Special Lecture series

Geospatial & big data analysis

Understanding shared e-scooter usage patterns

A Spatial modeling approach

Case of Stockholm, Sweden

Omkar Parishwad
Urban Mobility Group
Department of Architecture and Civil Engineering

CHALMERS
UNIVERSITY OF TECHNOLOGY

Shared Electric-scooters

Shared micro-mobility systems

- Bike-sharing/e-scooter sharing (micro-mobility) systems with minimal electric power assistance presented as a
  - low-budget alternative,
  - seamless transportation provision
  - decarbonization potential: environmentally friendly travel option for short-distance trips.

Planning for distribution within a city?

Shared electric scooters in Sweden

[Source: Ernest Ojeh / Unsplash]

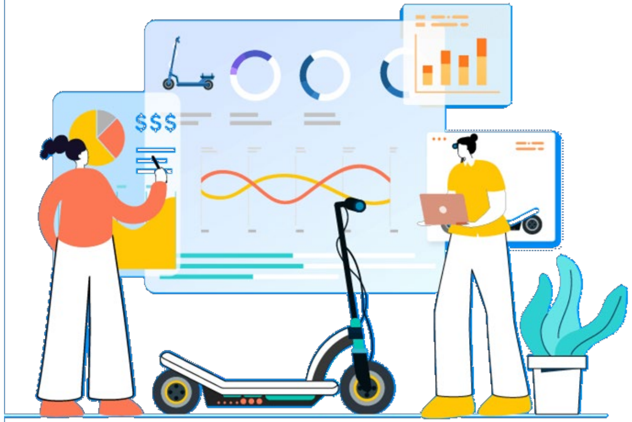
Shared bicycles in Chandigarh, India

[Source: Deepali Takkar, Panchkula in Hindustan Times]

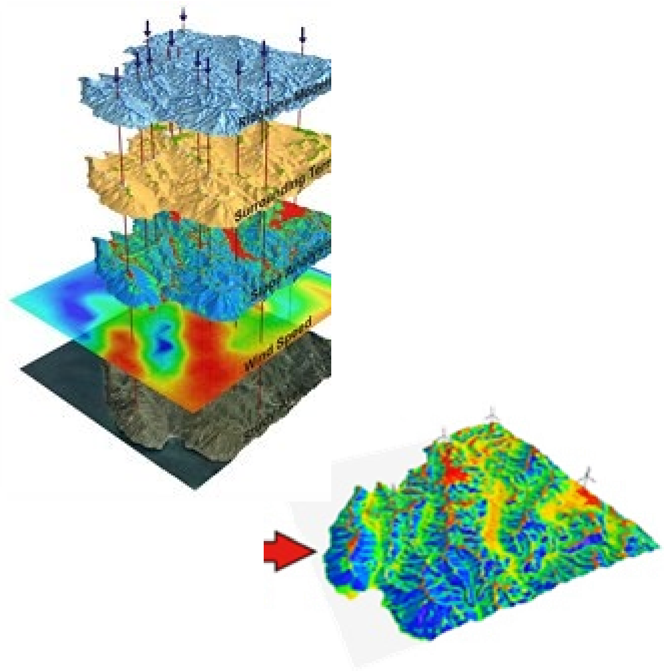
Analysis: Urban Planning context

Spatio-temporal demand estimation

- User behavior for shared e-scooter usage
- Understand usage wrt. our urban built environment
- Crucial for planning fleet distribution and operations
- Potential to enhance quality of life: more connected, accessible, sustainable cities.
- Climate change & our commitment to reducing greenhouse gas (GHG) emissions.



Need for spatial analysis for our cities



Urban Planning Challenges: traffic management, land use, infrastructure development, and environmental sustainability.

Spatial Complexity: patterns for specific locations, topography, and demographics of a city.

Analyzing Relationships: between variables and how they vary across different locations.

Data-Driven decisions: holistic view of urban dynamics for us to make more informed, efficient, and sustainable decisions- or evaluate policy impact.

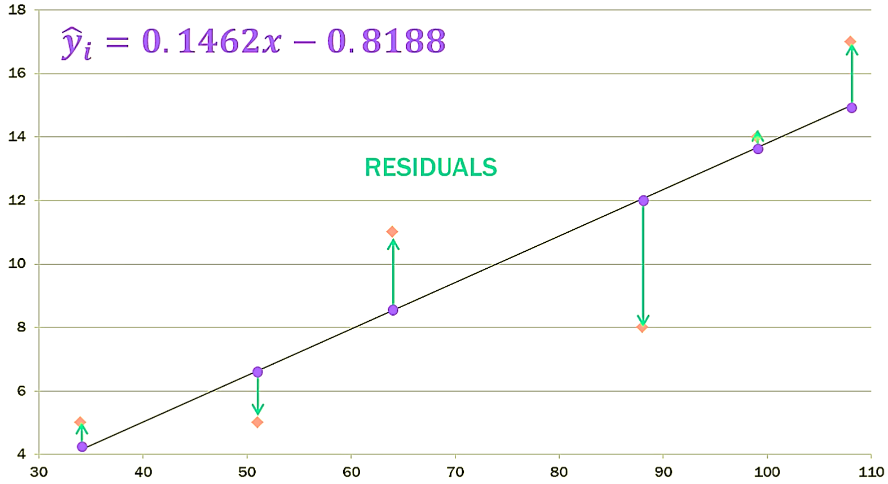
Traditional Modeling Approaches

Ordinary Least Squares (OLS) is a foundational linear regression method.

It assumes that relationships between variables are constant across the entire study area.

Geographically Weighted Regression (GWR) acknowledges that relationships between variables can vary across space.

It allows us to capture localized spatial patterns.

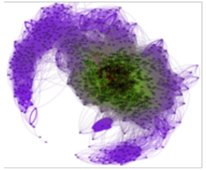


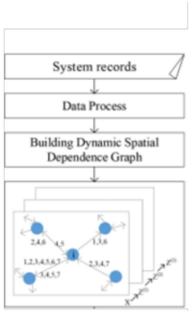
Ordinary Least Squares (OLS) approach

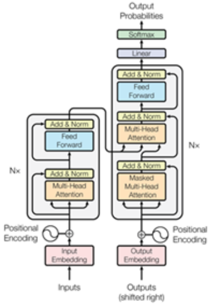
Modern Machine/deep learning models

Limitations with data-driven methods for predictions based on historical data (GNN, RNN)

- Data size and constraints (training time, overfitting, preprocessing)
- Overfitting vs model complexity (resource constraints)
- Interpretations, generalization for long term predictions.





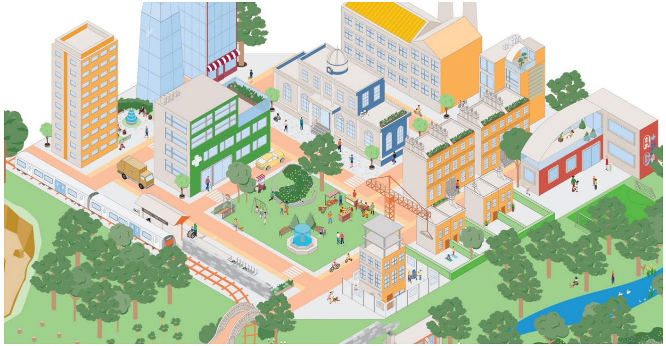


Beyond prediction, interpreting causal relationships

- How can we enhance/promote the usage in some areas?
- What are the travel/usage patterns related with, spatially-temporally?

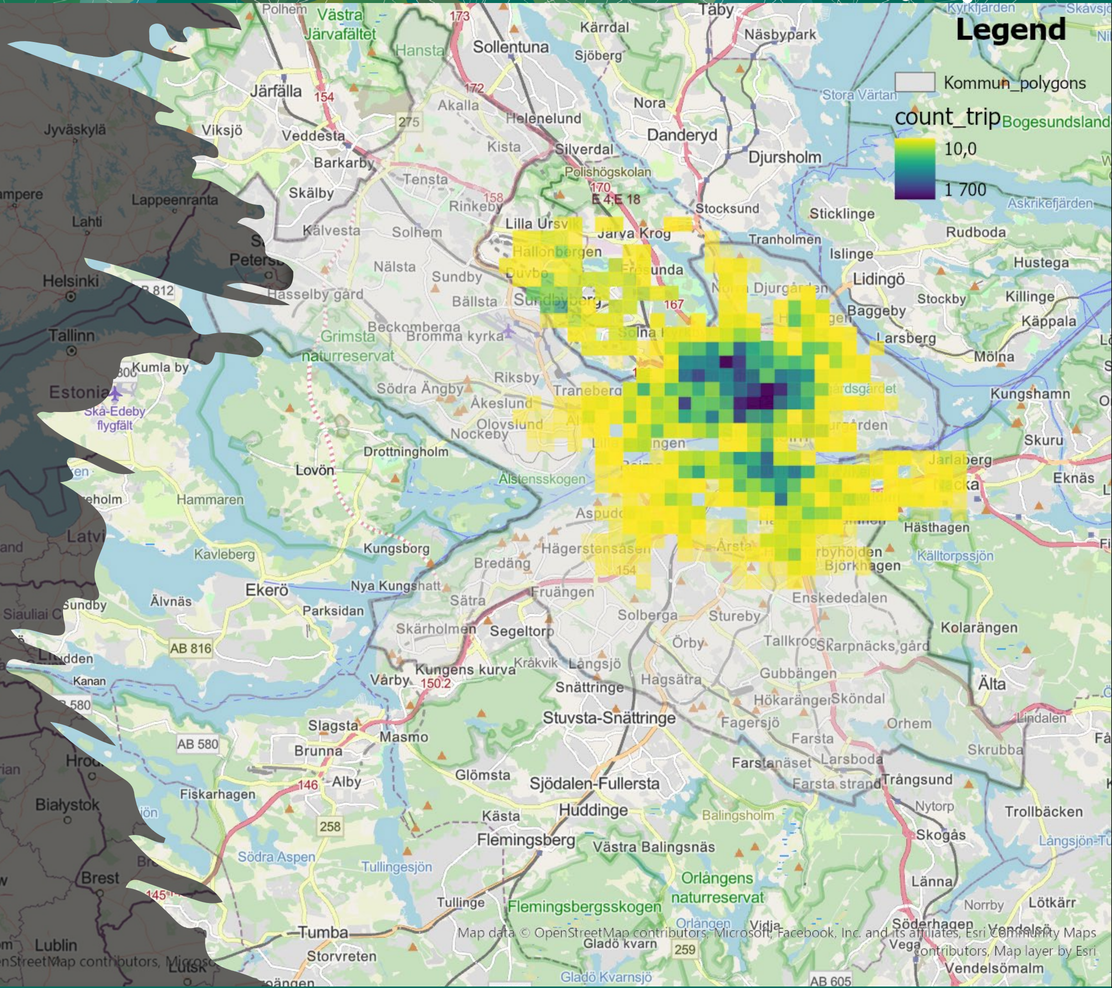
The effects/relevance of built environment on the demand for shared micro mobility systems

Land use, road/ transit network, Income, population density…



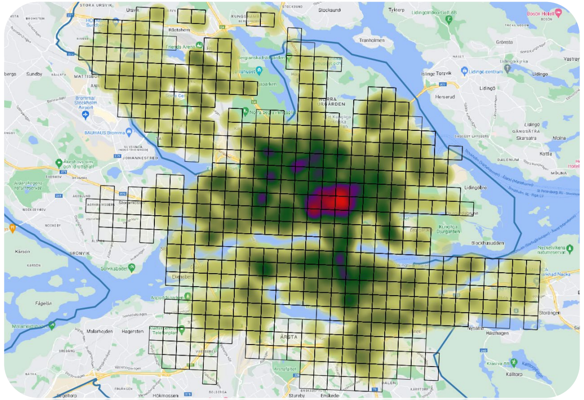
Case of Stockholm, Sweden

- Background
- Data
- Methodology
- Results/Summary



Data acquisition

- Shared **e-scooter trip data (2022)** from Tier and Voi (**Open-source API**)
- Spatially attributed data for socio-economic attributes (e.g., population, income..) from Swedish national repositories (**SCB**)
- Spatial data (shapefiles) for land use, transport infrastructure, facilities (POIs), attributed road network, and transit stations from Open Street Map (**OSM**).



Data: Approx. **2,300,000 trips** Stockholm, 12 months, **2022**

Study area: Järva and Norra Djurgården (59.39°N to 59.28°N) to Enskede Gard and Fruängen (18.18°E to 17.94°E), Stockholm City core area.

id	o_time	d_time	o_lat	o_lng	d_lat	d_lng	escooter_distance	escooter_time	transit_time	transit_totaltime	transit_walkdistance	transit_transitdistance
083ccd65-c2e4-4563-bcc1-856eff93c91c	2021-12-28 00:00:23	2021-12-28 00:06:24	59.342388	18.038866	59.343147	18.049500	608.781895	361	0.0	512.0	638.134534	0.00000
321399c3-07e4-43b6-b191-66660e9b5d50	2021-12-28 00:03:23	2021-12-28 00:45:31	59.330967	18.066385	59.337105	18.067427	685.049246	2528	0.0	634.0	799.175639	0.00000
415b85b8-a204-4927-9428-c5ef4974b519	2021-12-28 00:03:23	2021-12-28 00:09:25	59.289757	18.001770	59.293118	18.000690	378.693303	362	0.0	367.0	472.829030	0.00000

Data considerations

Aggregated to 24 hour trip count data for each

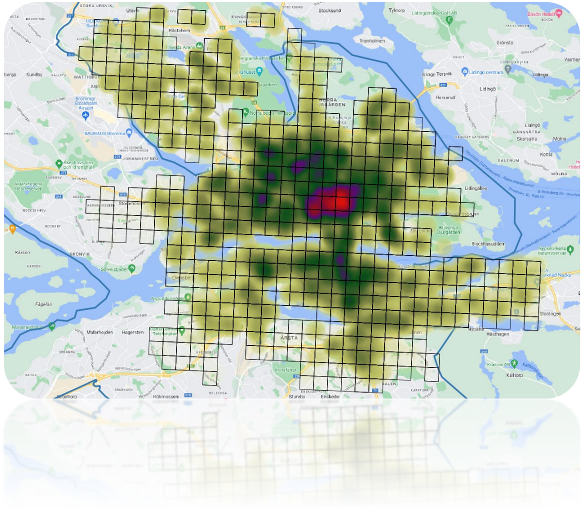
- season (summer, winter, spring, fall) (4)
- weekday/weekend (2)
- in spatial subdivision- fishnet grid (400x400m)

2,314,135 records after preprocessing for outliers

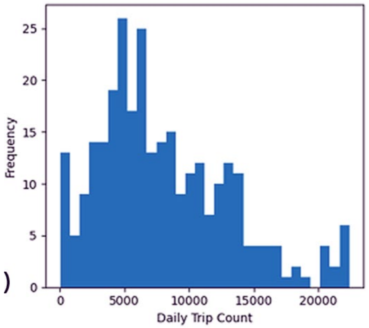
8,351 shared e-scooters in study area (Stockholm city core area)

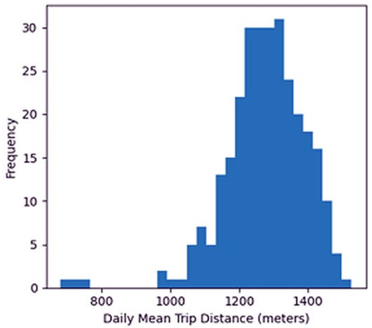
Considered region (12kmx12km) **499 grids**

Threshold of 5 trip count per day (**aggregated**)



Aggregated (24 hour) trips



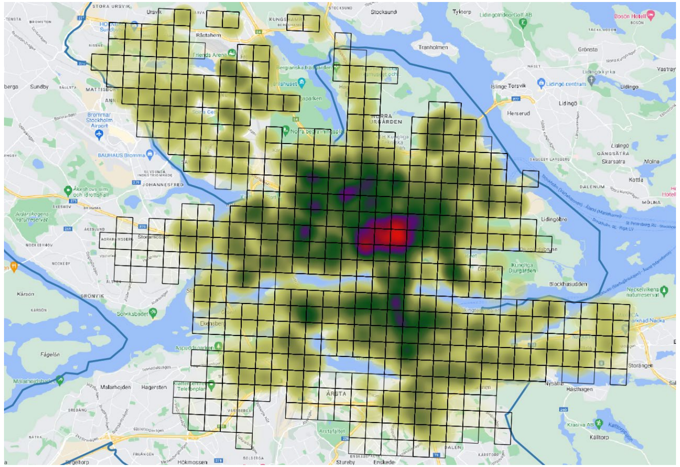


Usage pattern analysis

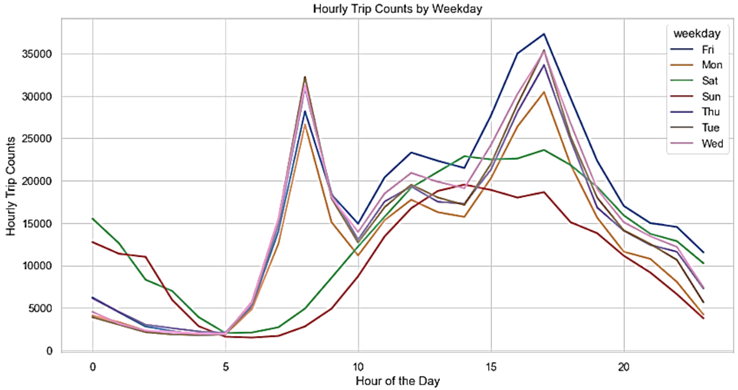
- Seasonal patterns: During the *summer*, the daily average trip counts surge by more than 35% compared to the 2022 daily average.
- Weekday patterns: Visible variation in usage pattern on weekend, as compared to weekdays.
- Spatial aggregation: Considering the temporal variations, we aggregate the trips for the fishnet grid zones. The potential spatial clustering of hotspots for the study area needs to be studied, as seen in the figure.

Reference: https://github.com/parishwadomkar/Usage-analysis-e-scooter/blob/main/00_EDA_Stockholm.ipynb

Spatial:



Temporal

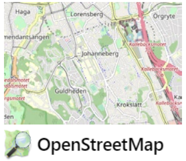


Data: Built environment factors

- Point of Interest (POI), road networks provided by OpenStreetMap, population census data

(Gao K. et.al., 2023)

Point of Interest Data



Transport facilities and network data

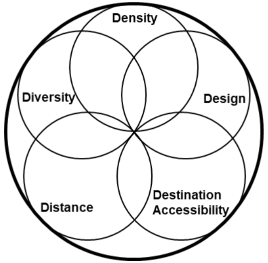
Figure source: Vector

Socio-economic data from open-access repositories

Other sources: documents

Semantic data fusion

Topic modeling

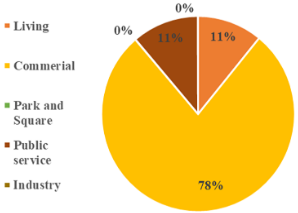


5-D transport-related built environment framework

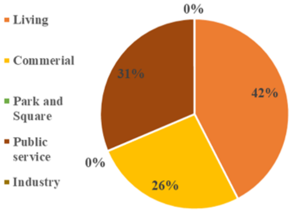
- ❑ Influencing factors
- ❑ Trip characteristics (spatiotemporal distributions)

Interpretable machine learning

- Data-driven methods to model complex nonlinear and interactions
- Explaining the effects of a factor in a data-driven way



Land use ratio based on POI number

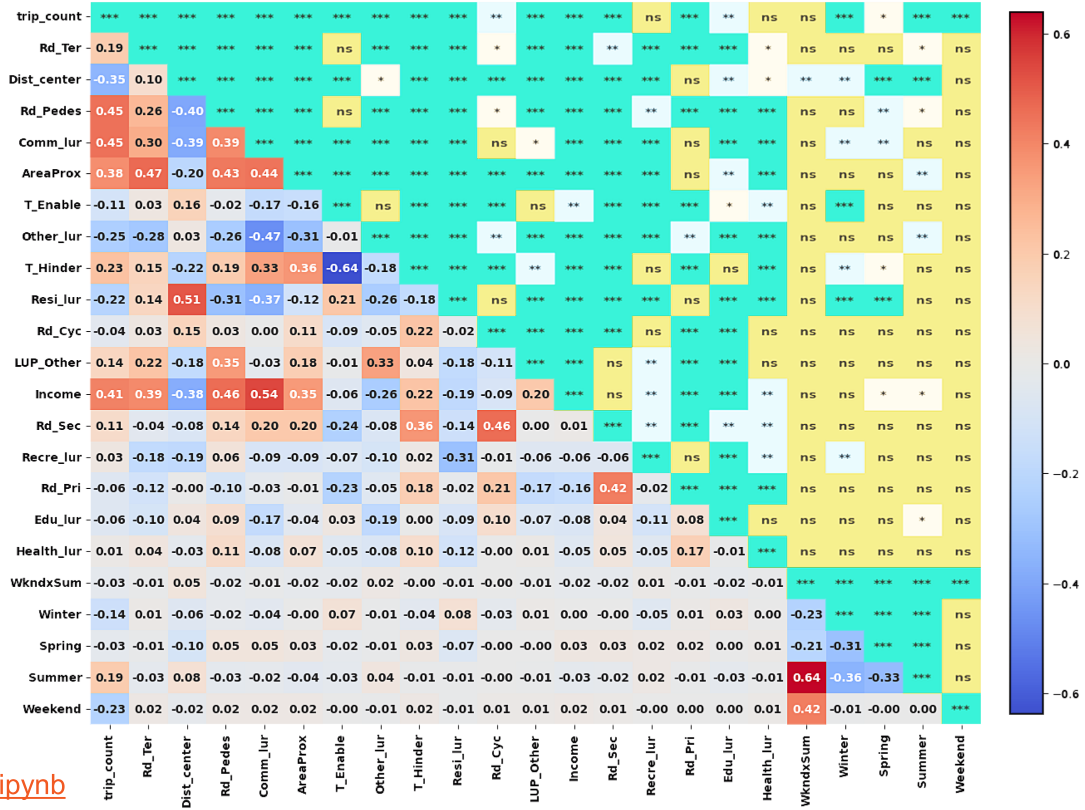


Land use ratio based on proposed TFIDF

Feature selection criterion

- VIF (Variance Inflation Factor) <15
- Spearman correlations <60%
- Spatial autocorrelations (Morans I)
- Local multicollinearity
- Dimentionality reductions
- Feature Interactions
- Local Condition Number (LCC),
- Collinearity Index (CN)

Selected features (22 out of 45)



Reference: https://github.com/parishwadomkar/Usage-analysis-e-scooter/blob/main/01_FeatureSelection_Sto.ipynb

Global ML Models

- Global machine learning models like random forest and XGBoost exhibit high accuracy, capturing complex interactions and non-linear relationships in high-dimensional datasets. However, they **don't** provide detailed insights into the **spatial distributions** and treat all observations equally.
- Global models are highly sensitive to hyper parameter tuning and optimizations.

Global Models*

GLM: Generalised **linear model**

DRF: Dist. **Random Forest**

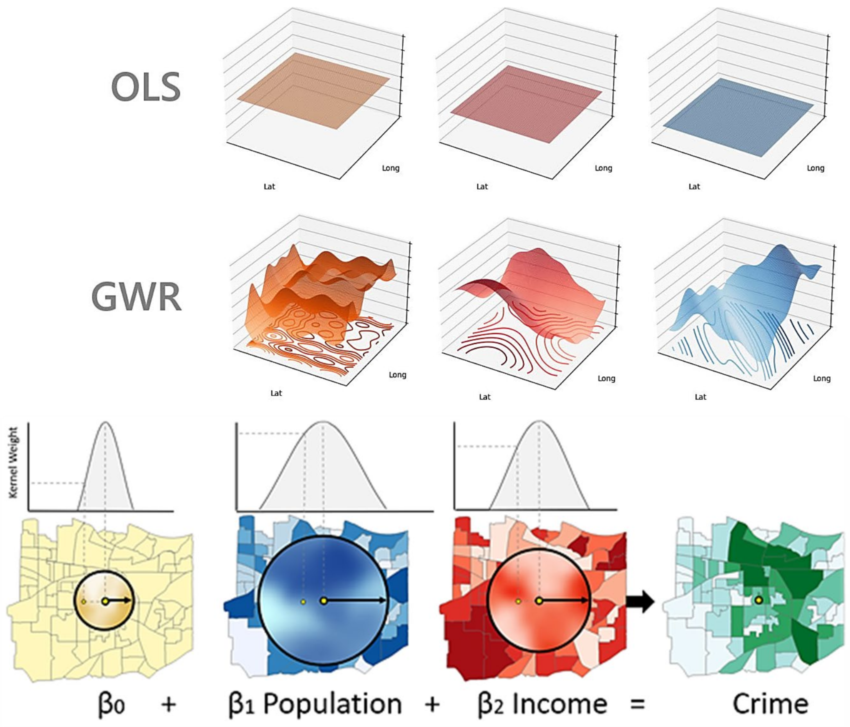
GBM: Gen. **XGBoost**ing model

MLP: Multi-layered Perceptron

Models	GLM	DRF	GBM	MLP
R-squared	0.391945	0.81008	0.901123	0.837036
Mean Squared Error	1720408.4409	537353.6515	279759.2719	461085.2397
Root Mean Squared Error	1311.6434	733.0441	528.9227	679.0325
Mean Absolute Error	597.2122	259.1466	578.1387	213.0738
Mean Residual Deviance	1720408.4409	537353.6515	279759.2719	461085.2397

**H_2O framework - open-source machine learning API*

Geographically Weighted Regression (GWR) model

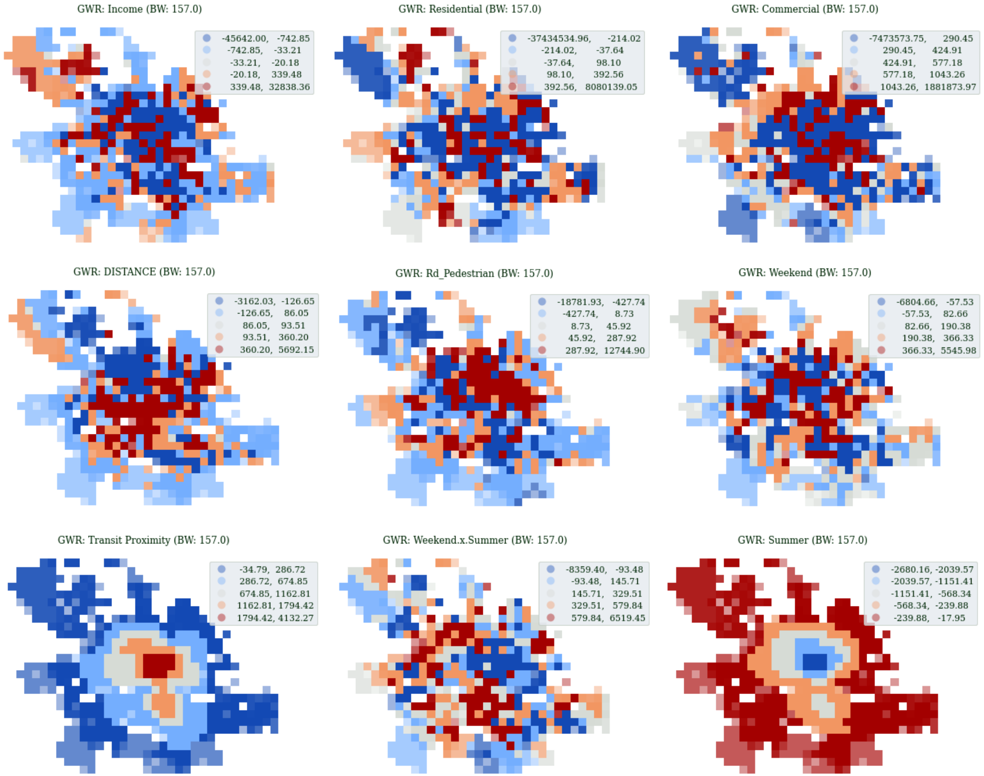



```
Global Regression Results
---------------------------------------------------------------------------
Residual sum of squares:                                     4416499153.393
Log-likelihood:                                                  -25479.975
AIC:                                                              51005.950
AICc:                                                             51008.355
BIC:                                                         4416475408.407
R2:                                                                   0.450
Adj. R2:                                                              0.446

Geographically Weighted Regression (GWR) Results
---------------------------------------------------------------------------
Spatial kernel:                                           Adaptive bisquare
Bandwidth used:                                                     157.000

Diagnostic information
---------------------------------------------------------------------------
Residual sum of squares:                                      558485849.574
Effective number of parameters (trace(S)):                          710.595
Degree of freedom (n - trace(S)):                                  2279.405
Sigma estimate:                                                     494.989
Log-likelihood:                                                  -22388.504
AIC:                                                              46200.200
AICc:                                                             46645.513
BIC:                                                              50471.928
R2:                                                                   0.930
Adjusted R2:                                                          0.909
```

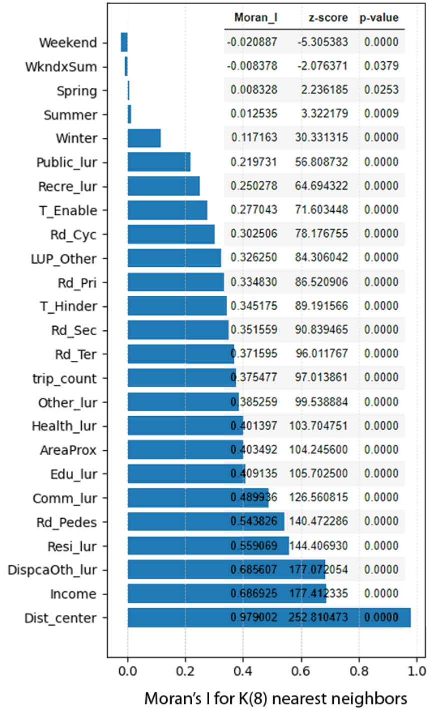
Reference: *https://github.com/parishwadomkar/Usage-analysis-e-scooter/blob/main/Sto_GWR.ipynb*

Geographically Weighted Regression (GWR) Model

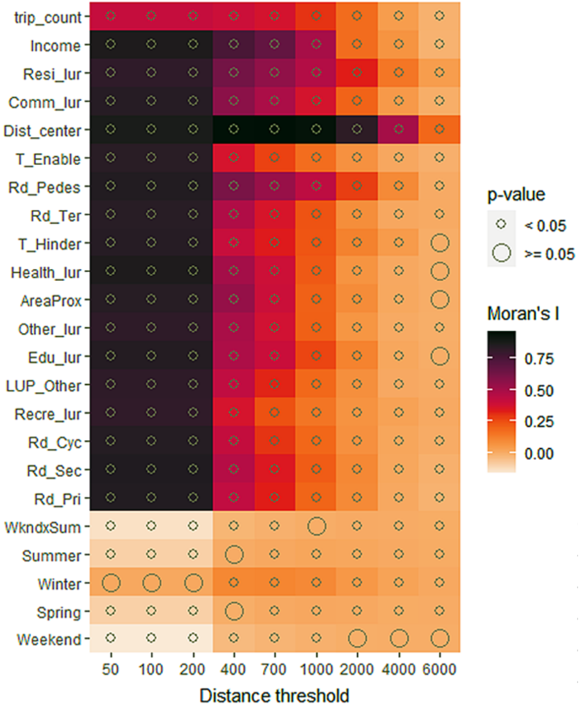
- GWR provides a great improvement in accuracy/model fit, as compared to its linear global model.
- But GWR is a linear model. We need to ensure if there exists a linear relation between the factors and the usage patterns.
- This would also convey the validity for the spatial patterns defined by the feature coefficients.

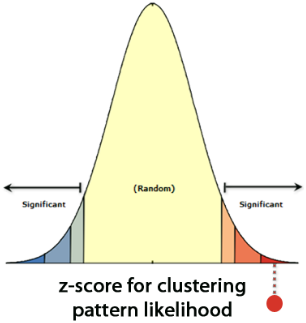


Applicability of Spatial models for analysis

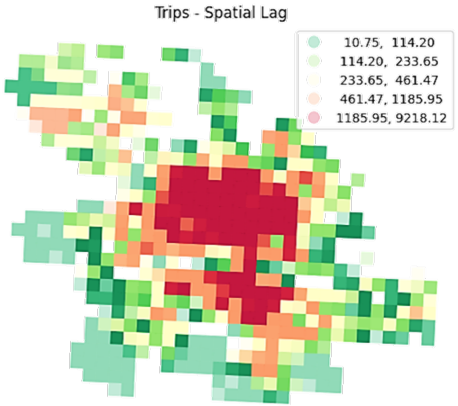


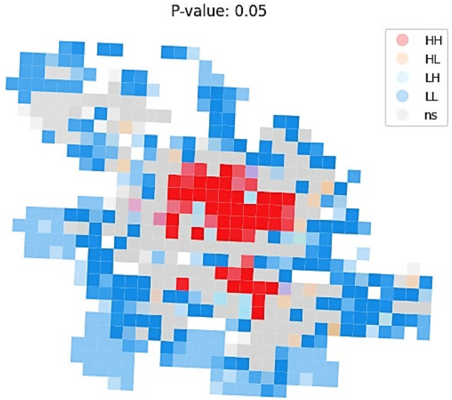
Variable	Moran_I	z-score	p-value
Weekend	-0.020887	-5.305383	0.0000
WkndxSum	-0.008378	-2.076371	0.0379
Spring	0.008328	2.236185	0.0253
Summer	0.012535	3.322179	0.0009
Winter	0.117163	30.331315	0.0000
Public_lur	0.219731	56.808732	0.0000
Recre_lur	0.250278	64.694322	0.0000
T_Enable	0.277043	71.603448	0.0000
Rd_Cyc	0.302506	78.176755	0.0000
LUP_Other	0.326250	84.306042	0.0000
Rd_Pri	0.334830	86.520906	0.0000
T_Hinder	0.345175	89.191566	0.0000
Rd_Sec	0.351559	90.839465	0.0000
Rd_Ter	0.371595	96.011767	0.0000
trip_count	0.375477	97.013861	0.0000
Other_lur	0.385259	99.538884	0.0000
Health_lur	0.401397	103.704751	0.0000
AreaProx	0.403492	104.245600	0.0000
Edu_lur	0.409135	105.702500	0.0000
Comm_lur	0.489936	126.560815	0.0000
Rd_Pedes	0.543826	140.472286	0.0000
Resi_lur	0.559069	144.406930	0.0000
DispcaOth_lur	0.685607	177.072054	0.0000
Income	0.686925	177.412335	0.0000
Dist_center	0.979002	252.810473	0.0000





Moran's Index	0.375477
Expected Index	-0.000335
Variance	0.000015
z-score	97.966011
p-value	0.000000
Input Field:	TRIP_COUNT
Conceptualization:	FIXED_DISTANCE
Distance Method:	EUCLIDEAN
Row Standardization:	True
Distance Threshold:	565.7420 Meters



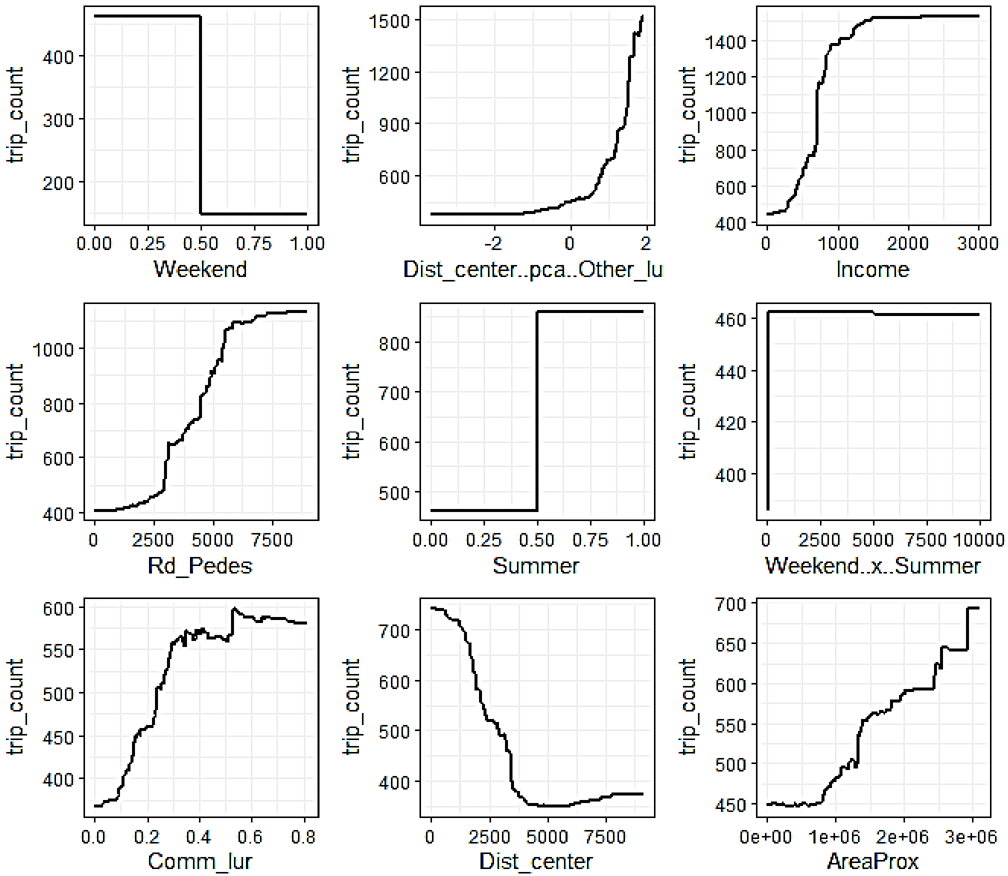


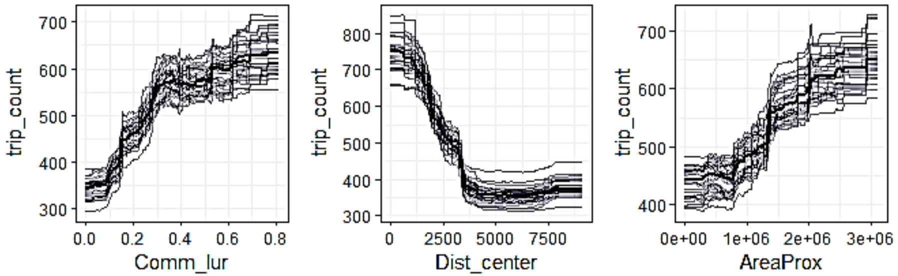
Local Morans I

Variable relation

Response curves or Partial dependence plots (PDP)

These plots help visualize how the predicted outcome changes as a function of one or more independent variables while keeping other variables constant.





Evolving the Random Forest spatially..

1. Bandwidth concept

A positive number that may be an integer in the case of an "adaptive kernel" or a real in the case of a "fixed kernel". In the first case, the integer denotes the number of nearest neighbours, whereas in the latter case the real number refers to the bandwidth.

(Georganos et al. (2019) ***Spatial extension of the random forest algorithm with parameter tuning*** doi:10.1080/10106049.2019.1595177).

Reference: https://github.com/parishwadomkar/Usage-analysis-e-scooter/blob/main/SpatialML_Sto.R

2. Moran's Eigenvector Maps (Dray, Legendre, and Peres-Neto 2006)

The MEMs are calculated by decomposing the Moran's I statistic into a set of orthogonal eigenvectors. These eigenvectors are then used as predictors in the spatial regression model.

(ref: B.M. Benito (2021). ***Easy Spatial Regression with Random Forest***. doi: 10.5281/zenodo.4745208)

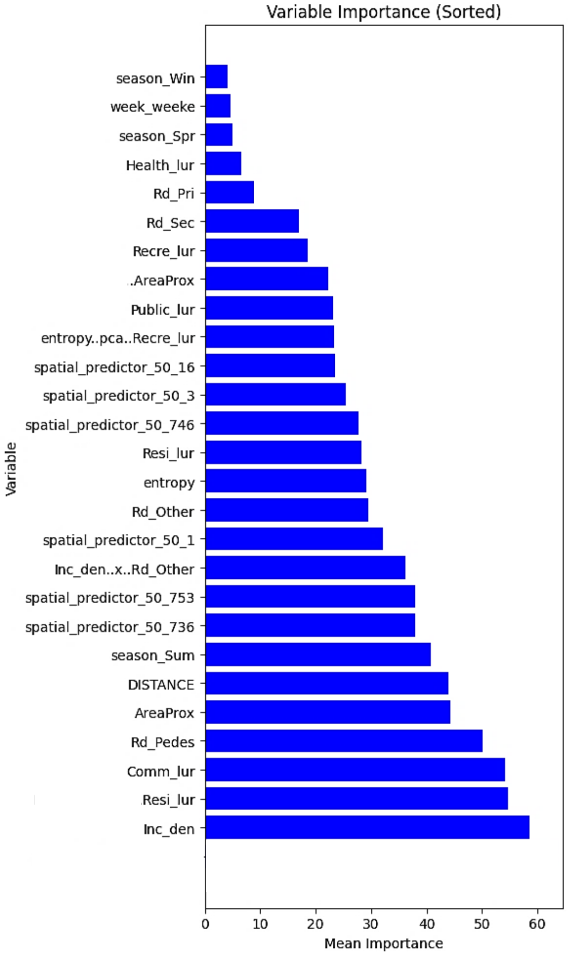
Reference: https://github.com/parishwadomkar/Usage-analysis-e-scooter/blob/main/SpatialRF_Sto.R

SpatialRF Model Performance

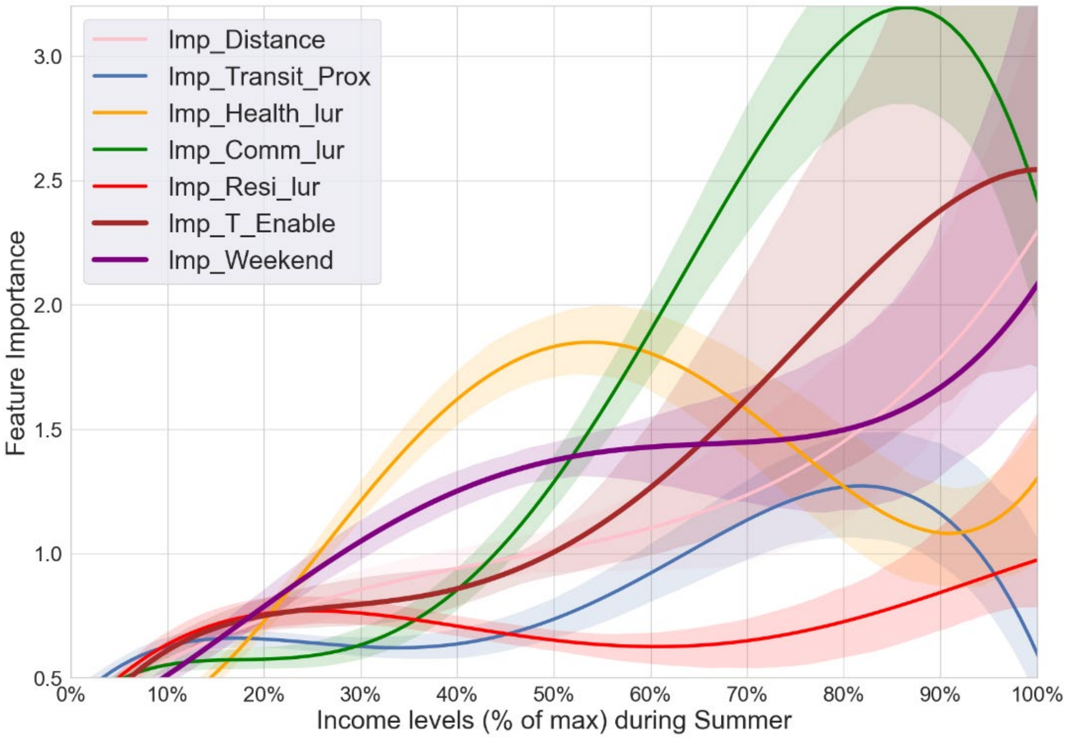
	Global (non-spatial) RF	GW-RF (bandwidth)	Spatial RF (MEMs)
R squared (oob)	0.8578374	0.6694	0.702 +/- 0.0018
R squared	0.9233252	0.8625	0.962 +/- 2e-04
RMSE (oob)	89.96485	192.226	37.098 +/- 0.3722
MSE (oob)	1597.0922	36951.602	1377.200 +/- 0.139
MSE	641.8974	2911.642	341.777 +/- 0.0212
RMSE	25.3357	58.0951	18.479 +/- 0.1458
Predictors	22	22	28

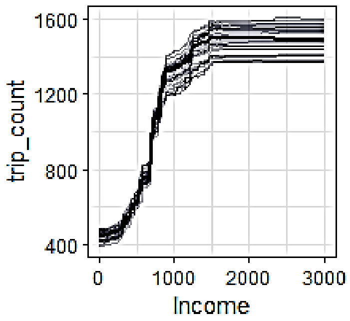
```
"Rd_Pedes"                    "Comm_lur"                    "AreaProx"                   "Resi_lur"
"DISTANCE"                    "season_Sum"                  "Inc_den"                    "entropy"
"Rd_Other"                    "Other_lur"                   "week_weeke"                 "Rd_Sec"
"season_Spr"                  "Recre_lur"                   "Public_lur"                 "LUP_Other"
"Rd_Ter"                      "season_Win"                  "Rd_Pri"                     "Health_lur"
"spatial_predictor_50_43"     "spatial_predictor_50_480"    "spatial_predictor_50_1"     "spatial_predictor_50_34"
```

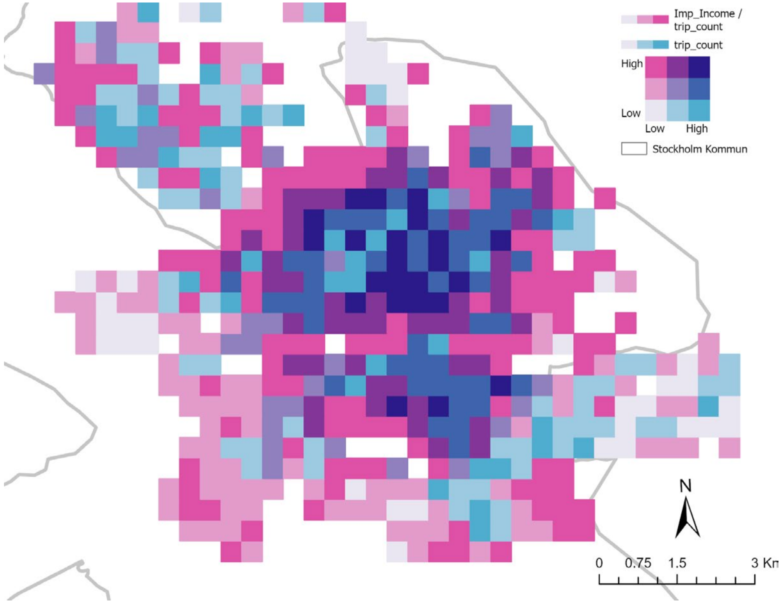
**Spatial_predictor_X_Y, where X is the neighborhood distance at which the predictor has been generated, and Y is the index of the predictor.*



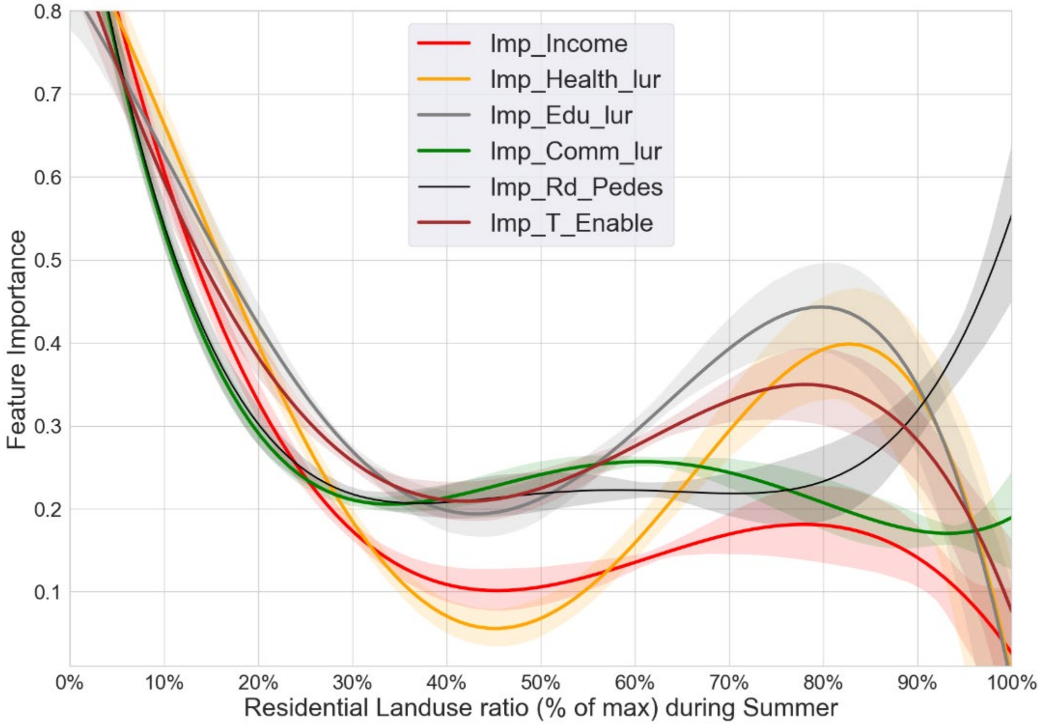
Income levels

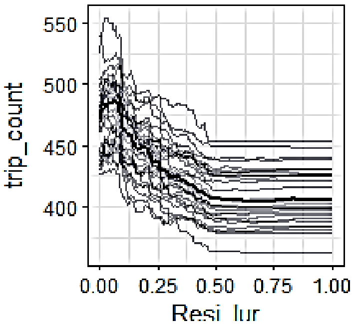


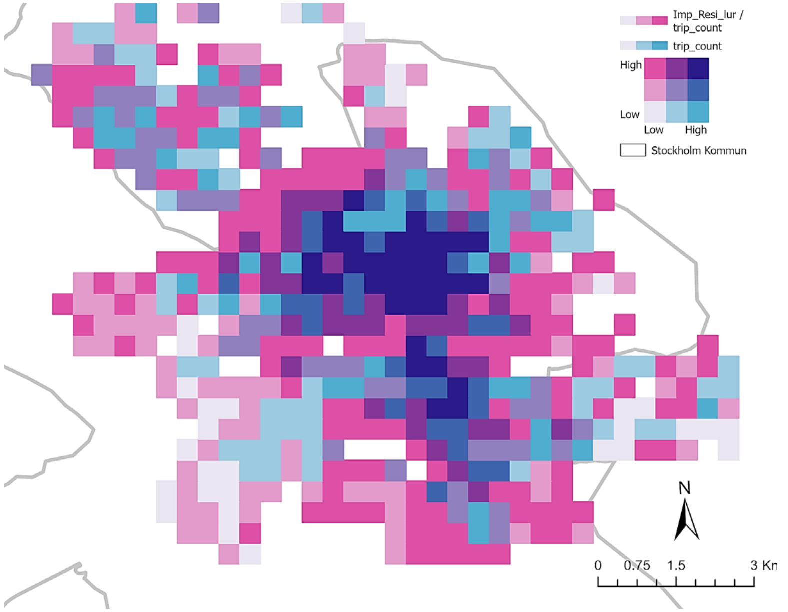




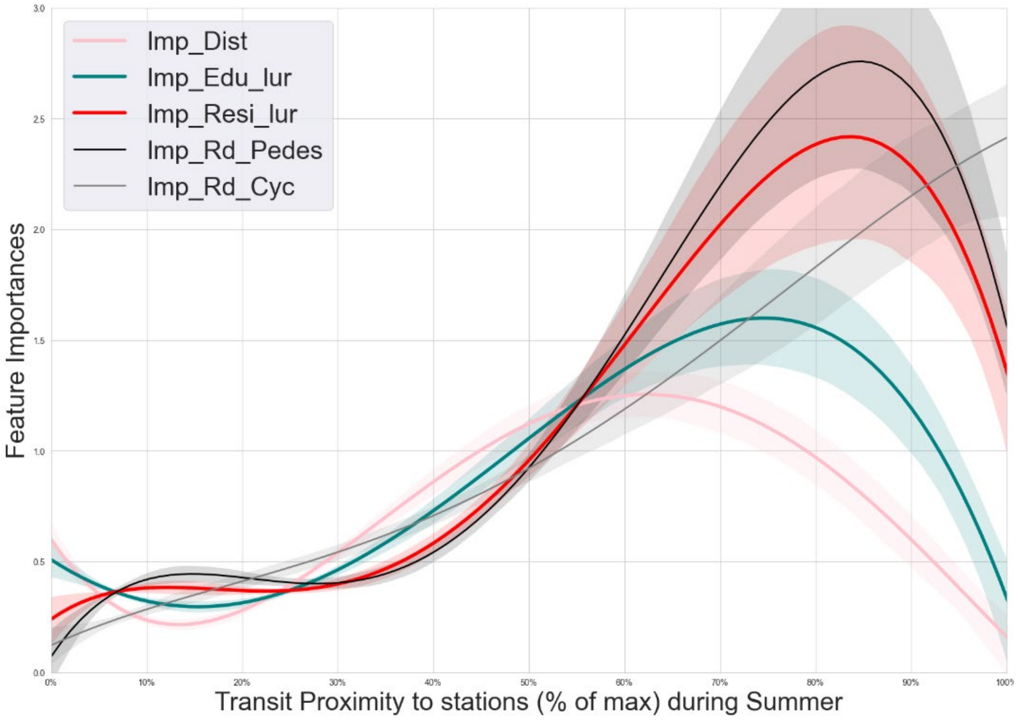
Residential LUR (tdifd)

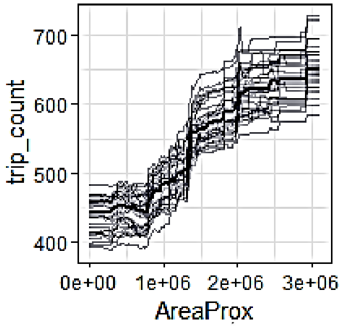


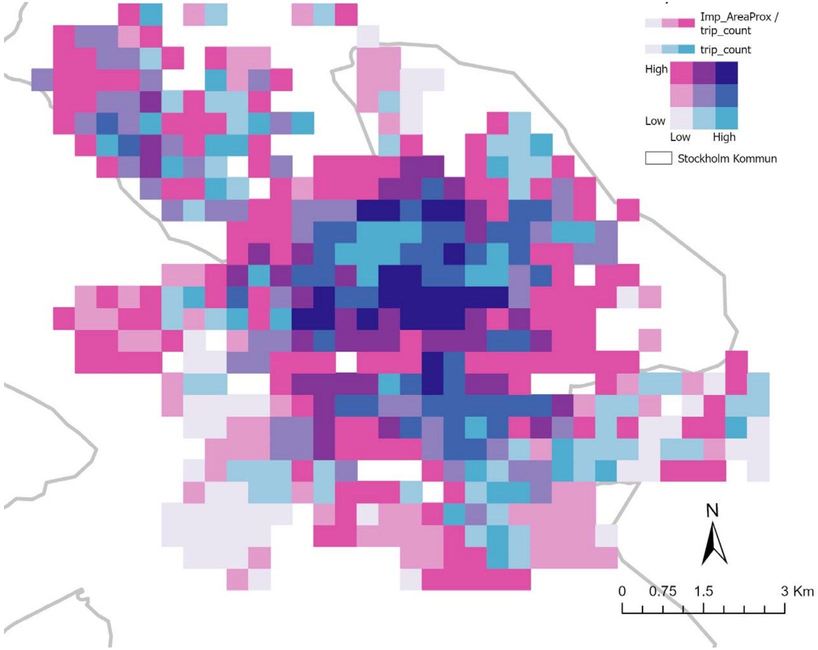




Proximity to transit stations







Summary and way forward

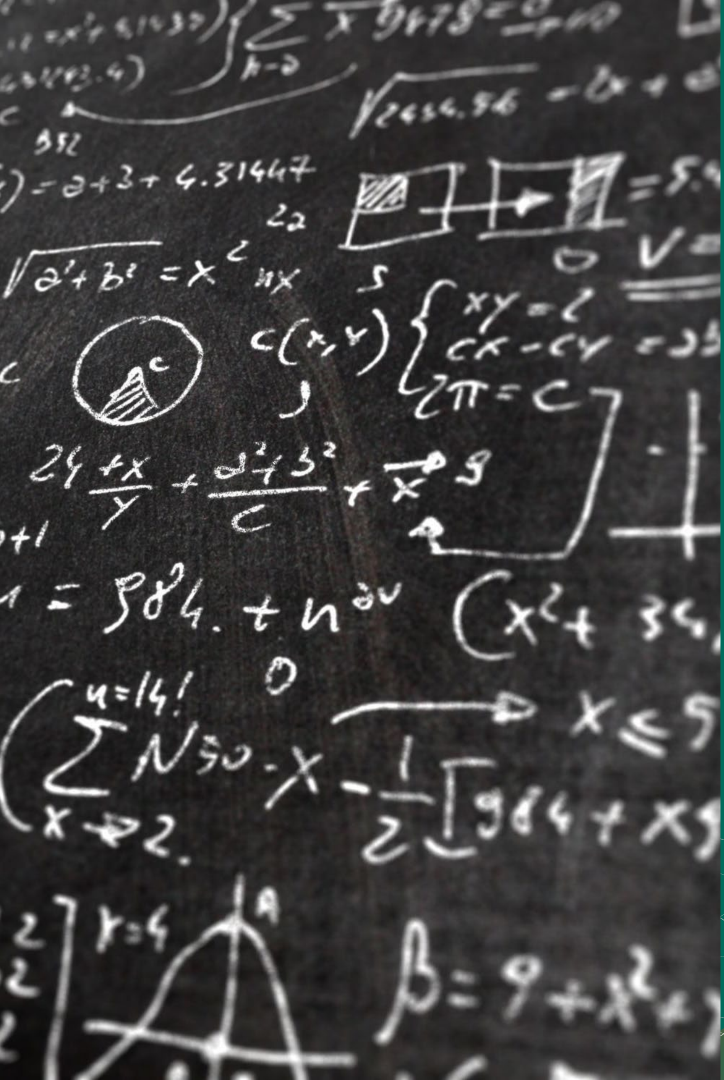
- **Intermodal Mobility Analysis**: Incorporating shared bikes, ride-sharing, and public transport to gain comprehensive insights into the interactions between different transportation modes and base a decision support system to analyze the policy implications.
- **Urban Design**: Developing suitability criteria for urban planning and infrastructure design, with a focus on integrating e-scooters seamlessly into the public transportation ecosystem.
- **Generalized Models**: Creating models that can be applied effectively across diverse urban contexts.

Thanks for Your Attention

Omkar Parishwad

*omkarp@chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY



Notable relevant research..

S. Georganos, T. Grippa, A. N. Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E.Wolff, S. Kalogirou, ***Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling***, Geocarto International 36 (2021) 121–136. doi:10.1080/10106049.2019.1595177

A. Hosseinzadeh, M. Algomaiah, R. Kluger, Z. Li, ***Spatial analysis of shared e-scooter trips***, Journal of Transport Geography 92 (2021). doi:10.1016/j.jtrangeo.2021.103016

Katja Schimohr, Philipp Doebler, J. Scheiner, ***Prediction of Bike-sharing Trip Counts: Comparing Parametric Spatial Regression Models to a Geographically Weighted XGBoost Algorithm***, Geographical Analysis (2022). doi:10.1111/gean.12354

J. Jiao, Junfeng Jiao, S. Bai, Shunhua Bai, ***Understanding the Shared E-scooter Travels in Austin***, TX, ISPRS international journal of geo-information 9 (2020) 135. doi:10.3390/ijgi9020135

S. Quinones, A. Goyal, Z. U. Ahmed, ***Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA.***, Scientific Reports 11 (2021) 6955–6955. doi:10.1038/s41598-021-85381-5