Exploring Commuting Inequalities through Inferential Machine Learning: Lessons for Spatial Mismatch Hypothesis

Subham Kharel[1], Soheil Sharifi[2], Lijuan Tang[3], Jiangling Li[4], Qisheng Pan[4,5]

[1] Senior Data and Analytics Planner, Lehigh Valley Planning Commission, Allentown, Pennsylvania

[2] Transportation data modeler, Houston Galveston Area Council, Houston, Texas

[3] Masters Student, Division of Data Science, College of Engineering, The University of Texas at Arlington, Texas

[4] Professor, Department of Public Affairs and Planning, The University of Texas at Arlington, Texas

[5] Director, Center for Transportation, Equity, Decisions and Dollars (CTEDD), Professor, Department of Public Affairs and Planning, The University of Texas at Arlington, Texas

THE UNIVERSITY OF TEXAS AT ARLINGTON

CENTER FOR TRANSPORTATION EQUITY, DECISIONS & DOLLARS

Introduction and Research Background (Part – 1)

- In U.S. cities, **residential segregation** and **suburban employment decentralization** adversely impacted **job prospects for racial minorities,** including African Americans in inner cities (Kain 1968).

- As an extension of Kain's observation, scholars later formalized this concept into the **Spatial Mismatch Hypothesis (SMH)**, which has since been widely studied, with **job accessibility** serving as a key performance indicator (Blumenberg & Manville, 2004).

- However, job accessibility **overlooks critical spatial disparities** (Shen, 1998; Cervero et al., 1995) and its **measurements drop to zero beyond the transit catchment**, failing to capture broader spatial mismatch challenges (Kharel et al., 2024; Sharifiasl et al., 2023).

- **Commuting distances** show more **nuanced inequity** patterns. For instance, studies reported **disparate commute distances** with significant disparities **between high- and low-income workers** (Blumenberg & King, 2019) and between **car users and public transit riders** (Kawabata & Shen, 2007).

- **Rising housing costs** have pushed **low-income households to suburbs** with limited transit, exacerbating transportation burdens for **families that own cars but cannot offset related financial burdens** (Allen & Farber, 2019; Blumenberg & King, 2019; Benner & Karner, 2016).

- Furthermore, many **studies overlook the implications of skill and income-based mismatches on commuting patterns**, which could provide deeper insights into how spatial mismatch disproportionately affects different population groups (Sharifiasl et al., 2023; Kharel et al., 2024).

Introduction and Research Background (Part - 2)

- The SMH has been examined using **various modeling approaches**, from **linear regression to spatial econometric models,** primarily focusing on metropolitan-scale analyses while often overlooking subregional variations (Grengs, 2010; Shen, 2000).

- The compounded challenges arising from the intersection of various factors at different locations within a region suggest that **spatial mismatch may not follow a simple linear pattern**. Yang et al. (2023) in Chengdu (China) **confirmed that spatial mismatch is non-linear**, though they **did not account for housing affordability in their analysis**.

- Kawabata and Shen (2007) in their study of **San Francisco** and Bautista-Hernández (2020) in their study of **Mexico City** both used **Spatial autoregressive models** to highlight significant disparities in commute times between transit and car users, revealing **localized patterns of inequality**, particularly affecting transit-dependent populations.

- Blumenberg and Siddiq (2023) and Blumenberg and Wander (2023) applied **spatial panel models** to analyze job-housing fit and its influence on commute distances, uncovering **significant regional differences driven by housing costs**.

- Hu (2015) and Antipova (2020) used **descriptive and ANOVA techniques** to demonstrate that **low-income workers tend to commute longer distances than higher-income workers**, underscoring income disparities in spatial mismatch.

- Gaps remain in understanding the complex, non-linear interactions between **housing costs, socio-demographic factors, the built environment, and commuting burdens in the U.S. context**. These gaps can be addressed using emerging AI-driven methodologies.

Research Questions

Given the focus of this study on **adopting AI-driven methodologies to explore spatial mismatch**, we propose the following three research questions:

- **How does a Random Forest model compare to the Random Effects models in predicting commute distances and spatial mismatch?**
- **How do various socioeconomic and built environment factors affect differently the commute distances for high-, medium-, and low-income workers?**
- What are the **implications of the findings from AI-driven approaches** for addressing spatial and socioeconomic disparities in urban planning?

Data Sources

- Socioeconomic Data: **American Community Survey 5-yr datasets** (2007-2011, 2012-2017, 2017-2021)
- Commuting Distance Variables: **Longitudinal Employer-Household Dynamics (LEHD) Dataset** (2010, 2015, 2020)

Methodology

1. Subregion Classification

- ✓ Use **Contiguity constraint** to establish initial boundaries for candidate central city BGs.
- ✓ Use **population density** and **pre-1970 housing density thresholds** to assess similarity with arbitrary thresholds (90% CI).
- ✓ Find **optimal cutoff thresholds** by maximizing similarity to arbitrary threshold areas based on the closest z-score (Liu et al., 2019).
- ✓ Use **Median Housing Age data** - identify cutoff scores via z-scores, distinguishing true city centers from candidate centers.
- ✓ Classify remaining candidate centers as **inner-ring suburbs**.

2. Commute Distance Measurement

$$ACD_a = \frac{1}{N}\sum_{j=1}^{N} {D_{ij}}^a$$

ACD_a = Avg. Commute Distance for income group 'a', ${D_{ij}}^a$ = Distance between i and j, N = total count of block groups

Note: ***Here, residential and workplace characteristics must be separated by income groups***

3. Housing Unaffordability Measurement

$$Housing\ Cost\ (Unaffordablility) = \frac{Mortgage\ Cost * OO}{OO + RO} + \frac{Median\ Gross\ Rent * RO}{OO + RO}$$

OO and RO = Owner Occupied and Renter Occupied Housing Units

4. Random Forest Regressor

$$y^x = \frac{1}{N}\sum_{n=1}^{N} y_n(x)$$

$y\hat{}(x)$ = Predicted output for instance x, N = total trees in the forest, $y_n(x)$ = Prediction of the n-th tree for input x

Note: ***SHapley Additive exPlanations (SHAP) plots are used to make the model more interpretable***

Local Interpretable Model-agnostic Explanations (LIME) are also developed and expanded for better interpretability

Dependent and Independent Variables

Name	Code	Variable Type	Relationship with DV	Mean	Median	SD
Commute Distance for Low-income Workers (kms)	se01	N/A	N/A	58.11766	54.88875	35.69589
Commute Distance for Medium-income Workers (kms)	se02	N/A	N/A	54.79858	50.9189	34.47894
Commute Distance for High-income Workers (kms)	se03	N/A	N/A	62.59104	54.12068	43.78447
Square Root of Commute Distance for Low-income Workers	srt_se01	Dependent	N/A	7.253526	7.408694	2.346087
Square Root of Commute Distance for Medium-income Workers	srt_se02	Dependent	N/A	7.046292	7.135745	2.269018
Square Root of Commute Distance for High-income Workers	srt_se03	Dependent	N/A	7.460047	7.356676	2.634173
Subregional Classification	new_class	Independent	Neutral	1.810902	2	0.914243
Area (sq. miles) if Block Group is in CBSA	cbsa_area	Independent	(-)	6.388439	0.395041	46.44377
Job Density (Jobs/Acre)	job_den	Independent	(-)	2.632375	1.87549	3.263775
Intersection Density (Intersections/Acre)	int_den	Independent	(-)	0.052203	0.030134	0.05668
Housing Density (Housing Units/Acre)	hous_den	Independent	Neutral	2.550495	1.61395	3.858934
Share of Pre-1970 Housing Units	oldhousing	Independent	(+)	0.276963	0.163121	0.284848
Share of Renter Occupied Housing Units	renterocc	Independent	(-)	0.364569	0.298587	0.266097
Ratio of Transit Users to Auto Users	tr_share	Independent	(-)	0.009938	0	0.032793
Share of Females	female	Independent	(+)	0.503999	0.506295	0.066791
Share of Population 65+ or more	seniors	Independent	(+)	0.128131	0.110057	0.089678
Share of African American Population	black	Independent	(+)	0.113836	0.044211	0.168523
Share of Asian Population	asian	Independent	(-)	0.0398	0.005527	0.078517
Share of Hispanic Populations	hispanic	Independent	(+)	0.38138	0.286858	0.299284
Share of Populations with Educational Attainment Below High School	bhs	Independent	(+)	0.182077	0.135466	0.15967
Share of Unemployed Populations	unemp	Independent	(+)	0.225057	0.227139	0.088067
Median Family Household Income ($)	mfhi	Independent	(-)	62102.32	55033	33011.73
Housing Cost ($)	housing_co	Independent	(-)	1158.815	1076	570.9819
Share of Low-income Jobs	lijobs	Independent	(+)	0.561041	0.566102	0.147458
Share of Manufacturing Jobs	manu	Independent	(+)	0.171208	0.160057	0.06977
Share of Retail Jobs	retail	Independent	(+)	0.15711	0.157937	0.031648
Share of Health Care Jobs	health	Independent	(-)	0.246816	0.235562	0.070428
Share of Arts and Entertainment Jobs	arts	Independent	(-)	0.012346	0.011594	0.007483
Share of Accomadation Jobs	accom	Independent	(-)	0.093645	0.090452	0.027961
Share of Finanancial and Information Jobs	finan_inf	Independent	(-)	0.107303	0.086116	0.070723
Square Root of Housing Cost	sqrt_hous_cost	Independent	(-)	25.75692	30.09983	14.03508
Time in Years	year	Independent	(-)	2	2	0.816504

Model Fit Statistics

LOW INCOME ACD MODEL

✓ **Random Effects Model**

- R^2 Within : 0.3966
- R^2 Between : 0.2670
- R^2 Overall : 0.3329

✓ **Random Forest Model**

- Validation Set - R^2: 0.691, RMSE: 1.300
- Test Set - R^2: 0.676, RMSE: 1.322

MEDIUM INCOME ACD MODEL

✓ **Random Effects Model**

- R^2 Within : 0.3966
- R^2 Between : 0.2670
- R^2 Overall : 0.3329

✓ **Random Forest Model**

- Validation Set - R^2: 0.650, RMSE: 1.341
- Test Set - R^2: 0.639, RMSE: 1.350

HIGH INCOME ACD MODEL

✓ **Random Effects Model**

- R^2 Within : 0.4252
- R^2 Between : 0.3663
- R^2 Overall : 0.3950

✓ **Random Forest Model**

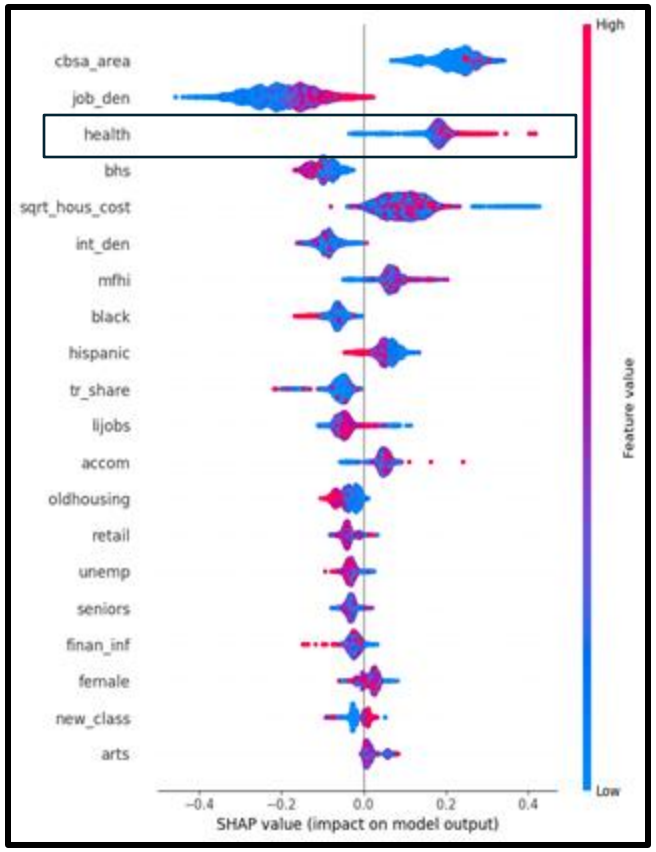
- Validation Set - R^2: 0.703, RMSE: 1.416
- Test Set - R^2: 0.706, RMSE: 1.423

❖ Clearly, the machine learning models used in the study provide a better fit to the data compared to random effects models, which are often used for handling variability across groups or clusters in the data.

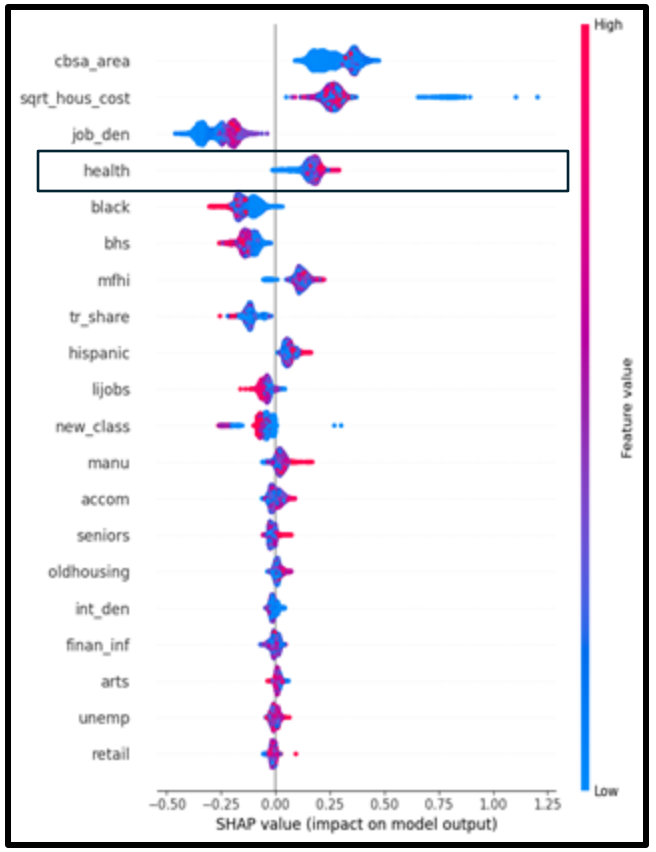
❖ A small difference between the model's performance on the validation and test datasets indicates that the model generalizes well to new, unseen data. This is a positive sign, suggesting that the model is not overfitting and can accurately make predictions beyond the data it was trained on.

❖ Although not displayed here, the year variable had the highest feature importance in the random forest model overall. However, when the model was broken down by year using LIME explanations, the impacts were minimal. This suggests that the model is effectively capturing temporal variation in a way similar to a random effects model, without relying too heavily on year as a predictor.

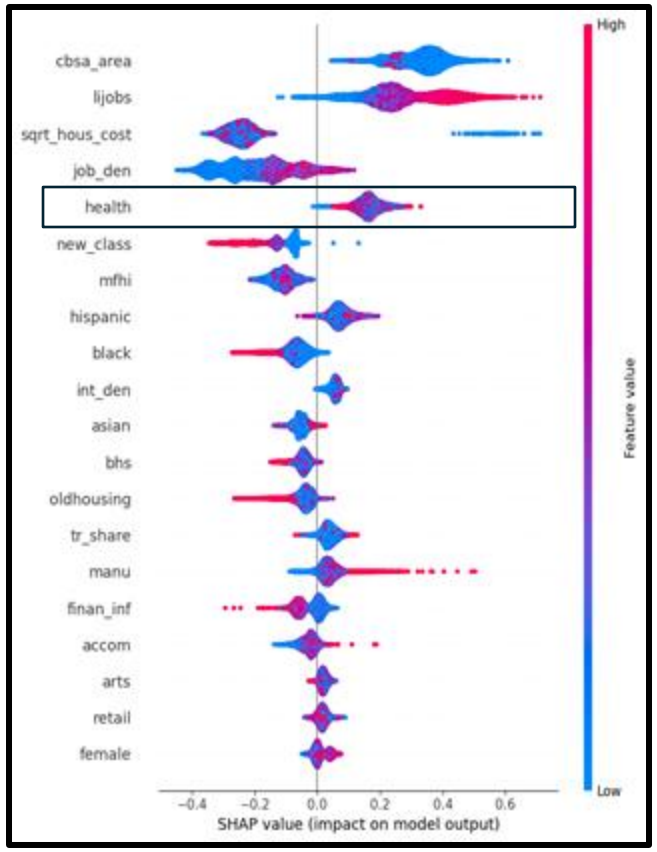
SHAP Explanations



LOW INCOME ACD MODEL



MEDIUM INCOME ACD MODEL



HIGH INCOME ACD MODEL

LIME Explanations

Variable	IMPACTS ON ACD OF EACH GROUP		
	Low Income	Medium Income	ACD (High Income)
cbsa_area	0.00586	0.02021	0.01308
job_den	0.17917	-0.01242	0.02614
int_den	0.00345	0.02498	0.01462
oldhousing	-0.00346	0.00033	-0.00008
renterocc	-0.00301	-0.00306	-0.01011
tr_share	-0.00019	0.00582	0.00364
female	0.00162	0.00322	0.00482
seniors	0.00511	0.01322	0.01626
black	-0.01306	-0.01409	-0.00248
asian	-0.00411	-0.00562	-0.01667
hispanic	-0.00504	-0.00406	-0.00043
bhs	-0.00363	0.00298	0.00325
unemp	0.00082	0.00558	0.00510
mfhi	-0.00120	0.01549	0.00235
lijobs	-0.01186	0.01876	0.10049
manu	0.01206	0.03540	0.04970
retail	-0.00031	-0.00163	0.00246
health	-0.00157	0.00142	0.01310
arts	0.00141	0.00343	0.00643
accom	0.00230	0.01201	0.02818
finan_inf	-0.26097	-0.04434	-0.01009
sqrt_hous_cost	-0.63286	-0.14689	-0.14567
2017	-0.00003	0.00000	0.00000
2021	-0.00003	0.00000	0.00000
Inner Ring Suburbs	-0.00001	-0.00002	-0.00002
Outer Ring Suburbs	0.00026	0.00027	0.00024
Rural	0.00032	0.00054	0.00044

- As CBSA area and intersection density increase, commuting distance rises for all groups, with the strongest impact on middle- and higher-income groups due to their greater reliance on driving. This confirms that low-income jobs are more dispersed in the suburbs while other types are close to the center
- Higher job density reduces commute distance for medium-income workers but increases it for other groups, with the strongest negative effect on low-income groups.
- As transit share increases, commute distance decreases for low-income groups but increases for other groups.
- In all three cases, a higher share of renter-occupied housing units reduces commute distance, with the strongest effects on high-income groups, indicating gentrification.
- Among Asians, who are more likely to work in high-income job sectors, the commute distance to high-income jobs is significantly lower than to low-income jobs. In contrast, for African Americans and Hispanics, especially African Americans, the trend reverses, suggesting no evidence of racial spatial mismatch.
- Unemployed individuals, females, those with less education, and seniors have the shortest commute distances to low-income jobs compared to other job types, with the strongest effects observed for seniors.
- As expected, low-income workers have shorter commutes to low-income jobs compared to other job sectors. However, as household income increases, commuting distance to low-income jobs decreases, while this trend reverses and becomes more pronounced for medium- and high-income groups.

LIME Explanations (Contd.)

Variable	IMPACTS ON ACD OF EACH GROUP		
	Low Income	Medium Income	ACD (High Income)
cbsa_area	0.00586	0.02021	0.01308
job_den	0.17917	-0.01242	0.02614
int_den	0.00345	0.02498	0.01462
oldhousing	-0.00346	0.00033	-0.00008
renterocc	-0.00301	-0.00306	-0.01011
tr_share	-0.00019	0.00582	0.00364
female	0.00162	0.00322	0.00482
seniors	0.00511	0.01322	0.01626
black	-0.01306	-0.01409	-0.00248
asian	-0.00411	-0.00562	-0.01667
hispanic	-0.00504	-0.00406	-0.00043
bhs	-0.00363	0.00298	0.00325
unemp	0.00082	0.00558	0.00510
mfhi	-0.00120	0.01549	0.00235
lijobs	-0.01186	0.01876	0.10049
manu	0.01206	0.03540	0.04970
retail	-0.00031	-0.00163	0.00246
health	-0.00157	0.00142	0.01310
arts	0.00141	0.00343	0.00643
accom	0.00230	0.01201	0.02818
finan_inf	-0.26097	-0.04434	-0.01009
sqrt_hous_cost	-0.63286	-0.14689	-0.14567
2017	-0.00003	0.00000	0.00000
2021	-0.00003	0.00000	0.00000
Inner Ring Suburbs	-0.00001	-0.00002	-0.00002
Outer Ring Suburbs	0.00026	0.00027	0.00024
Rural	0.00032	0.00054	0.00044

- Commute distance to manufacturing, arts, and accommodation jobs is higher across all income groups. While the effects are more pronounced for high-income groups, their reliance on cars offsets the cost. Spatial mismatch in this case is likely to affect low- and medium-income groups the most.
- As wholesale and retail jobs increase, commute distance decreases for medium- and low-income groups, with the strongest effects observed for medium-income groups.
- Surprisingly, as health & education and financial & information jobs increase, commute distance decreases for low- and medium-income groups, with the strongest effect on low-income workers, indicating a clear skill mismatch.
- Similarly, as housing costs rise, commute distance to all job types decreases, with the strongest effects on low-income groups—suggesting a mismatch driven by unaffordable housing
- While feature importance analysis identified time as the most influential factor, LIME results showed little effect of time, except for the low-income group. This suggests the model effectively captured random effects associated with the time variable.
- Trends in the subregional variable align with the actual distribution of jobs in U.S. metro areas, further validating the model's predictions.

Conclusion and Discussion

- Our study shows that combining machine learning models like Random Forest provides valuable insights into the complex factors driving spatial mismatch, accounting for non-linear relationships and data uncertainty.

- The main contribution of this paper is demonstrating how aggregating or disaggregating outputs from model-agnostic explanation techniques like LIME/SHAP can provide urban planners with deeper insights into the spatial and temporal dynamics of urban systems, enabling more targeted and context-sensitive decision-making.

- This study showed unique kinds of spatial mismatch patterns impacting different groups differently. While the findings related to Core Based Statistics Area (CBSA) and intersection density show signs of a sprawled urban pattern where higher- and middle-income groups generally try to live in suburban neighborhoods, due to their flexibility of driving and affordability of vehicles, the job density variable shows that it is the low-income groups being highly affected by living away from jobs.

- Despite higher transit usage among low-income groups, housing costs indicate that they are highly cost-burdened, often compelled to live near transit due to limited transportation options.

- A clear mismatch exists between job types and commute distances. White-collar jobs (e.g., finance, information, health, education) are negatively associated with low-income commute distance, suggesting they are less accessible to low-income workers. In contrast, blue-collar jobs (e.g., manufacturing, arts, accommodation) show a positive association, indicating that low-income workers must travel farther for these opportunities.

- These findings indicate a clear misalignment between affordable housing, job locations, and transportation options, especially for low- and medium-income groups.

- Future research should examine the applicability of geographically weighted machine learning models, like XGBoost, CatBoost, or neural networks, to capture the complex relationships underlying spatial mismatch. Similarly, additional variables like land use and commuting time can be analyzed, while decomposing SHAP outputs can provide deeper insights into the spatial and temporal dynamics of spatial mismatch.