A Data-Driven Approach to Enhancing Gravity Models for Trip Demand Prediction

Kamal Acharya*, Mehul Lad*, Liang Sun#, Houbing Song*

*University of Maryland Baltimore County, #Baylor University

April 30, 2026

Contents

- Trip demand prediction crucial for transportation planning
- Gravity model widely used but limited in complexity handling
- Need for integration of non-linear factors (geographical, economic, social)
- Objective: Enhance traditional gravity models using machine learning and data-driven approach

The key contributions of this paper are:

- First, we integrate a diverse array of datasets to enhance the traditional gravity model's ability to capture complex, non-linear relationships between variables.
- Second, we apply machine learning techniques to improve the accuracy and scalability of trip demand predictions.
- Third, we validate the enhanced model with real-world data from counties in TN and NY.

The traditional gravity model is expressed as:

$$T_{ij} = \frac{P_i \cdot P_j}{d_{ij}^{\beta}} \tag{1}$$

where

- T_{ij} is the interaction between the locations i and j,
- P_i and P_j represent the "population" of the two locations,
- d_{ij} is the distance between them, and
- β is the distance decay parameter, which reflects how interaction decreases with increasing distance.

The modified equation for gravity is as follows:

$$T_{ij} = k \frac{P_i^{\lambda} \cdot P_j^{\alpha}}{d_{ij}^{\beta}} \tag{2}$$

where k is a scaling constant that adjusts the model for different magnitudes of flow, such as daily or monthly movements.

In the expanded model, spatial interaction (T_{ij}) is formulated as a function of three distinct vectors:

$$T_{ij} = f(O_i, D_j, S_{ij}) \tag{3}$$

where:

- O_i: A vector of origin attributes, representing the characteristics of the originating location (e.g., population, economic output, or accessibility);
- D_j: A vector of destination attributes, representing the characteristics of the destination location (e.g., attractiveness, size, or infrastructure);
- S_{ij}: A vector of separation attributes, capturing the effects of spatial separation, such as distance, travel costs, and intervening opportunities.

Study	Methodology	Key Findings
[1]	Traditional Gravity Model $$T_{ij} = \frac{P_i \cdot P_j}{d_{ij}^2}$$	Basic trip distribution modeling
[2]	Modified Gravity Model $$T_{ij} = k \frac{P_i^{\lambda} \cdot P_j^{\alpha}}{d_{ij}^{\beta}}$$	Incorporation of income and infrastructure effects
[3]	Deep Gravity Model (Use Deep Learning Models) $$T_{ij} = f(O_i, D_j, S_{ij})$$	Geographic data-based mobility flow generation

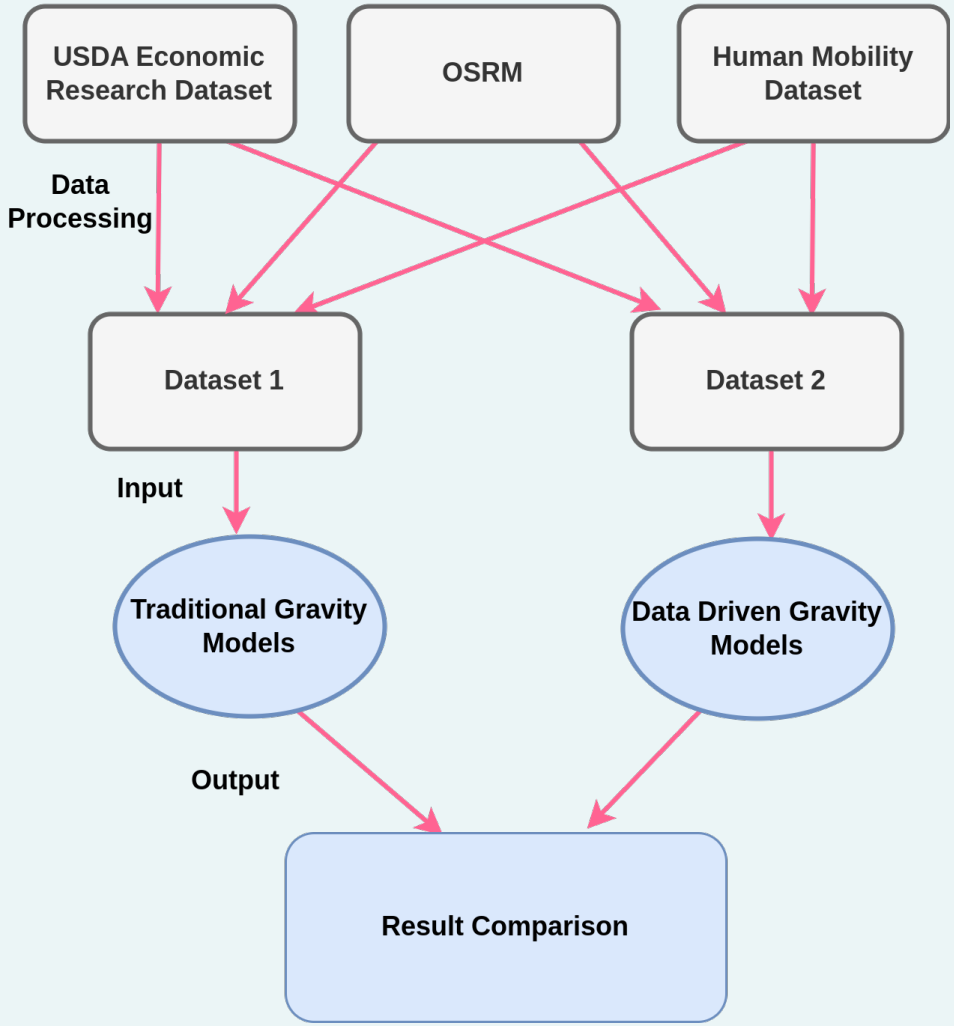

Figure: Research Process

- Two states selected: Tennessee (TN) and New York (NY)
- TN: Diverse geographic/demographic composition
- NY: High-density urban environments
- County-level analysis for stability and data consistency

Table: Datasets Used in the Research

Name	**Source**
Mobility Dataset	Multiscale Dynamic Human Mobility Flow Dataset
Population, Economic and Education Dataset	USDA Economic Research Service
Geographic Features, Distance and Time Between Counties	Open Source Routing Machine (OSRM)

Table: Description of Feature Categories

Category (Numbers)	Description
Land Use Counts (7)	F1: Natural, F2: Agricultural, F3: Residential, F4: Commercial, F5: Public, F6: Industrial, and F7: Military purposes.
Points of Interest (8)	F8: Education, F9: Commerce, F10: Public Services, F11: Healthcare, F12: Recreation, F13: Heritage, F14:Transport, and F15: Miscellaneous.
Roads (3)	F16: Highways, F17: Roadways, and F18: Streets.
Terminals (1)	F19: Includes the total count of airports, railway stations, bus stations, etc.
Structures (1)	F20: Total building count within each county.
Economic Features (5)	F21: Unemployment percentage rate, F22: Median Household Income, F23: % of State Median Household Income, F24: Percentage of people in poverty, and F25: Percentage of children (0-17) in poverty.
Education (1)	F26: Includes percentage of population completing college.
Population (1)	F27: Provides population count for each county.

We compile the two sets of datasets by merging the mobility dataset with the others.

- Dataset 1 is based on the original gravity model input, which only considers the population and distance of origin and destination counties as the input, along with the time of travel between the two counties.
- Dataset 2 contains the other geographical, economic and social features along with the previous features.

- Random Forest: Ensemble learning, optimized via RandomizedSearchCV
- Deep Neural Network: Five-layer architecture, optimized with Optuna
- Gradient Boosting: Ensemble boosting, hyperparameter tuning

- **R-squared (R^2)**: Variance explanation capability
- **Mean Absolute Error (MAE)**: Prediction accuracy
- **Common Part of Commuters (CPC)**: Reliability of flow predictions

Table: Comparison of Gravity Models for TN and NY with Percentage Improvement

Model	Metric	TN			NY		
		Traditional	Data-Driven	% Improved	Traditional	Data-Driven	% Improved
NNs	*MAE*	0.1026	0.0622	**39.38%**	0.0879	0.0320	**63.59%**
	R^2	0.8274	0.9406	**13.68%**	0.6444	0.9762	**51.48%**
	CPC	0.7206	0.8412	**16.73%**	0.6295	0.9085	**44.32%**
RF	*MAE*	0.0420	0.0402	4.29%	0.0255	0.0208	18.43%
	R^2	0.9746	0.9781	0.36%	0.9755	0.9874	1.22%
	CPC	0.9121	0.9169	0.53%	0.9291	0.9456	1.78%
GB	*MAE*	0.1070	0.0957	10.56%	0.0620	0.0527	15%
	R^2	0.8814	0.9180	4.15%	0.9395	0.9654	2.75%
	CPC	0.7554	0.7894	4.50%	0.8231	0.8631	4.86%

Metric	TN Improvement	NY Improvement	Best Model
MAE	39.38%	63.59%	Neural Networks
R^2	13.68%	51.48%	Neural Networks
CPC	16.73%	44.32%	Neural Networks

Table: Percentage improvement by data-driven models

Table: Top 10 Features for Models in TN and NY

State	Model	Ranked Features (1–5)	Ranked Features (6–10)
TN	Neural Networks	Distance, Time, F27-D, F27-O, F26-D	F4-O, F4-D, F25-O, F17-D, F20-D
	Random Forest	Time, F11-D, F14-O, F26-D, F27-D	F16-D, Distance, F18-O, F5-O, F8-O
	Gradient Boosting	Time, F26-D, F16-D, F27-D, F14-O	F11-D, F17-D, Distance, F8-O, F18-O
NY	Neural Networks	Time, Distance, F12-D, F5-D, F20-O	F5-O, F10-O, F8-D, F10-D, F14-D
	Random Forest	Time, Distance, F22-D, F8-D, F21-D	F26-D, F13-D, F23-D, F8-O, F19-D
	Gradient Boosting	Time, Distance, F26-D, F8-D, F21-D	F20-O, F23-D, F16-O, F16-D, F4-D

- Key features: Distance, travel time, population, education, public services
- Geographical and educational factors important in TN
- Public services, transportation facilities important in NY

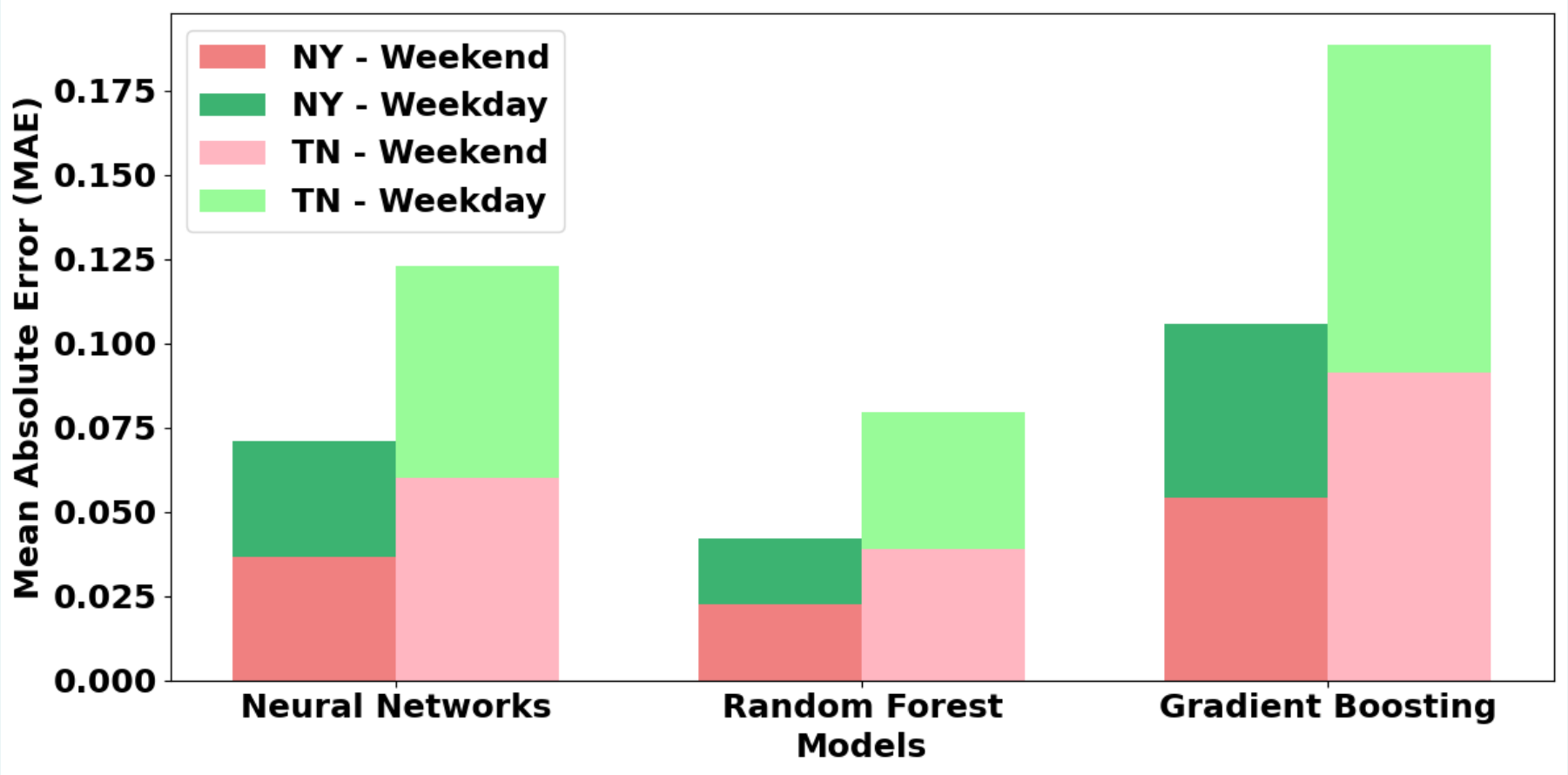

Figure: MAE Comparison of Models for NY and TN by Days

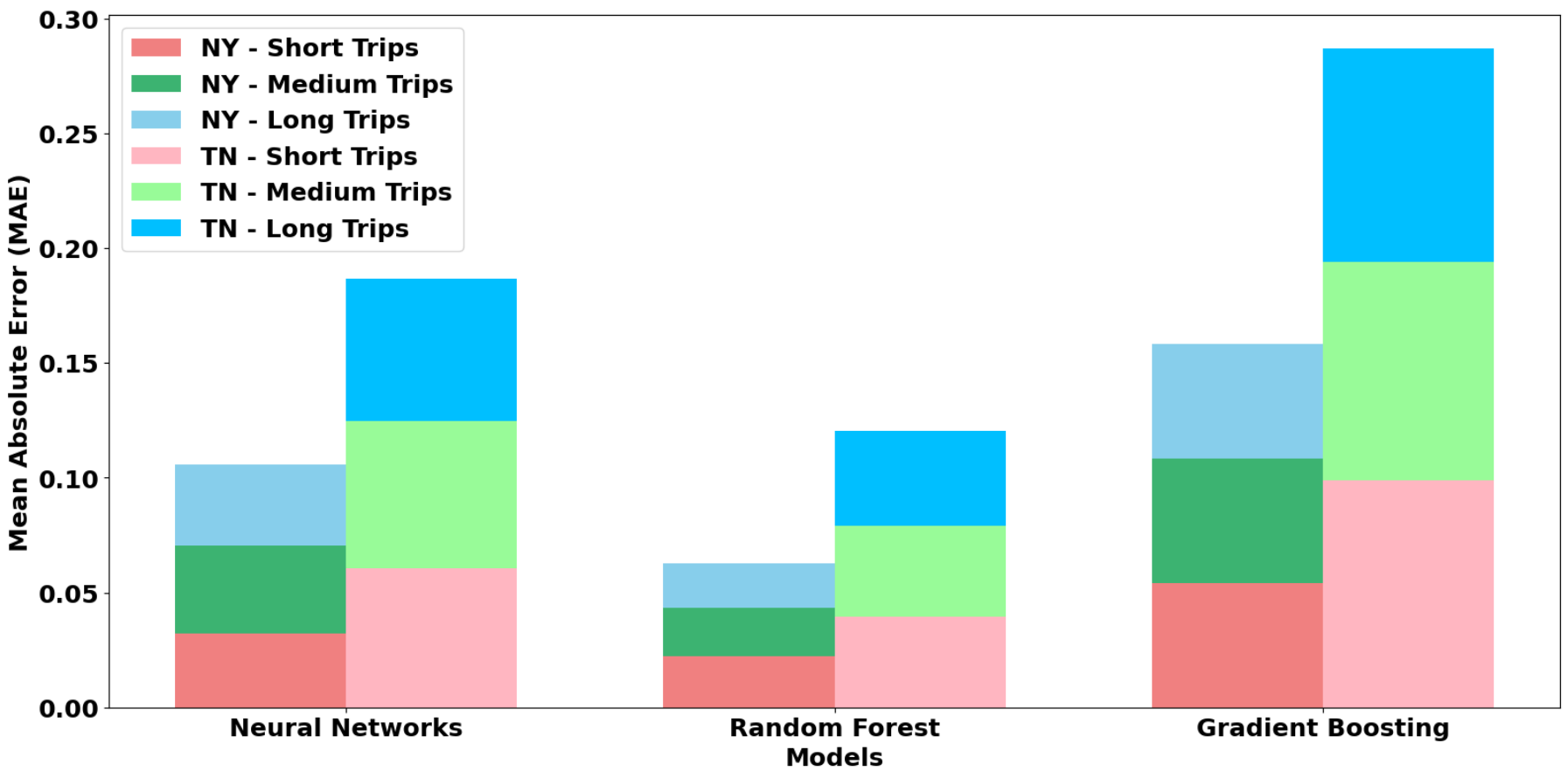

Figure: MAE Comparison of Models for NY and TN by Trip Distance

Conclusion

- Machine learning significantly enhances gravity model
- Data-driven models yield 4%-63% MAE improvement
- NNs best in urban areas; Random Forest versatile across regions
- Future direction: Incorporate real-time GPS and social media data

This material is based upon work supported by the NASA Aeronautics Research Mission Directorate (ARMD) University Leadership Initiative (ULI) under cooperative agreement number 80NSSC23M0059. This research was also partially supported by the U.S. National Science Foundation through Grant No. 2317117 and Grant No. 2309760.

S. Erlander and N. F. Stewart, *The gravity model in transportation analysis: theory and extensions*, vol. 3. Vsp, 1990.

W.-S. Jung, F. Wang, and H. E. Stanley, "Gravity model in the korean highway," *Europhysics Letters*, vol. 81, no. 4, p. 48005, 2008.

F. Simini, G. Barlacchi, M. Luca, and L. Pappalardo, "A deep gravity model for mobility flows generation," *Nature communications*, vol. 12, no. 1, p. 6576, 2021.

Thank You