Spatial Data to Predict Demand

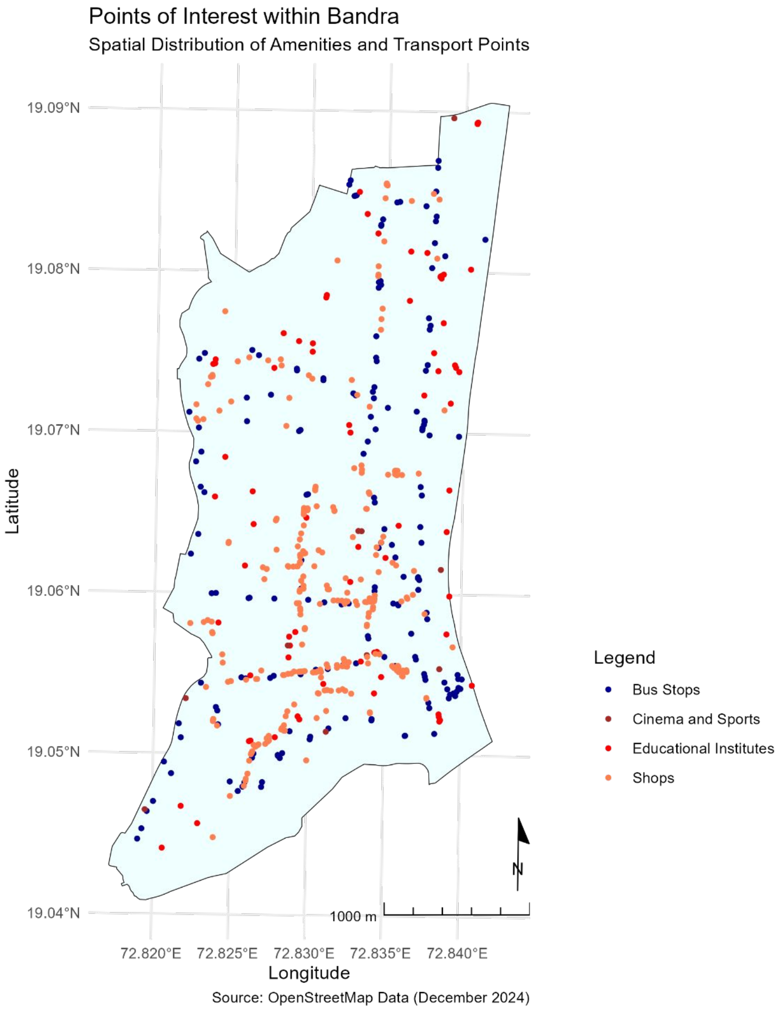
Generating Synthetic Geospatial Data

Manish Bilore

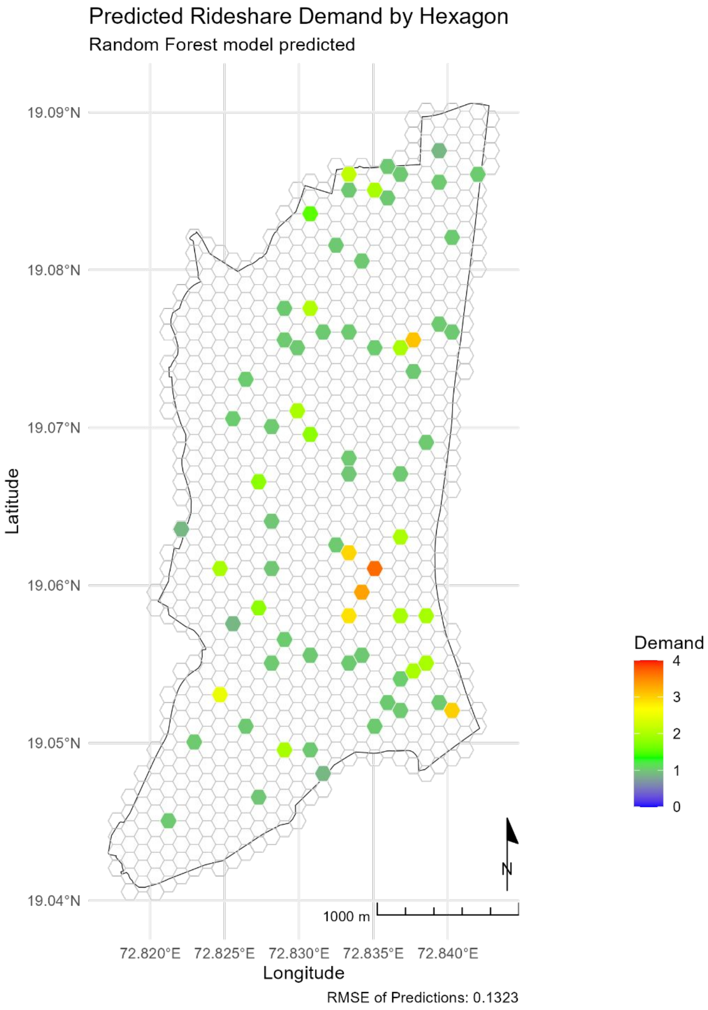
M.Tech IIT B

Objective

- Analyze and predict urban rideshare demand
- Use spatial and temporal models
- Leverage
    - Spatial Autocorrelation,
    - Random Forest techniques



Points of Interest within Bandra
Spatial Distribution of Amenities and Transport Points
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Legend
Bus Stops
Cinema and Sports
Educational Institutes
Shops
Source: OpenStreetMap Data (December 2024)



Predicted Rideshare Demand by Hexagon
Random Forest model predicted
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Demand
4
3
2
1
0
RMSE of Predictions: 0.1323

Step 01
Data Loading and Preprocessing

Input: Raw urban Point data.
Process:

- Load datasets for rideshare demand
- Data cleaning:
- Feature engineering:
    a. Add time-related features (hour, day, etc.) and
    b. Spatial features (H3 grid encoding).

Output: Cleaned and feature-enriched dataset.

Step 02
Predictive Model Development

Input: Preprocessed dataset
Process:

- Split data into training, validation, and test sets
- Train predictive models
- Hyperparameter tuning for optimal model performance
- Evaluate model performance using metrics

Output: Trained and validated predictive models.

Step 03
Visualization and Interpretability

Input: Model predictions and dataset
Process:

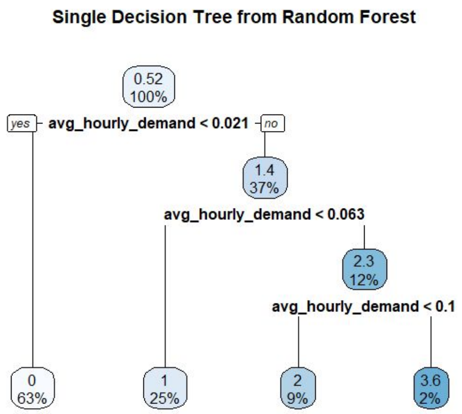
- Spatial visualizations of demand
- Interactive H3 grids
- Feature importance visualizations
- Build heatmap of temporal trend

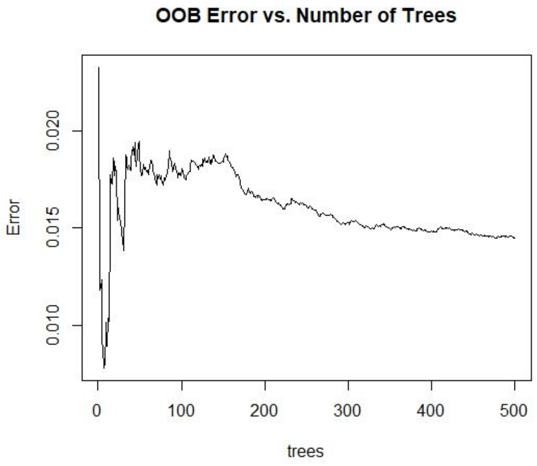
Output: Interactive and static visualizations, actionable insights.

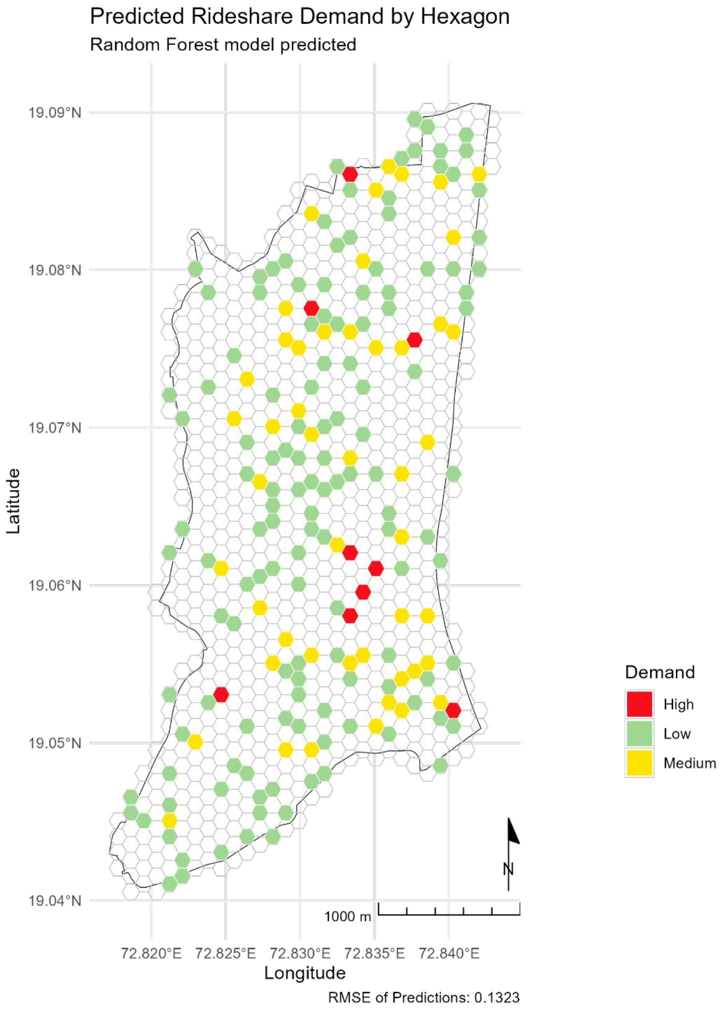
Methodology - Overall workflow

Random Forest Analysis

Using Spatial data to Predict Bike Rideshare demand







Predictive Model using Random Forest

Key Challenge

- Data availability
- Solution - Generation of Synthetic Data

Generating Synthetic Geospatial Data

Synthetic != Fake

Step 1
Setup and Preparation

Load district polygon and Point of Interest (POI) layers (e.g., bus stops, shops).

Clean and dissolve duplicate POIs.

Transform layers to a common CRS (UTM Zone 43N).

Step 2
Generate KDE

Combine Influence Points by merging POI layers into a single influence layer

Generate Kernel Density Estimation (KDE) to create spatial influence weights.

Normalize KDE values for probabilistic sampling

Step 3
Spatial Autocorrelation

Calculate Moran's I

Use K-nearest neighbors for spatial weights

Assess spatial autocorrelation of sampled points

Step 4
Generate Trip Data

On synthetic points perform KDE and normalise the values

Define trip start and end points based on KDE weights

Avoid self-loops (start = end)

Add travel times and timestamps

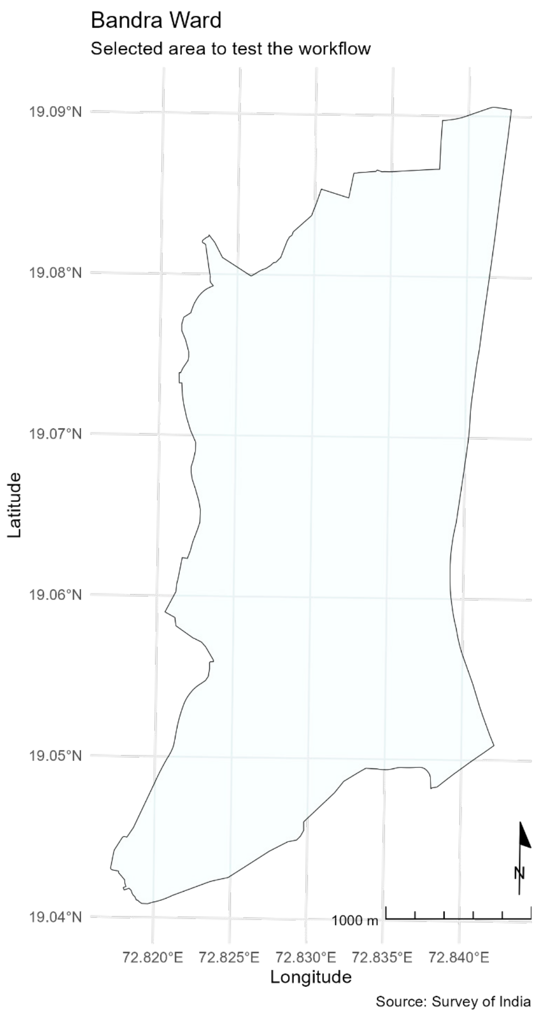
Step 5
Validation and Output

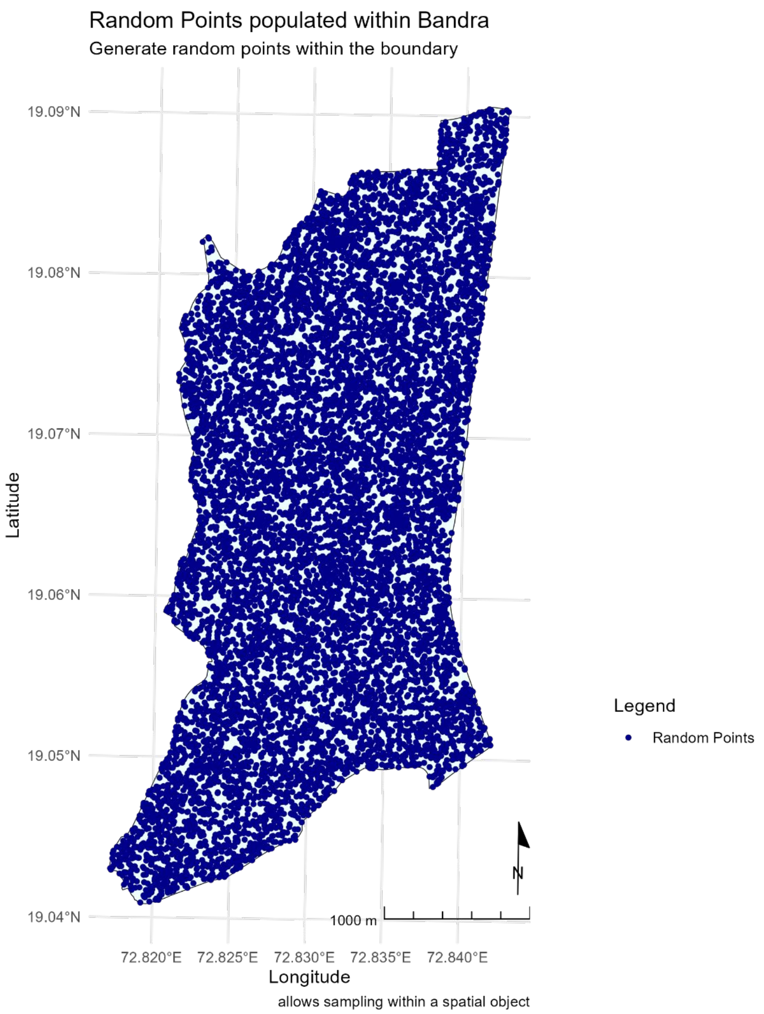
Verify Spatial Autocorrelation

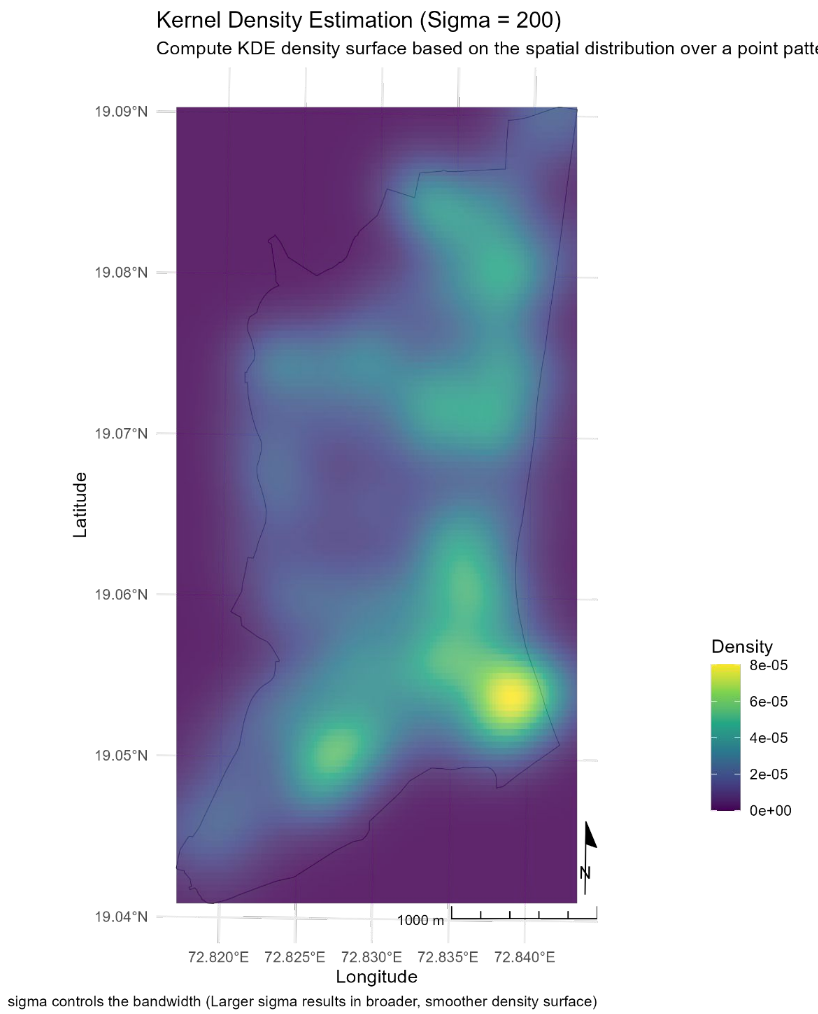
Calculate Moran's I for start and end points

Export synthetic trip data as CSV

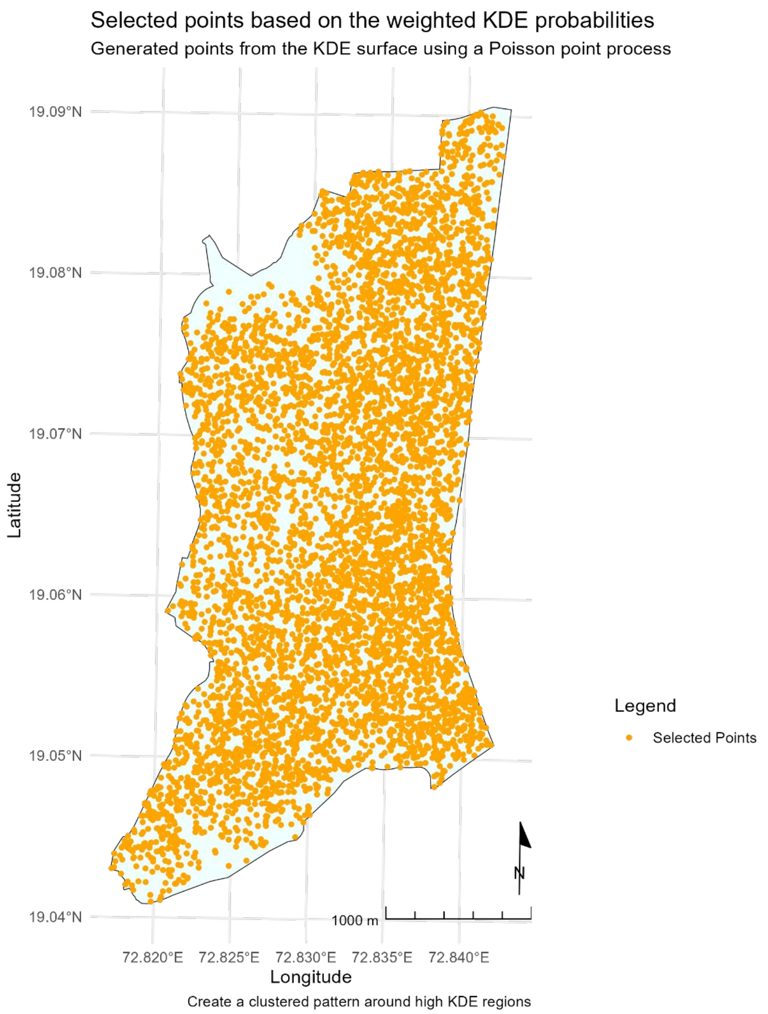
<u>Synthetic Data generation workflow</u>

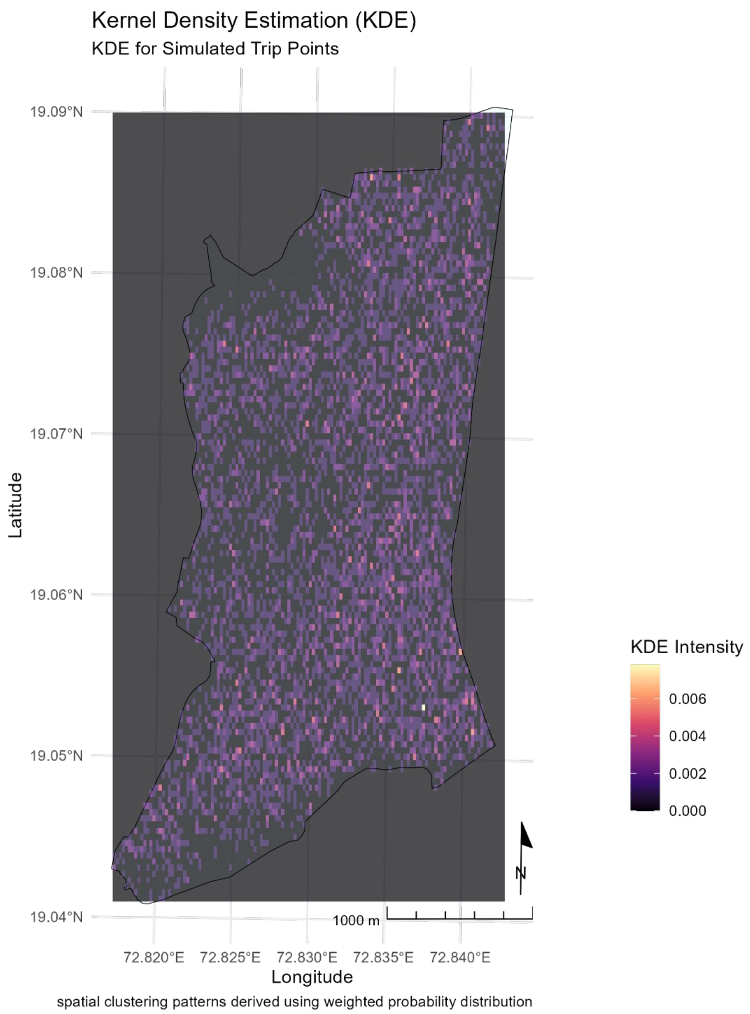

Bandra Ward
Selected area to test the workflow
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Source: Survey of India

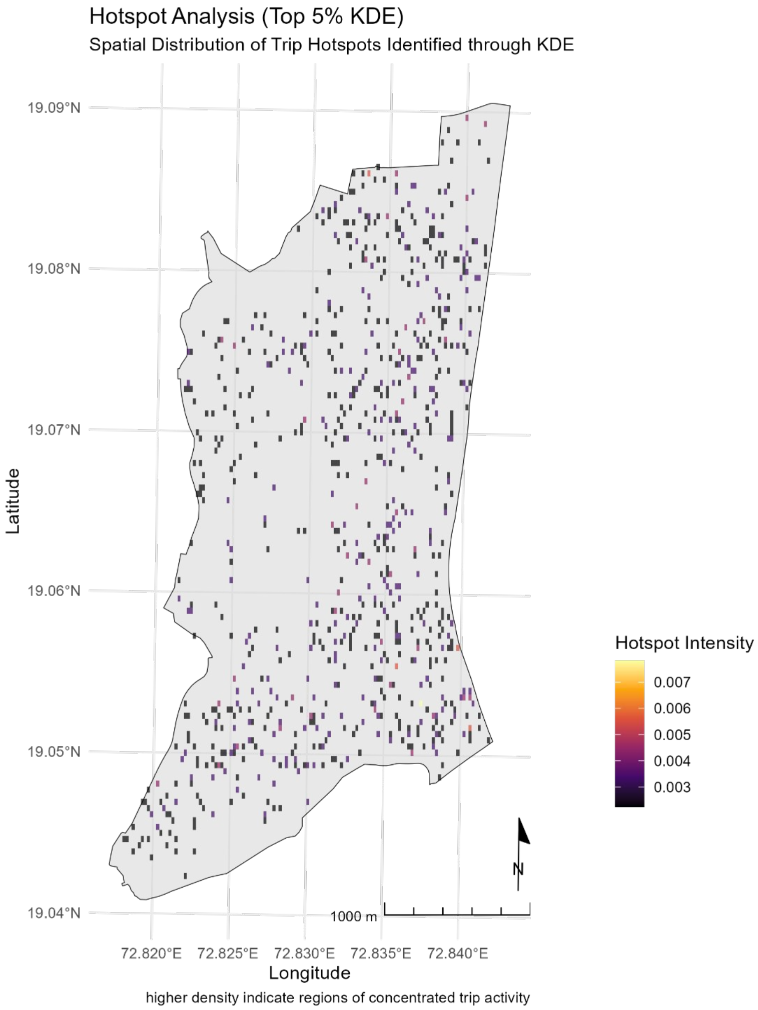

Random Points populated within Bandra
Generate random points within the boundary
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Legend
Random Points
allows sampling within a spatial object


Kernel Density Estimation (Sigma = 200)
Compute KDE density surface based on the spatial distribution over a point pattern
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Density
8e-05
6e-05
4e-05
2e-05
0e+00
sigma controls the bandwidth (Larger sigma results in broader, smoother density surface)

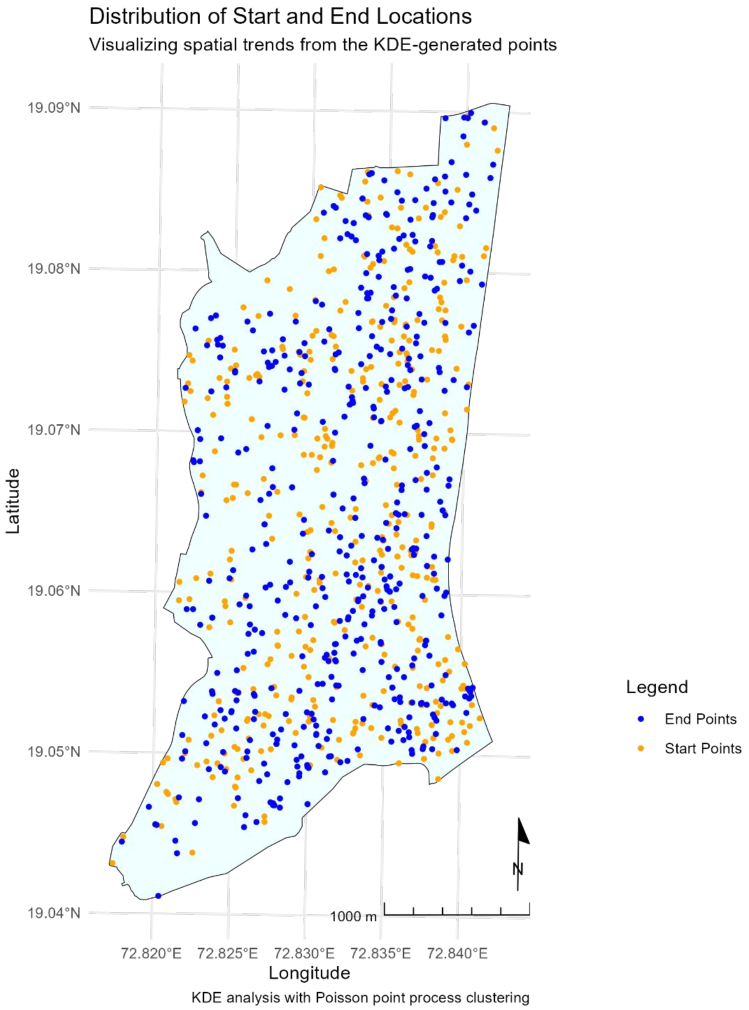

Buffer Zones Around Bus Stops in Bandra
Visualization of 235-meter buffer zones around bus stops
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Legend
Buffer
Bus Stops
Points within buffer receive additional weight simulating higher activity/influence


Selected points based on the weighted KDE probabilities
Generated points from the KDE surface using a Poisson point process
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Legend
Selected Points
Create a clustered pattern around high KDE regions


Kernel Density Estimation (KDE)
KDE for Simulated Trip Points
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
KDE Intensity
0.006
0.004
0.002
0.000
spatial clustering patterns derived using weighted probability distribution

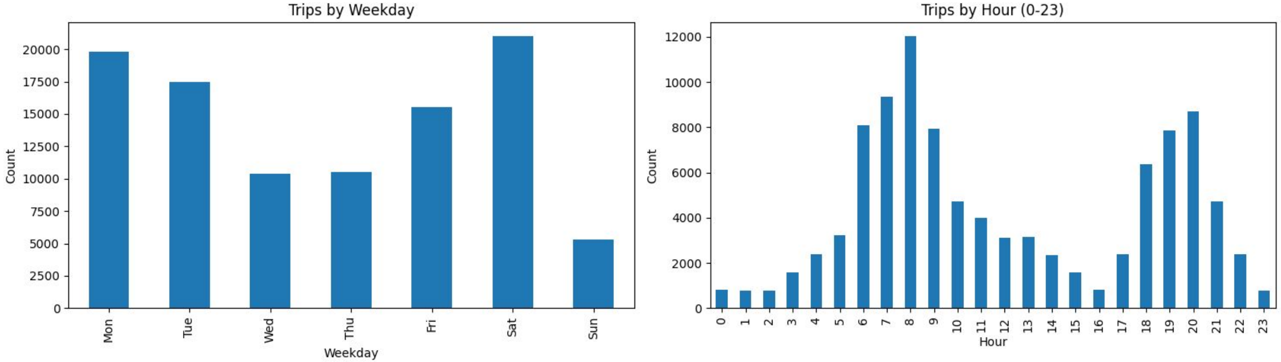


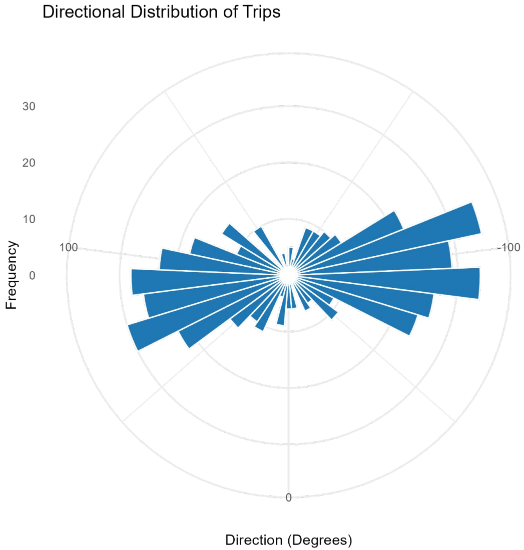
Hotspot Analysis (Top 5% KDE)
Spatial Distribution of Trip Hotspots Identified through KDE
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Hotspot Intensity
0.007
0.006
0.005
0.004
0.003
higher density indicate regions of concentrated trip activity

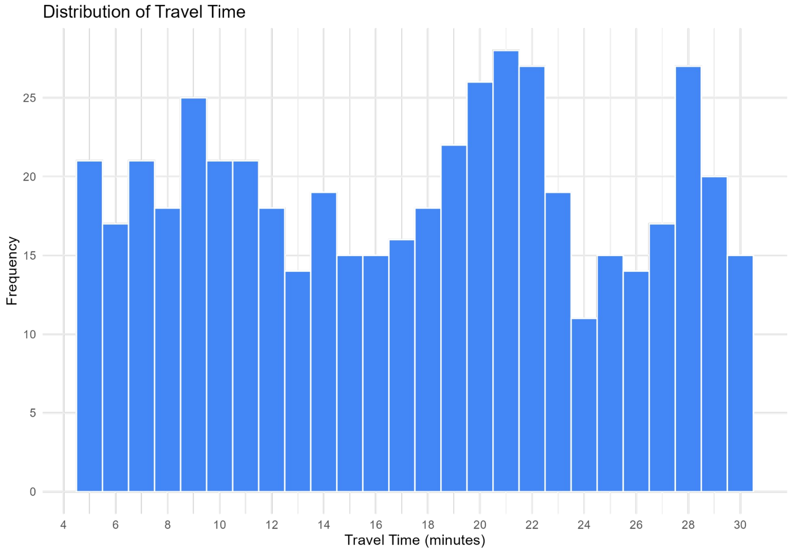


Distribution of Start and End Locations
Visualizing spatial trends from the KDE-generated points
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Legend
End Points
Start Points
KDE analysis with Poisson point process clustering

Distribution of trips over the week (left) and per hour of a day (right)







Nature of Generated Data

Spatial Autocorrelation

Relationship between spatial locations and their attribute values

Moran's I is a measure used to assess spatial autocorrelation

Positive spatial autocorrelation:

- Nearby locations have similar values
- clustered patterns like high-high or low-low

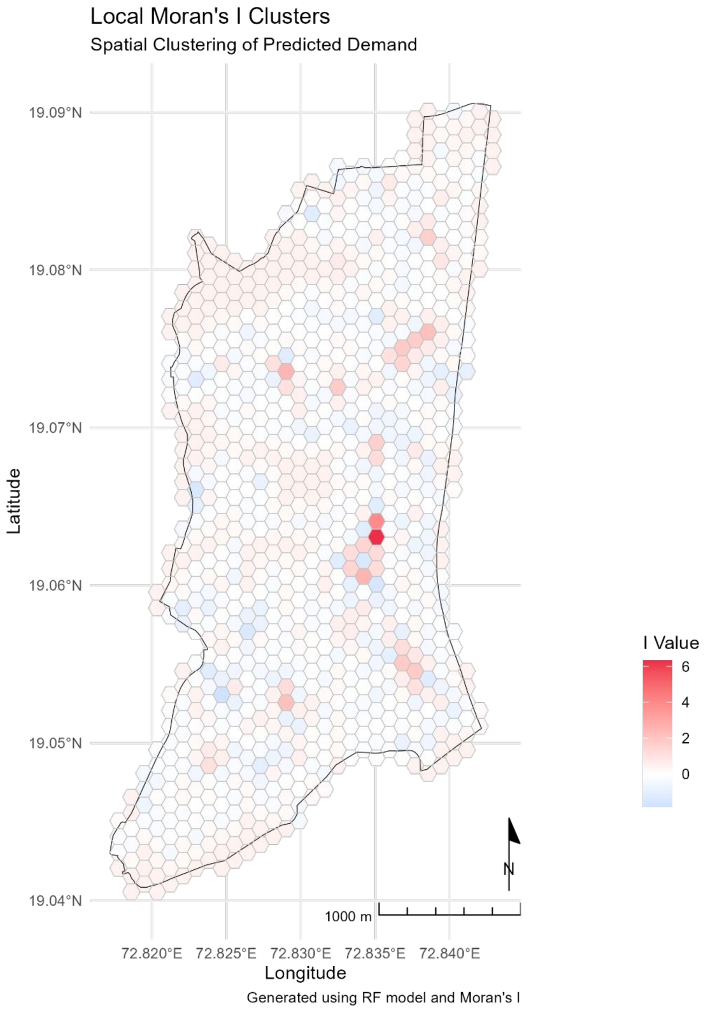
Negative spatial autocorrelation:

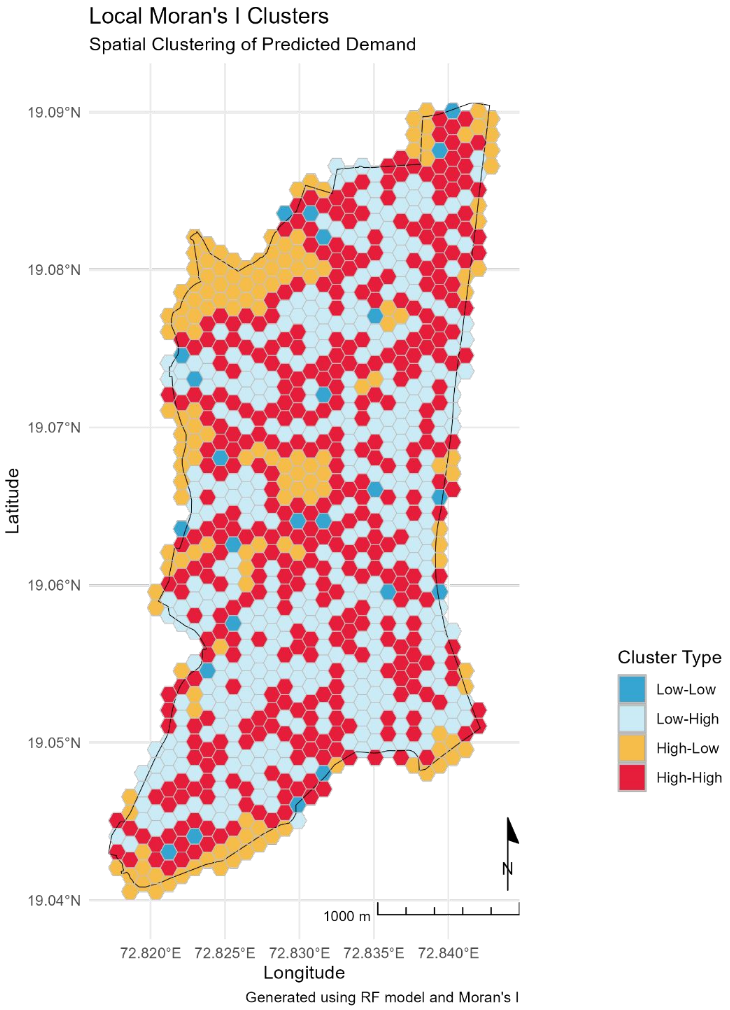
- Nearby locations have dissimilar values
- high-low or low-high patterns

No spatial autocorrelation

- Values are randomly distributed across space

Moran's I statistic for the generated data	
standard deviate	4.3842
p-value	5.82e-06
alternative hypothesis	greater
Moran I statistic	0.0943549750
Expectation	-0.0010482180
Variance	0.0004735254


Local Moran's I Clusters
Spatial Clustering of Predicted Demand
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
I Value
6
4
2
0
Generated using RF model and Moran's I


Local Moran's I Clusters
Spatial Clustering of Predicted Demand
19.09°N
19.08°N
19.07°N
19.06°N
19.05°N
19.04°N
Latitude
72.820°E
72.825°E
72.830°E
72.835°E
72.840°E
Longitude
1000 m
N
Cluster Type
Low-Low
Low-High
High-Low
High-High
Generated using RF model and Moran's I

Summary

- **Realistic Data Generation:**
  - KDE weighting ensured trip aligned with urban hotspots
- **Spatially Informed Modeling**:
  - Spatial models captured influence of spatial lag
  - ML predictions based on spatial and temporal features
- **Validation and Insights**:
  - Moran's I metrics confirmed the reliability of model

Future work

- **Scale the operation from one day to one year:**
  - Add weather data and compare with field data
- **Compare other ML models**
  - Identify best performing model