# Generating Synthetic Geospatial Trip Data for Urban Analytics

Predicting rideshare demand in Bandra, Mumbai

## Introduction: Synthetic Geospatial Data for Urban Analytics

Synthetic geospatial data has emerged as a transformative tool in urban analytics, particularly for contexts where granular, real-world data are unavailable or sensitive. Synthetic data replicates the statistical characteristics of real datasets while preserving privacy and avoiding reidentification risks. This approach has proven invaluable for applications in mobility analysis, infrastructure planning, and policy evaluation, offering a sandbox for "what-if" scenarios that are not feasible with real-world data alone.

In the context of cities in the Global South, such as Mumbai, the scarcity of detailed mobility data poses significant challenges for urban planning. Synthetic data generation provides a solution by simulating realistic, fine-grained urban behaviors. For instance, generating trip data for areas like Bandra in Mumbai requires adapting synthetic data techniques to reflect local travel patterns, socio-economic diversity, and infrastructural constraints. This entails the use of agent-based models, machine learning, and statistical simulations to produce data that align with real-world dynamics without relying on sensitive or unavailable datasets.

In this project, three scripts have been developed to generate synthetic trip data for Bandra. The first script constructs a synthetic population representative of Bandra's demographic and socio-economic attributes. The second simulates mobility patterns using probabilistic models that incorporate land use, transit availability, and behavioral assumptions. Finally, the third script evaluates and calibrates the synthetic data by comparing aggregated metrics against limited available real-world data. Together, these scripts demonstrate how synthetic data can be leveraged to overcome data gaps and inform evidence-based urban planning in resource-constrained environments.

## Methodology

This project leverages **Points of Interest (POI)** data from OpenStreetMap and spatial analysis tools such as QGIS to predict rideshare demand. Key POIs include bus stops, educational institutions, recreational areas, shops, and railway stations, forming the basis for identifying relevant scopes in the study area. By applying **Kernel Density Estimation (KDE)** and **spatial autocorrelation**, the synthetic data generation mimics real-world spatial dynamics, ensuring the results align with observed urban mobility patterns.

The three scripts sequentially perform a comprehensive spatial data analysis and modeling workflow for a city district. Here's a summary of each script:

### Script 1: Spatial Data Preparation and Synthetic Point Generation

1. **Data Loading and Preprocessing:**
    a. Loads polygon and point shapefiles representing a city district and various amenities (e.g., bus stops, education institutions).
    b. Cleans and reprojects all spatial layers to a common CRS (UTM Zone 43N).
2. **Spatial Analysis:**
    a. Combines influence points into a single dataset and creates a Kernel Density Estimation (KDE) surface to represent spatial intensity.
    b. Generates random points within the district, weights them using KDE and proximity to key amenities, and samples points to form a clustered spatial pattern.
3. **Quality Assurance:**
    a. Ensures valid geometries and point types.
    b. Introduces spatial autocorrelation testing (Moran's I) to evaluate clustering.
4. **Output:**
    a. Saves the processed points as a shapefile for further use.

### Script 2: Synthetic Trip Generation

1. **KDE-Based Weighting:**
    a. Uses KDE values calculated from points in Script 1 to weight trip origin and destination selection.

2. **Synthetic Trip Simulation:**
    a. Simulates trips by randomly pairing start and end points while avoiding self-loops.
    b. Generates random travel times and timestamps within a day.
3. **Spatial Autocorrelation Analysis:**
    a. Assesses spatial clustering of start and end points using Moran's I.
4. **Output:**
    a. Creates a synthetic trip dataset as an sf object and saves it to a CSV file for modeling.

## Script 3: Spatial Modeling and Predictive Analysis

1. **Spatial Join and Demand Aggregation:**
    a. Maps synthetic trips to a hexagonal grid overlaying the district.
    b. Extracts temporal features and aggregates trip demand by grid cell.
2. **Spatial Econometric Models:**
    a. Fits Spatial Lag Model (SLM) and Spatial Durbin Model (SDM) to analyze and predict rideshare demand.
    b. Incorporates spatial lags of demand and population density.
3. **Predictive Modeling:**
    a. Trains a Random Forest model to predict demand using temporal and spatial features.
    b. Evaluates model performance using RMSE.
4. **Visualization and Output:**
    a. Visualizes predicted demand using a hexagonal grid map.
    b. Saves the plot and residual validation metrics for further analysis.

## Results

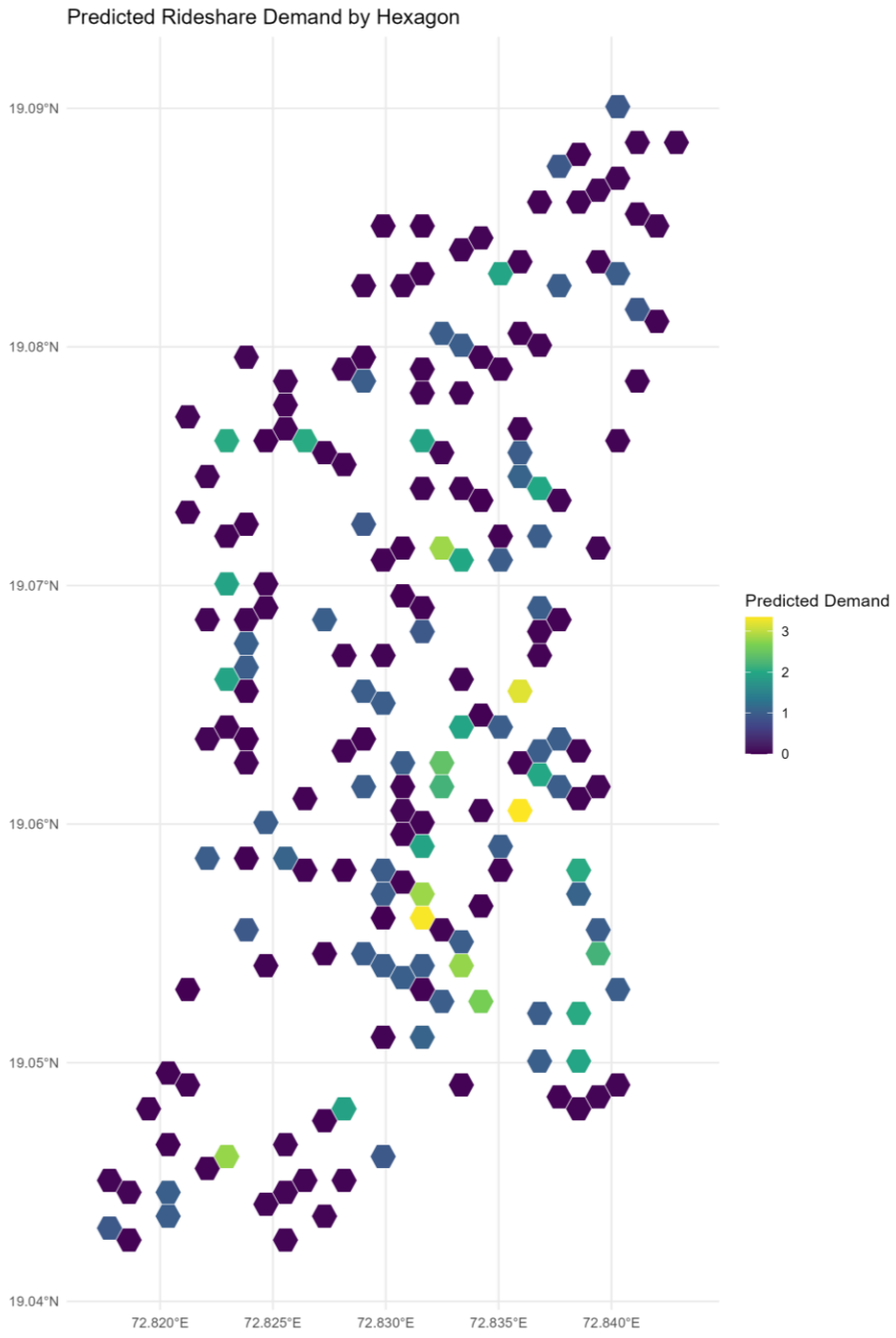


Figure 1 – Predicted Ride Share Demand

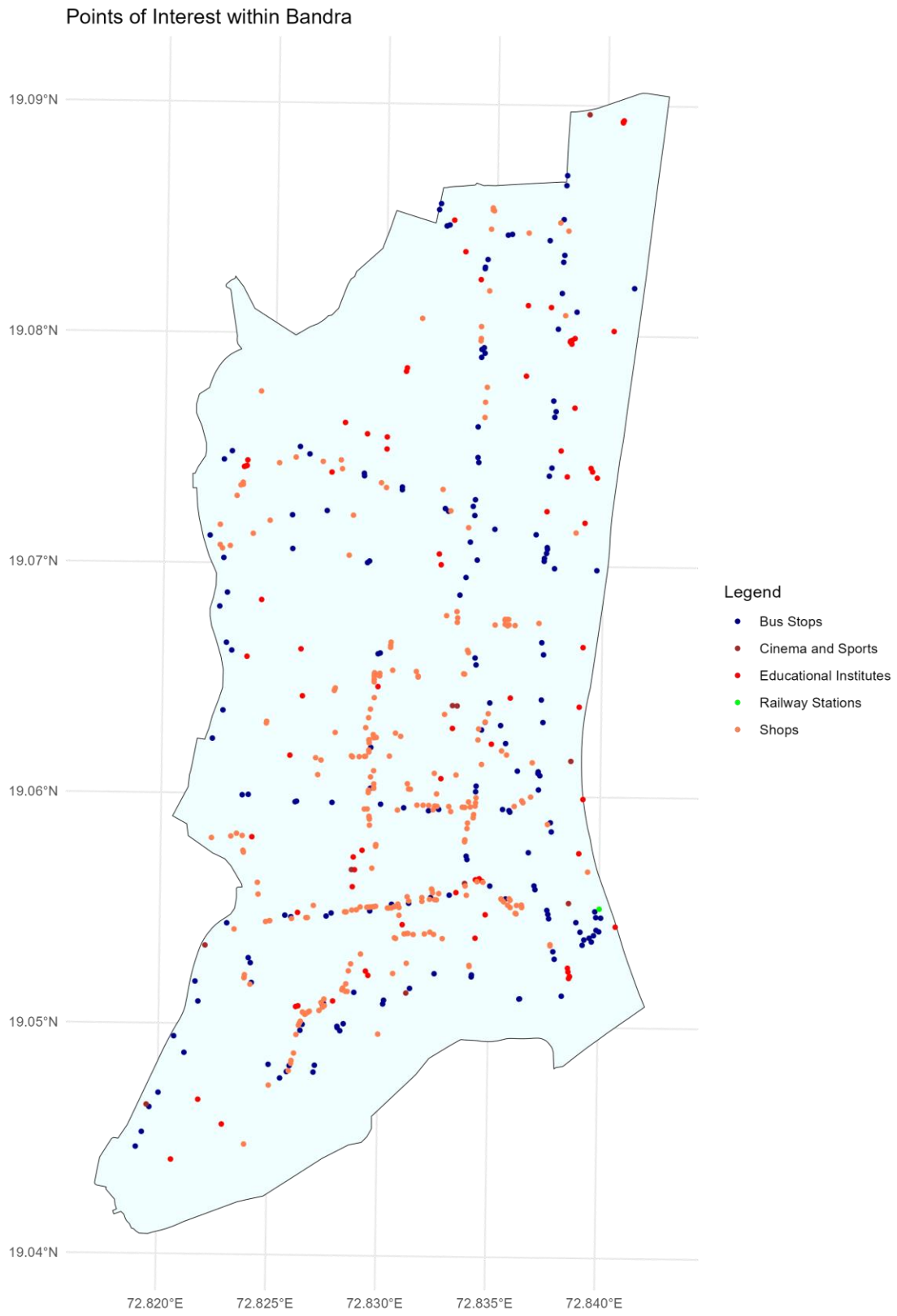


Figure – 2 Points of Interest in Bandra

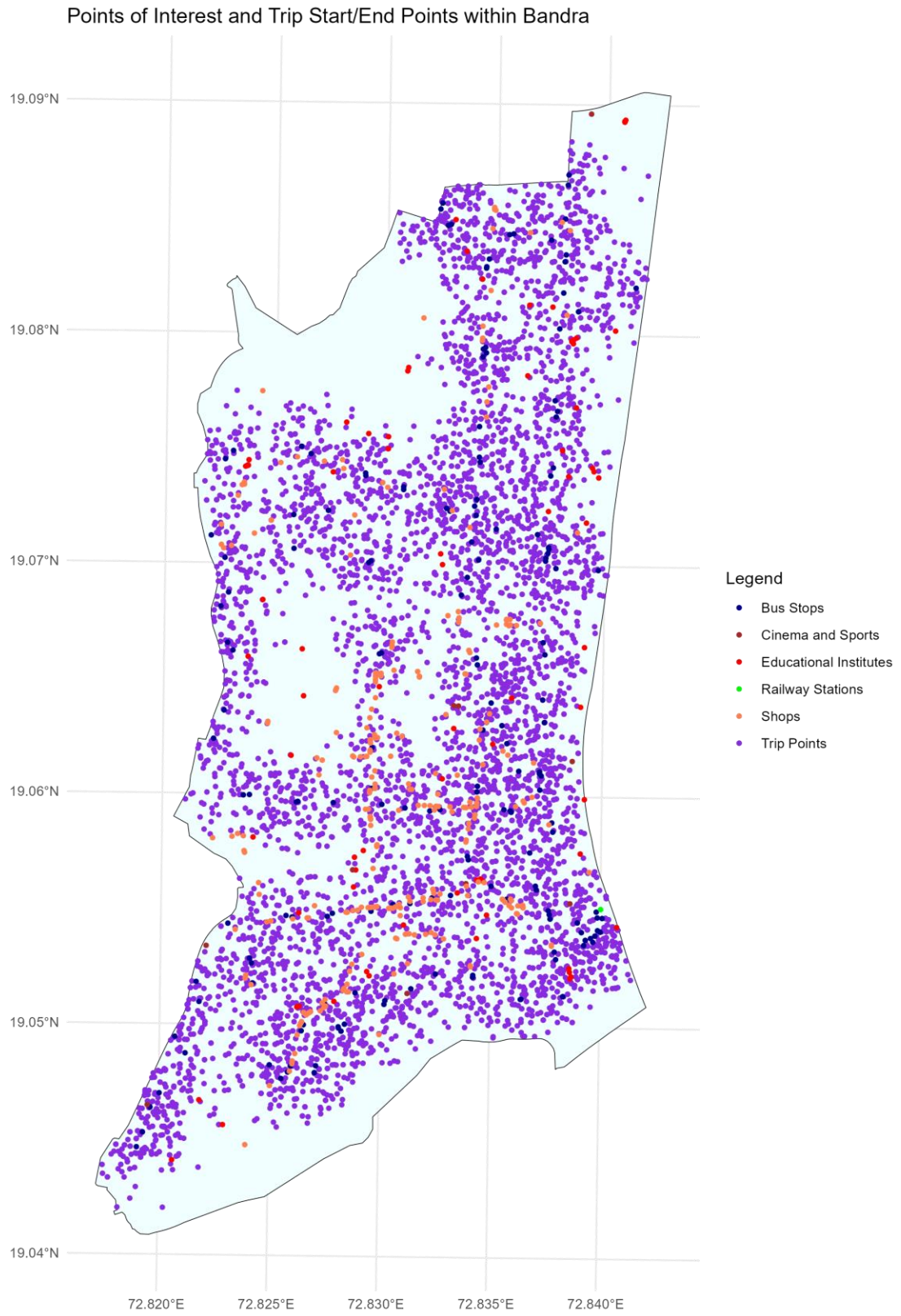


Figure – 3 All the synthetic trip start end point along with the Point of Interest

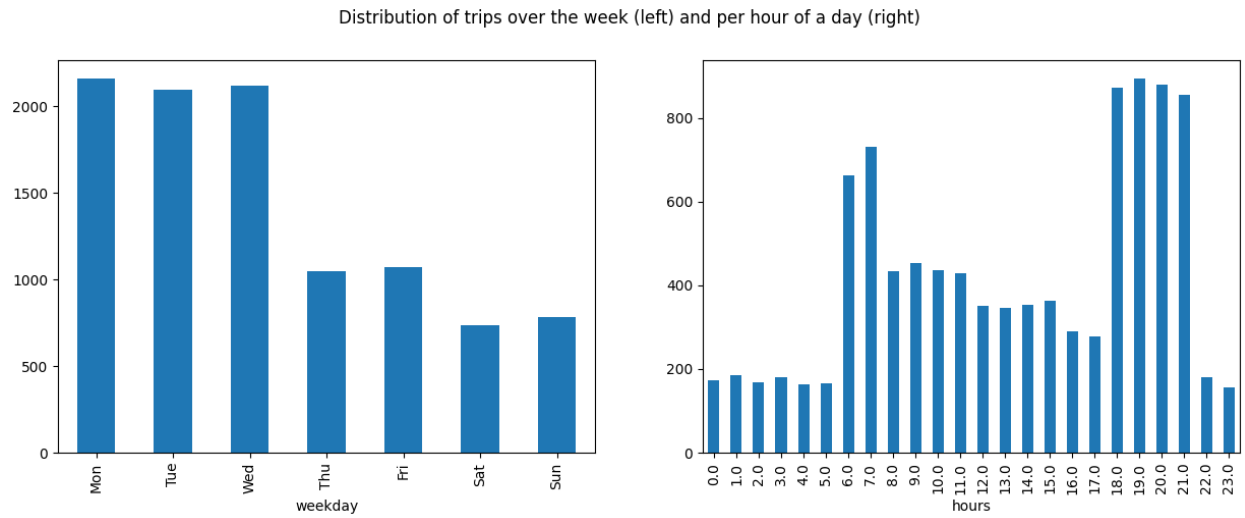


Figure – 4 Generated Trip data' temporal resolution

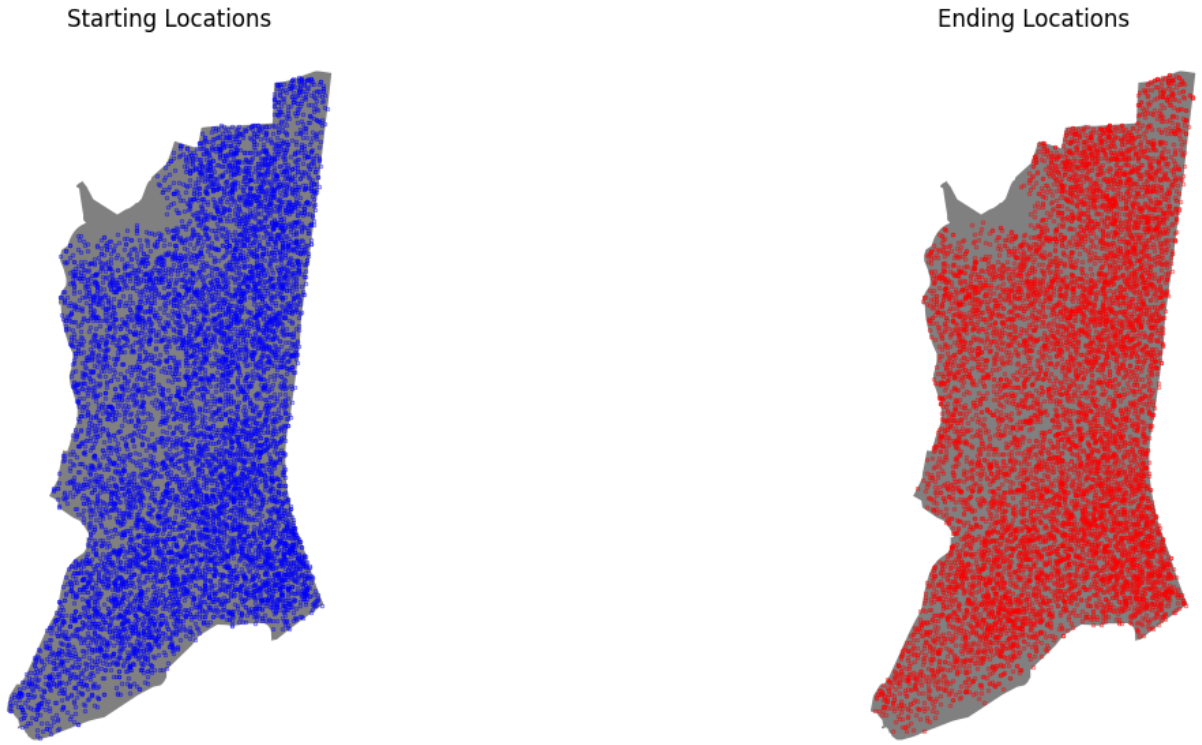


Figure – 5 Distribution of Start and End trip location (sampled 5000 points)

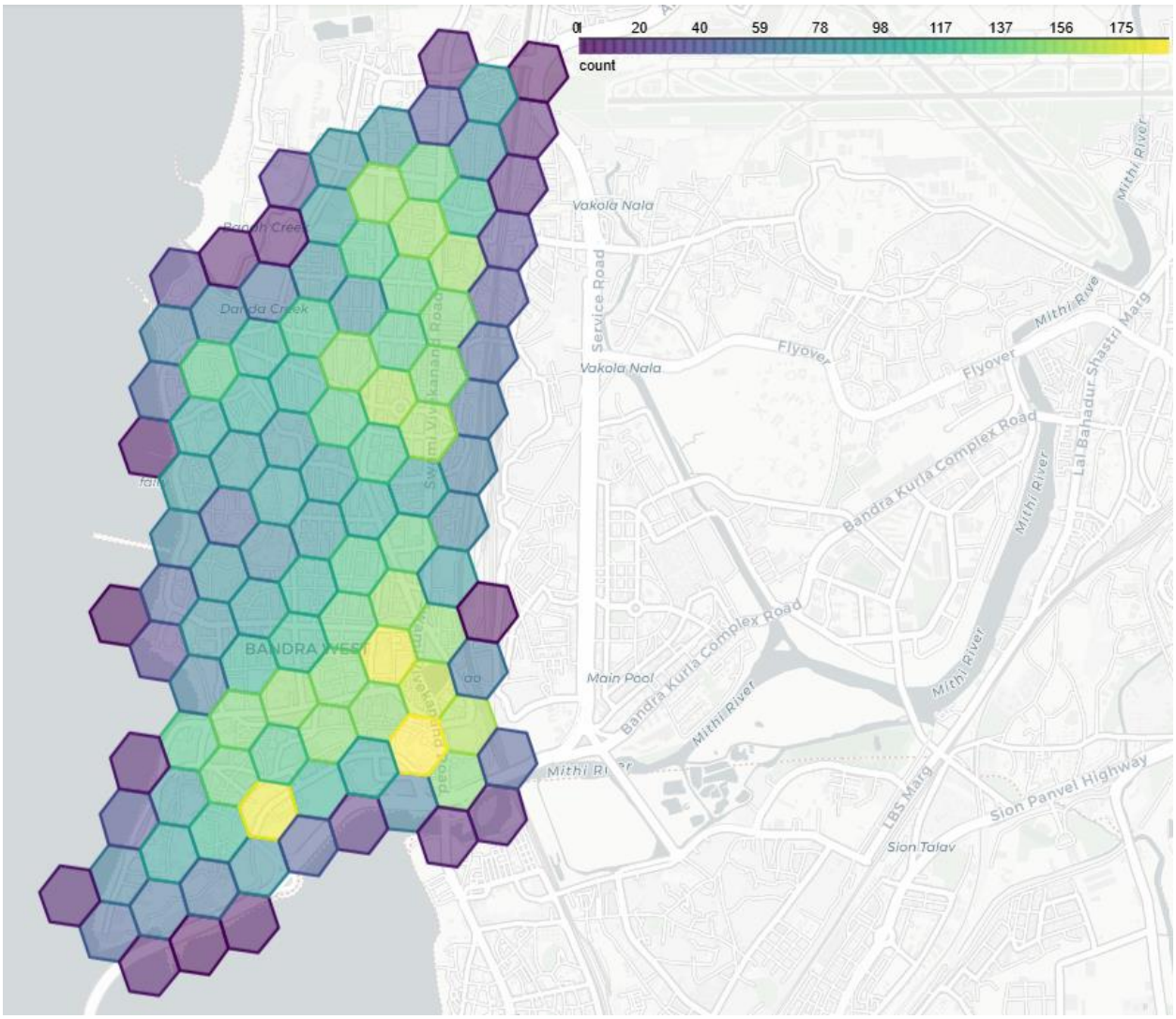


Figure – 6 H3: Uber's Hexagonal Hierarchical Spatial Index used for testing

- **Moran I Statistic for the trip starting point**:

```
> print(moran_test_start)

        Moran I test under randomisation

data:  start_coords[, 1]
weights: weights

Moran I statistic standard deviate = 36.818, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.9745812970     -0.0020040080      0.0007035547
```

- **Moran I Statistic for the trip end point**:
- **Moran's I test** results for the start points indicate a significant degree of spatial autocorrelation:

## • Results Analysis:

- **Moran I Statistic**:
- $I=0.97458I = 0.97458$I=0.97458: A value close to 1 indicates strong positive spatial autocorrelation (similar values are spatially clustered).
- **p-value**:
- $p<2.2e−16p < 2.2e\text{-}16$p<2.2e−16: This is highly significant, meaning the observed spatial autocorrelation is extremely unlikely to have occurred by chance.
- **Expectation and Variance**:
  - Expectation: $−0.002004\text{-}0.002004$−0.002004: The expected value under random distribution.
  - Variance: $0.00070360.00070360$0.0007036: The spread of the test statistic under randomization.

## • Interpretation:

- This result suggests that the **start points** exhibit a very strong spatial autocorrelation. This outcome is desirable if the goal was to create a realistic spatial distribution of start points that aligns with the underlying KDE-based density and autocorrelation model.

```
> print(moran_test_end)

        Moran I test under randomisation

data:  end_coords[, 1]
weights: weights_end

Moran I statistic standard deviate = 36.818, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.9745812970     -0.0020040080      0.0007035547
```

**Moran's I test** results for the **end points** show identical statistics to those for the start points, indicating a strong spatial autocorrelation:

## Results Summary:

- **Moran I Statistic**:
- $I=0.97458I = 0.97458$I=0.97458: This indicates a very high positive spatial autocorrelation among the end points, similar to the start points.
- **p-value**:

$p<2.2e−16p < 2.2e-16$p<2.2e−16: This confirms the result is highly statistically significant.

- **Expectation and Variance**:

Same as the start points, with an expected value of $−0.002004-0.002004$−0.002004 under randomization and variance of $0.00070360.0007036$0.0007036.

## Interpretation:

- Both the **start points** and **end points** show strong spatial clustering, suggesting that the modifications using KDE for point weights and spatial autocorrelation in endpoint generation have been successfully implemented.
- This outcome implies that the dataset now reflects realistic urban mobility dynamics, where start and end points are spatially correlated and adhere to the KDE-based density patterns.

## *Spatial Lag Model (SLM)*

```
Call:lagsarlm(formula = total_demand ~ lag_demand, data = hex_grid_with_demand,
listw = hex_weights, method = "eigen")

Residuals:
     Min       1Q   Median       3Q      Max
-1.32090 -0.51125 -0.30883  0.42128  7.08393

Type: lag
Coefficients: (asymptotic standard errors)
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 0.308832   0.046380  6.6588 2.761e-11
lag_demand  0.404828   0.060869  6.6508 2.915e-11

Rho: -1.6958e-08, LR test value: -4.0927e-12, p-value: 1
Asymptotic standard error: 0.047123
    z-value: -3.5987e-07, p-value: 1
Wald statistic: 1.295e-13, p-value: 1

Log likelihood: -1209.889 for lag model
ML residual variance (sigma squared): 0.7378, (sigma: 0.85896)
Number of observations: 955
Number of parameters estimated: 4
AIC: 2427.8, (AIC for lm: 2425.8)
LM test for residual autocorrelation
test value: 517.39, p-value: < 2.22e-16
```

### *Model Interpretation*

1. **Coefficients:**
    a. **Intercept**: The baseline predicted value of `total_demand` when `lag_demand` is 0. It is statistically significant (`p < 0.001`).
    b. **Lagged Demand (`lag_demand`)**: The coefficient (0.4048) represents the impact of demand in neighboring hexagons on the hexagon's demand. It is statistically significant (`p < 0.001`), indicating that spatial autocorrelation is an important factor in predicting demand.
2. **Rho (ρ):**
    a. The spatial autoregressive coefficient (`rho`) is approximately 0, which suggests weak or no spatial dependence in the residuals of this model. However, this is unusual given significant `lag_demand`.
3. **Residual Variance:**
    a. The residual variance (`sigma squared`) is 0.7378, suggesting the model's unexplained variation.
4. **Log-Likelihood and AIC:**
    a. Log-Likelihood: `-1209.889`, indicating the model's fit to the data.

b. AIC: 2427.8. While a lower AIC suggests a better fit, this should be compared with other models, such as a **Spatial Durbin Model (SDM)**, to confirm which model fits best.

5. **Residual Autocorrelation:**
   a. The LM test for residual autocorrelation has a **test value of 517.39** with a **p-value < 2.22e-16**, indicating strong residual spatial autocorrelation. This suggests the model does not fully account for spatial dependencies.

## Spatial Durbin Model (SDM) using the lagsarlm function

```
Call:lagsarlm(formula = total_demand ~ population + lag_population,
    data = hex_grid_with_demand, listw = hex_weights, method = "eigen")

Residuals:
     Min       1Q   Median       3Q      Max
-1.00340 -0.51540 -0.42027  0.45623  7.24242

Type: lag
Coefficients: (asymptotic standard errors)
                 Estimate  Std. Error z value Pr(>|z|)
(Intercept)     2.2727e-01  1.3641e-01  1.6661  0.09570
population     -6.5661e-05  1.0825e-04 -0.6065  0.54416
lag_population  3.9628e-04  2.1931e-04  1.8069  0.07077

Rho: 0.21717, LR test value: 23.58, p-value: 1.1983e-06
Asymptotic standard error: 0.043506
    z-value: 4.9917, p-value: 5.9849e-07
Wald statistic: 24.917, p-value: 5.9849e-07

Log likelihood: -1218.255 for lag model
ML residual variance (sigma squared): 0.74259, (sigma: 0.86174)
Number of observations: 955
Number of parameters estimated: 5
AIC: 2446.5, (AIC for lm: 2468.1)
LM test for residual autocorrelation
test value: 0.31265, p-value: 0.57606
```

### *Key Results*

1. **Intercept**:
   a. **Estimate**: 0.22727
   b. **p-value**: 0.0957
   c. The intercept is marginally insignificant at the 0.05 level. This value represents the baseline `total_demand` when all other variables are zero.
2. **Population**:
   a. **Estimate**: -6.5661e-05
   b. **p-value**: 0.54416

   c. The local population's effect on rideshare demand is negative but statistically insignificant.
3. **Lagged Population**:
   a. **Estimate**: 3.9628e-04
   b. **p-value**: 0.07077
   c. The spatially lagged population has a positive effect, nearing significance at the 0.05 level. This suggests that population in neighboring hexagons might slightly influence demand.
4. **Rho (Spatial Autocorrelation)**:
   a. **Estimate**: 0.21717
   b. **p-value**: 5.9849e-07 (highly significant)
   c. Indicates significant spatial dependence. Positive `rho` suggests that demand in one hexagon is positively associated with demand in neighboring hexagons.
5. **Model Fit**:
   a. **Log likelihood**: -1218.255
   b. **AIC**: 2446.5 (compared to 2468.1 for a simple linear model)
   c. AIC indicates that the spatial lag model is better than a non-spatial regression model.
6. **Residual Diagnostics**:
   a. **LM test for residual autocorrelation**:
      i. **p-value**: 0.57606
      ii. Suggests no significant residual spatial autocorrelation, indicating the model adequately accounts for spatial dependence.

## Interpretation and Next Steps

- **Model Explanation**:
  - While the direct effect of `population` is not significant, the spatially lagged population (`lag_population`) suggests that surrounding areas' population might influence rideshare demand.
  - Significant `rho` indicates strong spatial dependence, justifying the use of a spatial lag model.
- **Next Steps**:
  - **Improving Predictors**:
    - Explore additional covariates, such as income levels, land use, or accessibility to transit hubs.

        - Include temporal features like peak hours or weekday/weekend differentiation.
    - **Validation**:
        - Cross-validate the model to assess predictive accuracy.
        - Compare with other spatial models (e.g., Spatial Error Model or GWR).
    - **Visualization**:
        - Map residuals to identify areas where the model might underperform.
        - Plot predicted demand to examine spatial patterns.

## Summary

This project successfully developed a robust pipeline to analyze and predict rideshare demand using spatial data and advanced modeling techniques. By leveraging **Points of Interest (POI)** data, **Kernel Density Estimation (KDE)**, and **spatial autocorrelation**, the methodology ensured that synthetic data generation closely mimicked real-world urban mobility patterns.

The workflow systematically integrated spatial data preparation, synthetic trip simulation, and predictive modeling. Key achievements included:

1. **Realistic Data Generation**: KDE and proximity-based weighting ensured trip origins and destinations aligned with urban activity hotspots.
2. **Spatially Informed Modeling**: Spatial econometric models captured the influence of spatial lags on demand, while machine learning provided robust predictions based on spatial and temporal features.
3. **Validation and Insights**: Moran's I and demand aggregation metrics confirmed the reliability of spatial patterns and predictive models.

This comprehensive framework provides a scalable approach to studying urban mobility, offering valuable insights into rideshare demand dynamics that can support urban planning, transportation management, and policy development.

## Further work

1. **Refining Spatial Autocorrelation**: Refining the way to define spatial dependencies between points. Consider experimenting with different values for k (the number of nearest neighbors) or use a more tailored approach for generating correlated end points based on proximity.
2. **Validate End Points**: Check whether the **end points** show similar spatial autocorrelation. Strong spatial patterns in both start and end points would indicate the effectiveness of adjustments.
3. **Visualize Spatial Patterns**: Use plots to verify the clustering:
    a. **Start Points Distribution**: Plot the KDE or points spatially to visually confirm clustering.
    b. **End Points Distribution**: Overlay the end points and compare their patterns.

## Reference

Willing, Christoph, Konstantin Klemmer, Tobias Brandt, and Dirk Neumann. “Moving in Time and Space – Location Intelligence for Carsharing Decision Support.” Decision Support Systems 99 (July 2017): 75–85. https://doi.org/10.1016/j.dss.2017.05.005.

Wagner, Felix, Nikola Milojevic-Dupont, Lukas Franken, Aicha Zekar, Ben Thies, Nicolas Koch, and Felix Creutzig. “Using Explainable Machine Learning to Understand How Urban Form Shapes Sustainable Mobility.” Transportation Research Part D: Transport and Environment 111 (October 2022): 103442. https://doi.org/10.1016/j.trd.2022.103442.

Geng X, Li Y, Wang L, Zhang L, Yang Q, Ye J, Liu Y (2019) Spatiotemporal multigraph convolution network for ride-hailing demand forecasting. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 3656–3663