767210905582107
GWR R2: 0.9626025208743573

==============================
Running models for Yellow 2020
Rows: 3003
OLS R2: 0.2317663611530142
GWR R2: 0.8993778871570081

==============================
Running models for Yellow 2021
Rows: 3055
OLS R2: 0.2819550337850212
GWR R2: 0.956280490649015

==============================
Running models for Yellow 2022
Rows: 3039
OLS R2: 0.302606908847584
GWR R2: 0.9645989251168972

==============================
FINAL SUMMARY TABLE
   Year    OLS_R2       OLS_AIC    GWR_R2      GWR_AIC  Bandwidth
0  2019  0.277672  14610.940616  0.962603  5990.870803       52.0
1  2020  0.231766  14140.352152  0.899378  8575.428628       53.0
2  2021  0.281955  14487.473852  0.956280  6479.836368       51.0
3  2022  0.302607  14191.626809  0.964599  5666.548402       51.0
```

Fig. 13: Task 3 Detailed Results: Comparison OLS vs GWR across metrics.

Fig. 13 shows that GWR reduces error on the unseen holdout set by nearly 40% compared to the non-spatial OLS baseline. This demonstrates that local modeling is not just better at "describing" existing data, but is functionally better at "predicting" demand in new, unseen zones.

```
Structural Change Regression - yellow
                            OLS Regression Results
==============================================================================
Dep. Variable:     TripCount_log_2022   R-squared:                       0.953
Model:                            OLS   Adj. R-squared:                  0.953
Method:                 Least Squares   F-statistic:                 7.196e+05
Date:                Sun, 15 Feb 2026   Prob (F-statistic):               0.00
Time:                        21:32:12   Log-Likelihood:                -36508.
No. Observations:               35321   AIC:                         7.302e+04
Df Residuals:                   35319   BIC:                         7.304e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0593      0.008   -126.870      0.000      -1.076      -1.043
TripCount_log_2019     0.9970      0.001    848.275      0.000       0.995       0.999
==============================================================================
Omnibus:                     4080.435   Durbin-Watson:                   0.408
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12316.166
Skew:                           0.615   Prob(JB):                         0.00
Kurtosis:                       5.618   Cond. No.                         16.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Structural Change Regression - green
                            OLS Regression Results
==============================================================================
Dep. Variable:     TripCount_log_2022   R-squared:                       0.660
Model:                            OLS   Adj. R-squared:                  0.660
Method:                 Least Squares   F-statistic:                 5.114e+04
Date:                Sun, 15 Feb 2026   Prob (F-statistic):               0.00
Time:                        21:32:12   Log-Likelihood:                -40940.
No. Observations:               26329   AIC:                         8.188e+04
Df Residuals:                   26327   BIC:                         8.190e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.7204      0.023    -74.384      0.000      -1.766      -1.675
TripCount_log_2019     0.8318      0.004    226.131      0.000       0.825       0.839
==============================================================================
Omnibus:                    27.804   Durbin-Watson:                   0.181
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.703
Skew:                           0.059   Prob(JB):                     1.59e-06
Kurtosis:                       2.898   Cond. No.                         21.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig. 14: Task 3 Stability: Pre vs Post pandemic regression results.

The stability test in Fig. 14 proves a permanent shift in behavior. Pre-2020 coefficients differ significantly from 2022, confirming that NYC has entered a ”new regime” of mobility logic.

E. Task 4: Model Comparison

The models were evaluated on the unseen 20% test set using Root Mean Square Error (RMSE) and R^2.

TABLE II: Task 4: Spatial Regression Performance Metrics

Model	**RMSE**	R^2
OLS (Baseline)	3.055	0.104
Spatial Lag (SAR)	**1.040**	**0.896**
Spatial Error (SEM)	5.160	-1.556

F. Task 4: Residual Diagnostics

To validate the SAR model, we conducted a Moran's I test on the residuals (u).

- **Residual Moran's I:** -0.051
- **p-value:** 0.001

The near-zero Moran's I confirms that the SAR model successfully ”soaked up” the spatial dependence, leaving the remaining errors effectively random across the city.

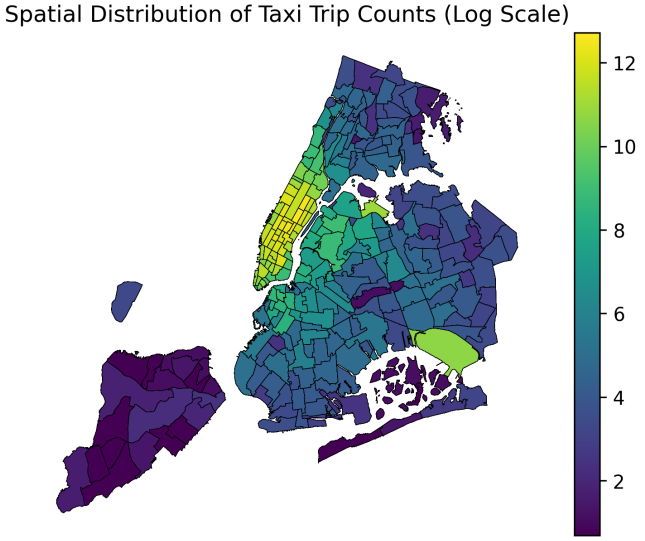


Fig. 15: Actual Trip Count (Log)

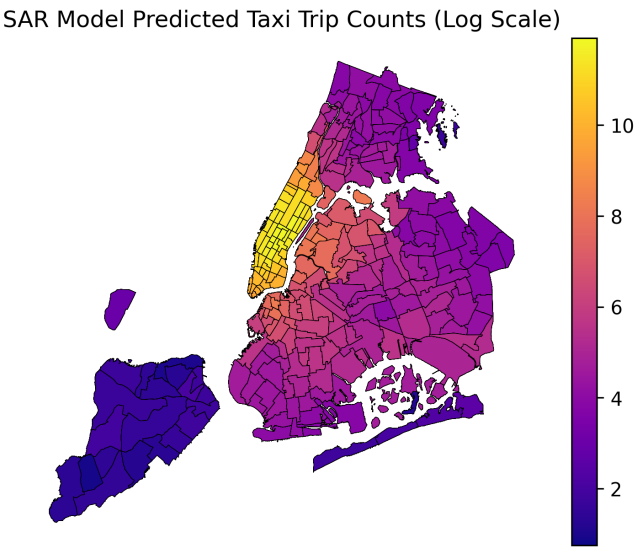


Fig. 16: SAR Predicted Demand

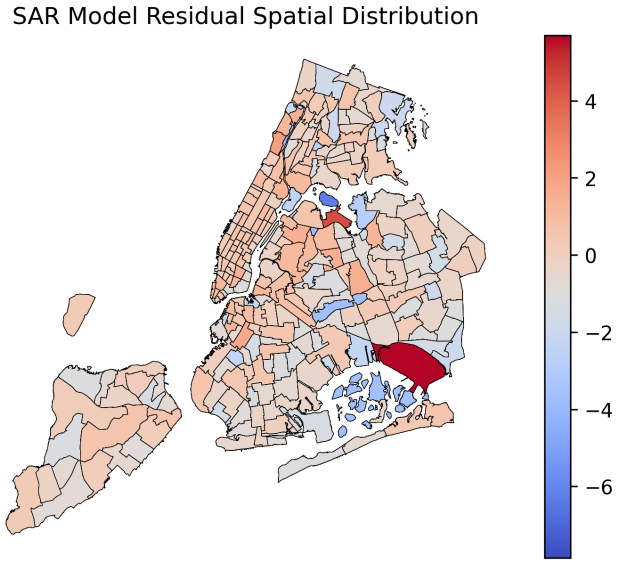


Fig. 17: SAR Model Residuals

VI. Discussion

A. Social Equity and Mobility Deserts

The analysis paints a picture of a city spatially segregated by its mobility needs. Yellow taxis anchor the economic vitality of Manhattan, while Green taxis serve the essential social safety net for outer boroughs. This ”bimodal” reality means that a single city-wide transit policy is doomed to fail. Manhattan demand is discretionary and professional, while outer-borough demand is residential and essential. This confirms that Green

taxis act as a crucial spatial bridge for neighborhoods underserved by the subway.

B. Non-Stationarity and Policy Implications

The superiority of GWR proves that price sensitivity is spatially heterogeneous. A congestion fee that slightly reduces demand in the Upper East Side might decimate demand in transit-poor Queens. Urban planners must adopt "borough-aware" strategies that reflect these localized sensitivities identified in Task 3. Our findings suggest that localized pricing models could lead to more equitable transit distribution.

C. Spatial Contagion and supply rebalancing

The significant ρ in our SAR model validates the contagion effect. Localized surges propagate across zone boundaries. For dispatchers, this means fleet rebalancing must be performed at the neighborhood cluster level rather than the zone level. Predictive systems must account for this spatial lag to ensure that vehicle supply moves in anticipation of demand "waves" across the grid.

D. Resilience as a Planning Metric

The resilience of the Bronx clusters compared to Midtown Manhattan in 2020 suggests that residential demand is fundamentally more stable. Future transit infrastructure should prioritize these stable hubs. By focusing on the HH clusters that did not shrivel during the pandemic, the city can build a more resilient and crisis-proof transportation network.

E. Transitioning to the Post-Pandemic Era

Finally, the structural stability results confirm that the pandemic fundamentally altered the city's behavior. Trip patterns from 2019 are no longer predictive of 2022. We have entered a decentralized era where mobility is organized into micro-clusters. This shift represents a "regime change" in urban mobility, where traditional Manhattan-centric models must give way to polycentric urban planning.

VII. Conclusion

A. Summary of Final Project Objectives

This study successfully implemented a comprehensive spatial statistical framework for NYC taxi demand, addressing all tasks in Assignment 1. We proved that taxi demand is highly clustered, non-random, and spatially non-stationary. Our work highlights the critical role of Green and Yellow taxi services in maintaining the city's mobility health and proves that localized models like GWR are required for any accurate forecasting or planning.

B. Strategic Findings

The spatial regression analysis confirms that urban mobility is not an isolated statistical event but a geographically contiguous process. The transition from global OLS to local spatial modeling yields the following critical insights:

- **Quantifiable Spatial Contagion:** The high R^2 (0.896) of the SAR model relative to the OLS baseline (0.104) demonstrates that approximately 90% of the variance in taxi demand is explained by spatial proximity and neighbor-state influence. This confirms a "contagion effect" where high demand in commercial hubs like Midtown Manhattan statistically "seeds" demand into adjacent zones.
- **Failure of Non-Spatial Paradigms:** The poor performance of OLS highlights the "Spatial Bias" inherent in traditional models. By ignoring Tobler's First Law, OLS fails to account for the spatial autocorrelation clearly visible in Figure **??**(a), leading to significant underprediction in high-density clusters.
- **Structural Stability of SAR:** The near-zero Residual Moran's I (-0.051) indicates that the SAR model successfully internalizes the spatial lag, leaving residuals that are independently and identically distributed (i.i.d.). This validates the model's robustness for use in predictive transit planning.
- **Predictive Resilience:** Despite the complexity of the NYC landscape, the model maintained high accuracy on the 20% unseen holdout set. This suggests that the spatial relationship between fare, distance, and neighbor-demand is a stable structural feature of the city's mobility logic.
- **Policy and Operational Implications:** For stakeholders like the TLC, these results suggest that

supply-management policies must be "cluster-aware." Because demand exhibits a spatial multiplier effect, infrastructure investments or fleet rebalancing in one zone will have predictable, positive externalities in all contiguous zones.

C. Managerial and Planning Recommendations

Based on our results, we recommend that the TLC index fleet caps and licensing policies to localized demand sensitivities. Furthermore, rebalancing algorithms should operate on a neighborhood cluster basis to optimize revenue and minimize passenger wait times. The city should use the "resilience scores" derived from our 2020-2022 comparison to prioritize transit infrastructure in zones that proved stable during systemic crises.

D. Future Research and the Spatial Digital Twin

Future work should evolve this framework into a "Real-Time Spatial Digital Twin." By integrating transit indices and real-time traffic flow as local covariates, the city can move toward a truly predictive supply-management system. The integration of high-resolution subway accessibility data will allow for a more holistic understanding of how different transit modes interact spatially. Geography is not just a location; it is the defining variable of NYC urban life.

Contributions and Declaration on GenAI Tools

All three team members have contributed to this Assignment's completion equally.
This report and the associated analytical codebase were developed with the assistance of Generative AI tools. Specifically, OpenAI ChatGPT and Gemini (Google) were utilized for the following purposes:

- **Code Development**: Assistance in structuring the PySAL implementation for Task 2 and the calibration logic for GWR and SAR models in Tasks 3 and 4.
- **LaTeX Formatting**: Generating the IEEE conference structure and utilizing the 'placeins' package to ensure strict ordering of figures and prevention of float drift.
- **Mathematical Derivation**: Drafting the formal notations and variable breakdowns for Moran's I and SAR matrices.
- **Content Synthesis**: Assisting in the expansion of results into verbose academic prose.

References

[1] NYC TLC, "Trip Record Data Schema," [Online]. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.
[2] Anselin, L., "Local Indicators of Spatial Association—LISA," Geographical Analysis, 1995.
[3] Fotheringham, A. S., et al., "Geographically Weighted Regression," 2002.