Creating an ML Open-source Tool for Estimating Transit Ridership Based on Network and Operation Data

Jorge M. Diaz-Gutierrez, Helia Mohammadi-Mavi, Andisheh Ranjbari

The Pennsylvania State University

2025 Modeling Mobility Conference

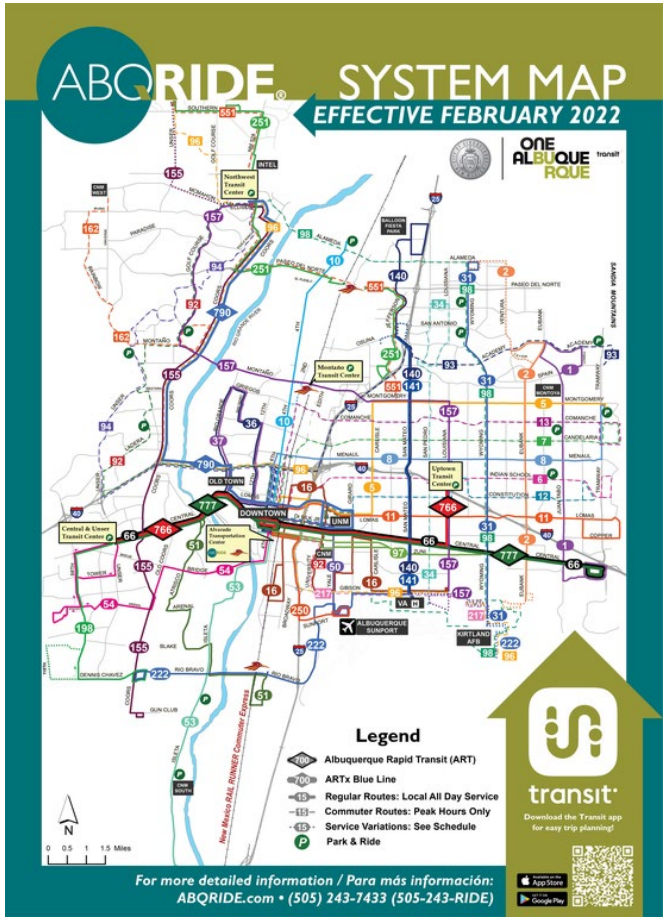
Background and Research Motivation

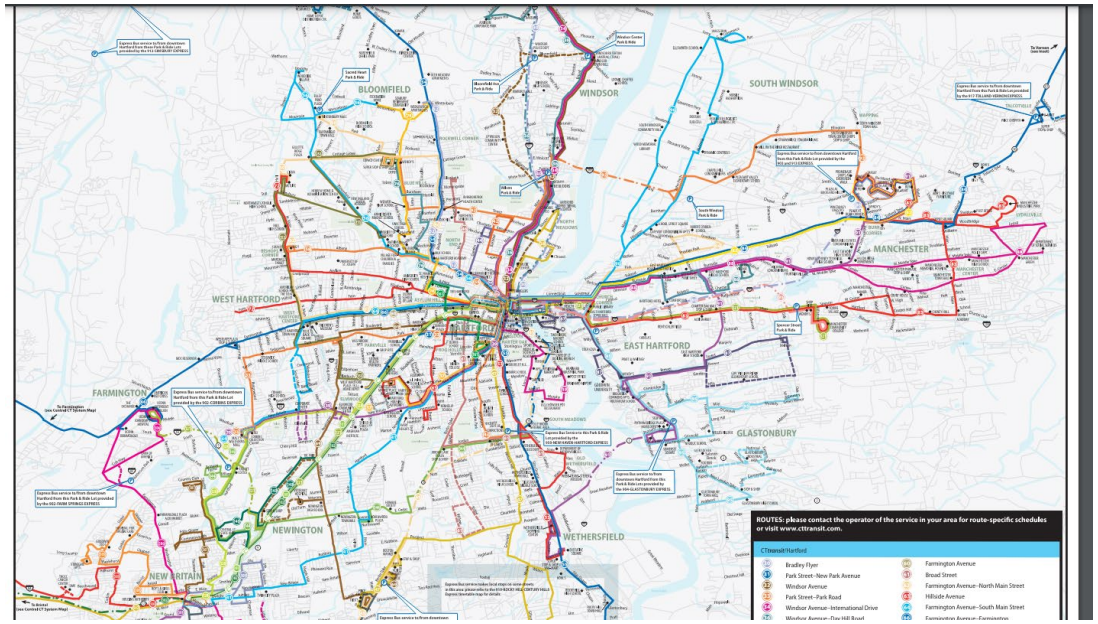
- Transit agencies have **limited resources**
- To efficiently allocate these resources, agencies must analyze the outputs of **ridership demand models**
- Ridership **Machine Learning models (ML)** deemed useful for dealing with large amounts of data

Background and Research Motivation

Benefits of ML models	Limitations of ML models
✓ Offer accurate predictions	× Low interpretability
✓ Manage large datasets	× Overfitting
✓ No manual adjustments	× Lack of generalization
✓ No data assumptions	× Data consuming

Background and Research Motivation





Objective

ML limitations

1. Overfitting and lack of generalization
2. High data consumption

ML problem

1. Predictions on unseen data may lack accuracy
2. ML is limited by funding and data availability

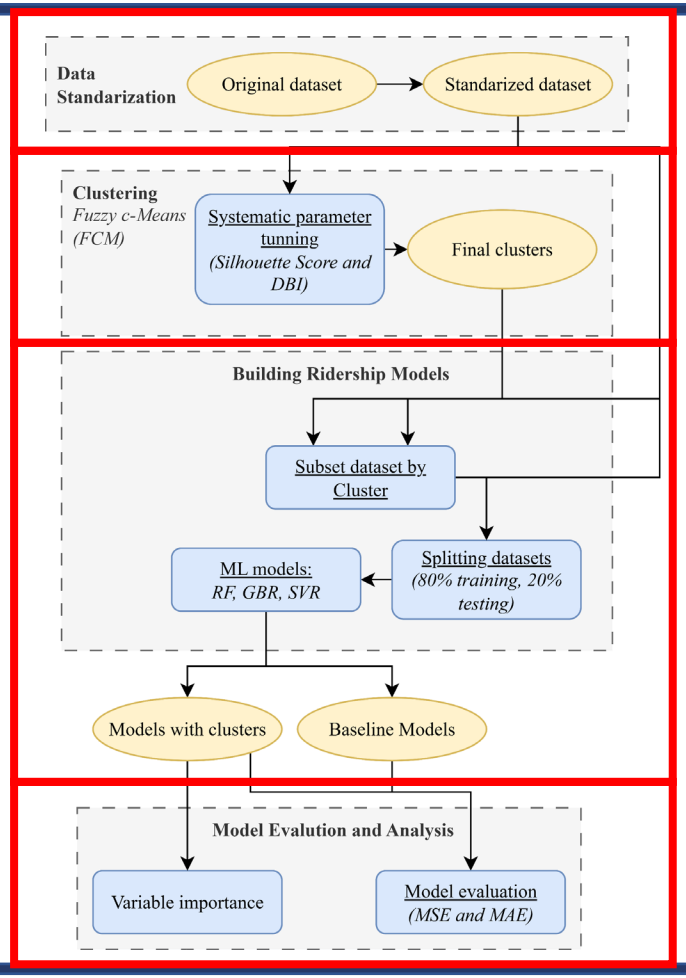
Solutions

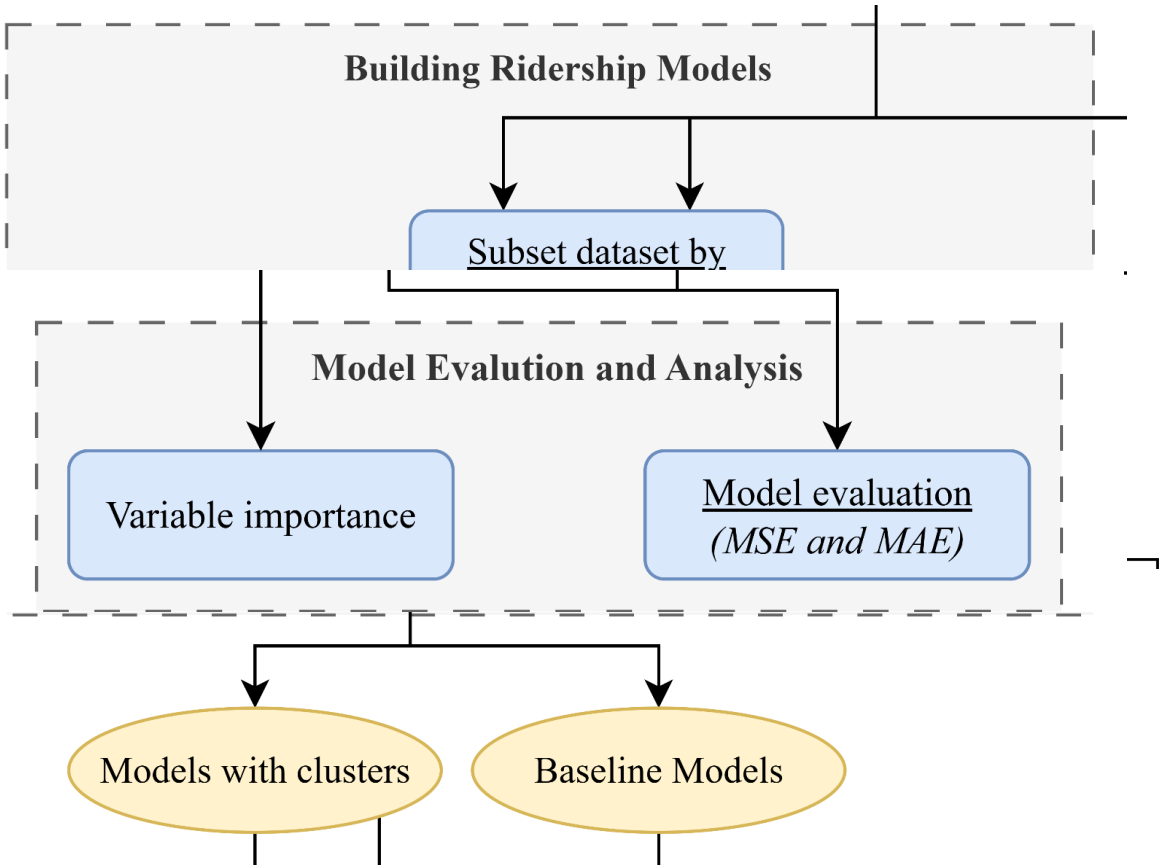
1. Two-Step ML approach: clustering agencies and modeling
2. Create publicly available ML model

Objective

Develop an ML-based open-source tool for estimating transit ridership based on large sets of network and service data

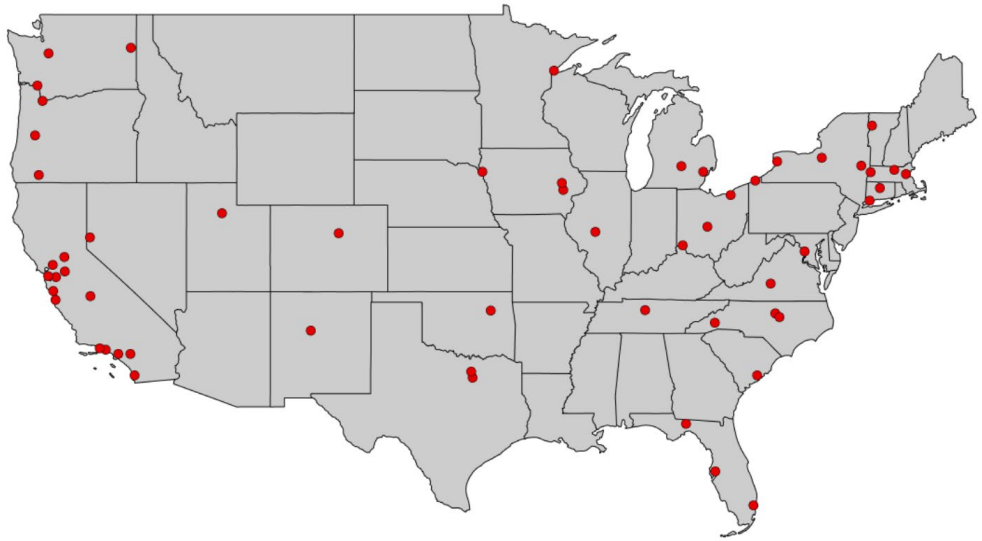
What we did?


Data Standarization
Original dataset
Standarized dataset
Clustering
Fuzzy c-Means (FCM)
Systematic parameter tunning
(Silhouette Score and DBI)
Final clusters
Building Ridership Models
Subset dataset by Cluster
ML models:
RF, GBR, SVR
Splitting datasets
(80% training, 20% testing)
Models with clusters
Baseline Models
Model Evalution and Analysis
Variable importance
Model evaluation
(MSE and MAE)


Building Ridership Models
Subset dataset by
Model Evalution and Analysis
Variable importance
Model evaluation
(MSE and MAE)
Models with clusters
Baseline Models

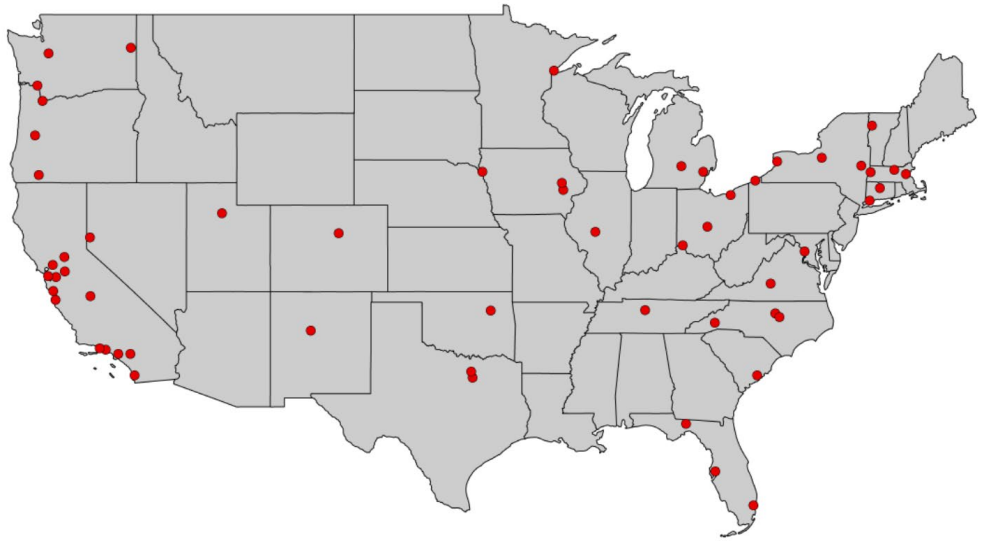
Data Sources

- 2022-2023 data from multiple sources
- 64 transit agencies, 69 variables, and 1,754 monthly observations
- 3 public data sources

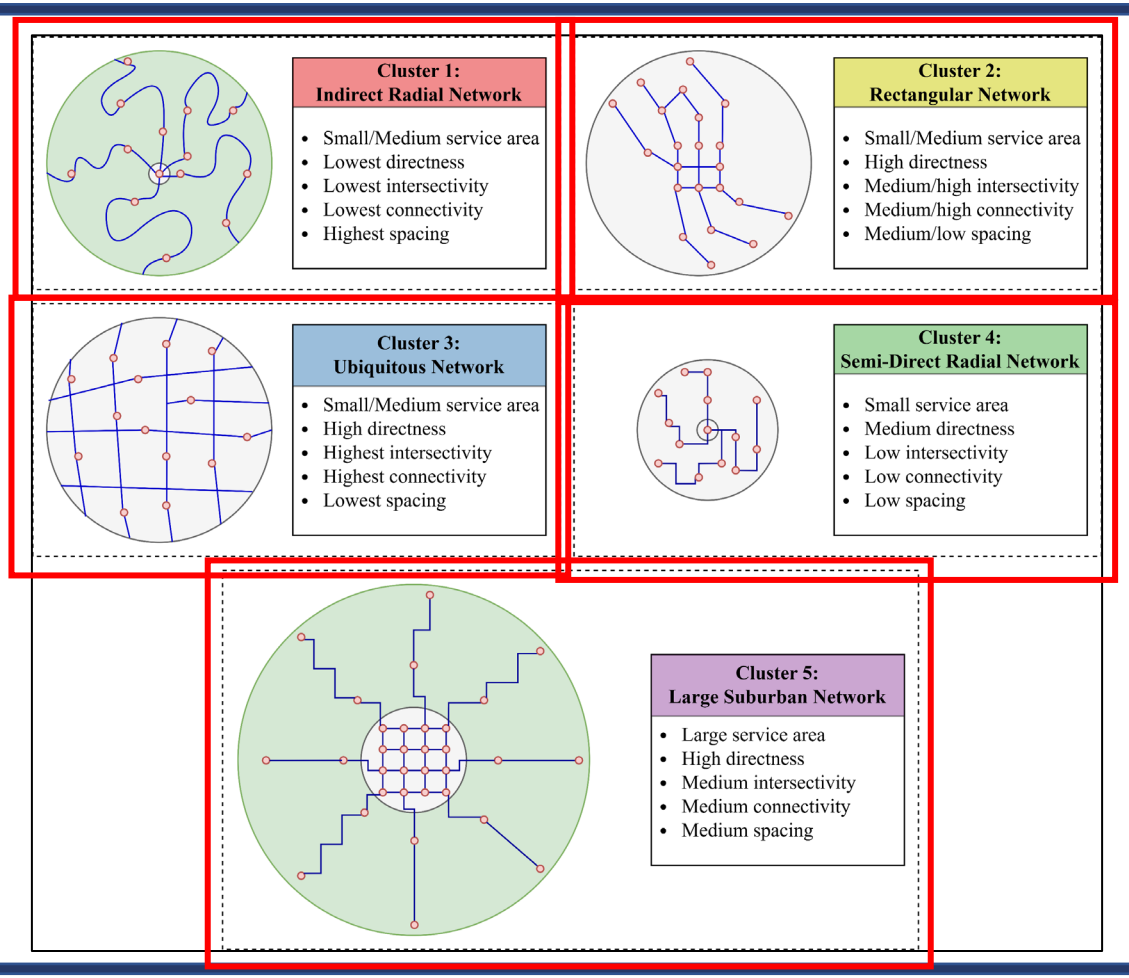


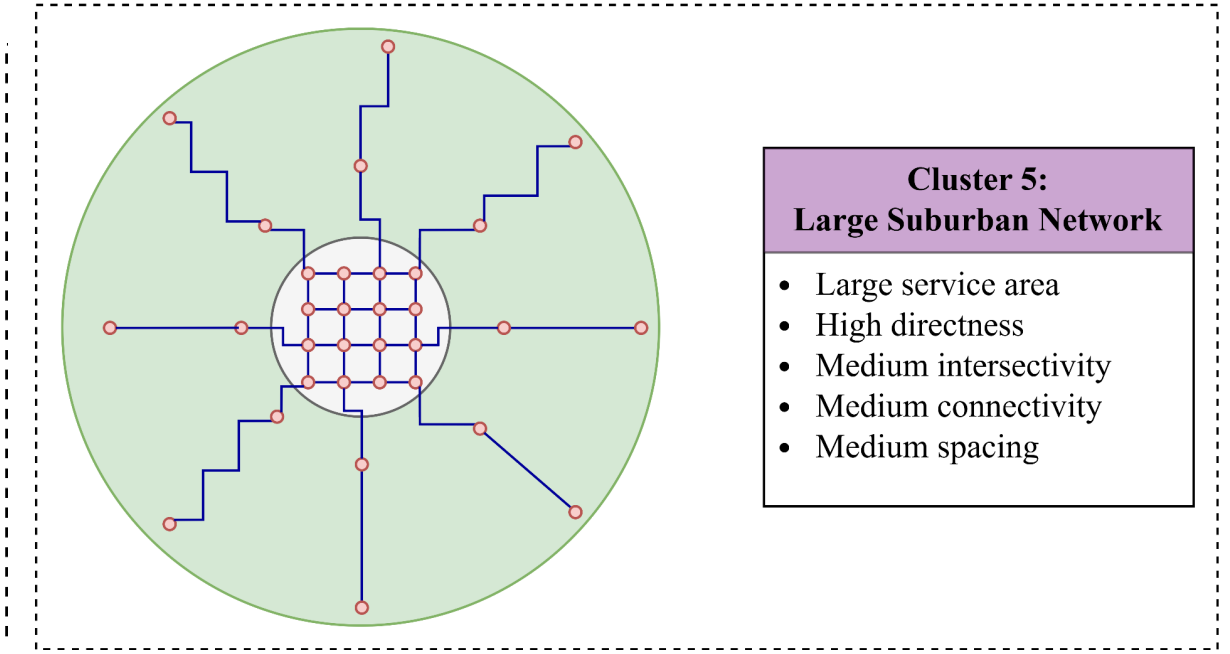
Data Sources

- **General Transit Feed Specification**: Stop spacing, coordinates, frequencies and speeds:
  - Intersectivity, connectivity, and directness
- **National Transit Database**: UPT, VRM, VRH, VOMS, fare, service area
- **5-year American Community Service**: Sociodemographic data

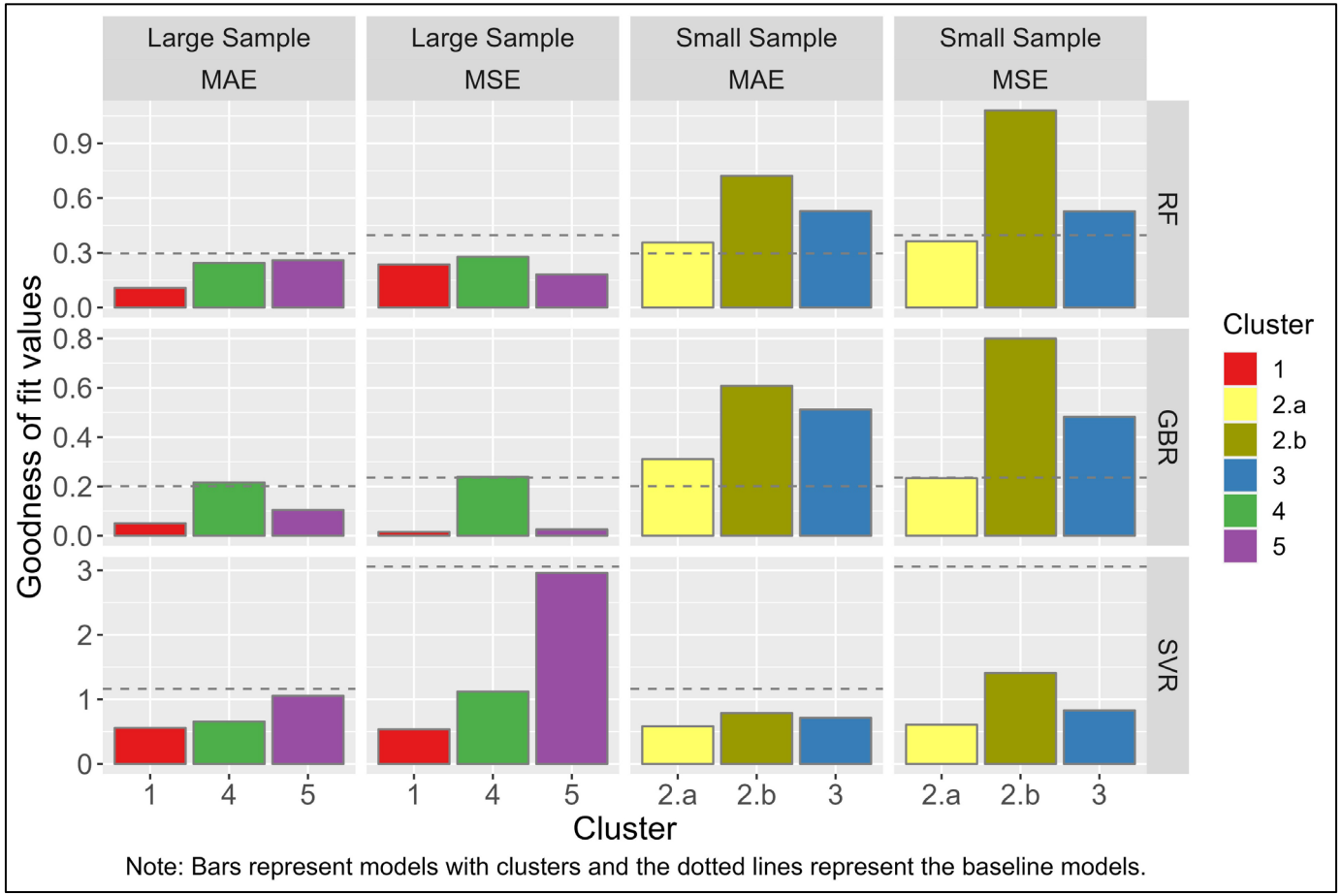


Clustering results



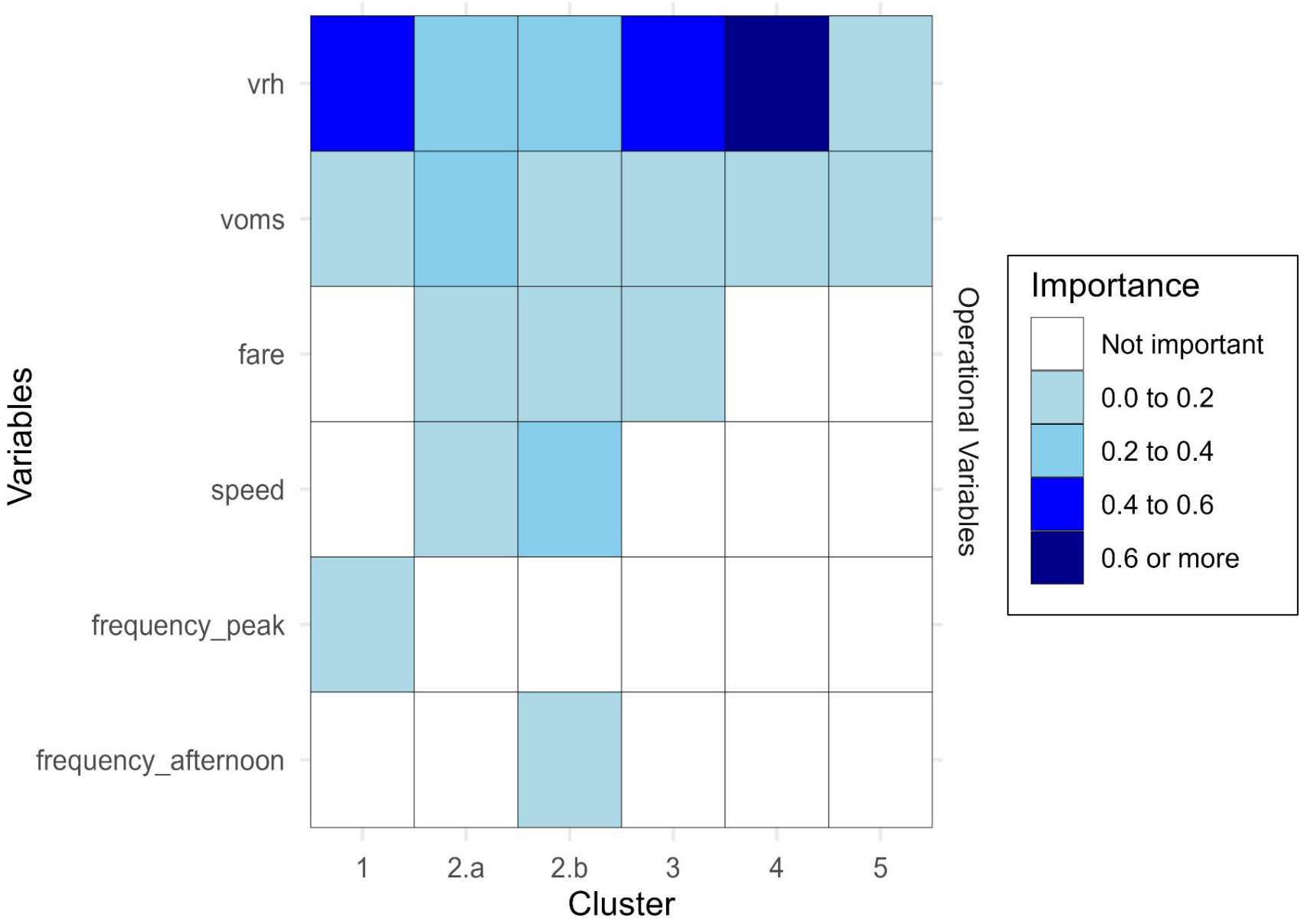


Evaluation of the ML Models



Note: Bars represent models with clusters and the dotted lines represent the baseline models.

Variable Importance



Conclusions

- The results showed acceptable to very high performance values for all the models with clusters compared to those without clusters.
- The variable importance results corroborated that different clusters showed distinct important predictors for ridership.
- Our models are publicly available on GitHUB.
- Transit agencies can use this tool by simply inputting their network and service data