The Journal of Transport and Land Use jtlu.org
Vol. 18 No. 1 [2025] pp. 397–424

JTLU

Data aggregation impacts on built environment-mode share models around public transit stations

Seyed Sajjad Abdollahpour (corresponding author)
Virginia Polytechnic Institute and State University
sabdullahpur@vt.edu

Huyen T. K. Le
The Ohio State University
le.253@osu.edu

Ralph Buehler
Virginia Polytechnic Institute and State University
ralphbu@vt.edu

Steve Hankey
Virginia Polytechnic Institute and State University
hankey@vt.edu

Abstract: This study examines how data aggregation influences the relationship between the built environment (BE) and mode share around 2,794 rail and BRT stations in the United States, using both inferential and machine learning methods. The results indicate that data aggregation impacts the outcomes of BE-mode share models, regardless of the data analysis approach. Models using network buffers are less affected by data aggregation compared to those using circular buffers, Thiessen polygons, or administrative boundaries (block groups). In addition, the optimal buffer sizes for capturing BE effects and minimizing sensitivity to data aggregation for active and public transit modes are 800 meters for BRT stations and 1000 meters for rail stations, while 1200 meters is effective for private vehicle mode share at both rail and BRT stations. Furthermore, key BE features in commuting mode share models—such as employment density, jobs per household, intersection density, residential density, distance from the central business district, job accessibility (active), and regional population density—remain robust against data aggregation. We recommend that urban and transportation planners account for aggregation biases and apply multiple methods when evaluating BE's impact on mode share around public transit stations to inform more effective policy recommendations.

Keywords: travel behavior, land use, urban form, zoning and scale effects, modifiable areal unit of problem

Article history:
Received: February 6, 2025
Received in revised form: April 3, 2025
Accepted: April 9, 2025
Available online: May 19, 2025

1 Introduction

Land-use planning around public transit stations is essential for encouraging active modes of travel and promoting transit-oriented development (TOD) (Li et al., 2025; Yang

https://doi.org/10.5198/jtlu.2025.2676
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

& Chang, 2025; Yin et al., 2025). To develop effective policies that support sustainable travel modes, it is crucial to accurately define public transit catchment areas and apply appropriate data analysis methods (Yang et al., 2021). The built environment (BE) surrounding transit stations influences individuals' travel behavior, specifically commuting mode share (Park et al., 2018), which provides insights into individuals' travel preferences and a deeper understanding of how effectively transportation infrastructure is being utilized (Babapourdijojin et al., 2024; Cui et al., 2020; Henao et al., 2015). Researchers commonly analyze BE-mode share relationships to identify planning strategies that reduce car dependency (De Vos et al., 2021; Jian et al., 2023; Khalil & Fatmi, 2025; Nakshi & Debnath, 2021). However, the way BE data is aggregated can influence study outcomes, making it essential to evaluate how different aggregation approaches affect model results.

While previous studies have focused on transit ridership to assess the impact of the built environment (BE) around public transit stations (Liu et al., 2023; Zhang et al., 2023), our study emphasizes mode share modeling to capture multi-modal demand. Unlike direct ridership models, mode share analysis considers all travel modes, providing a more comprehensive view of transportation behavior (Pan & Sharifi, 2024). This is particularly important in transit-oriented developments (TODs), where reducing car dependency and promoting active transportation are key goals (Jamme et al., 2019). Understanding mode share helps urban and transportation planners evaluate policies aimed at shifting travelers from private vehicles to sustainable options like public transit, biking, and walking (Papadakis et al., 2024; Wey & Huang, 2018).

To identify key predictors of commuting mode share around public transit stations, previous studies commonly use fixed-size circular buffers or administrative boundaries to define catchment areas and explore the relationship between BE features and mode share (Nasri & Zhang, 2019; Park et al., 2018; Wu et al., 2021). This approach offers an easier data extraction process compared to other types of data aggregation methods (Li et al., 2020; Liu et al., 2023). However, circular buffers may not accurately represent public transit catchment areas as people's movements and activities may not adhere to circular patterns or administrative boundaries (Liu et al., 2023). Additionally, selecting the appropriate size of public transit catchment area is crucial, as it can impact the findings regarding the association of BEs and mode share (Clark & Scott, 2014). Consequently, the findings regarding BE and mode share around public transit stations according to fixed-size buffers or administrative boundaries may be biased.

Some existing literature suggests that different buffer sizes (or public transit catchment areas) have a limited impact on model results (Guerra et al., 2012). However, another set of studies indicates that various sizes and shapes of catchment areas can notably affect model outputs (Akbari & Bafarasat, 2024; Hong et al., 2014; Mitra & Buliung, 2012; Yang et al., 2019). While previous literature has explored the influence of BE on travel behavior based on different buffer sizes (Boarnet & Crane, 2001; Gehrke & Clifton, 2014; Laviolette et al., 2022; Zhang & Kukadia, 2005), gaps still remain.

First, many existing studies primarily focus on rail catchment areas to examine the effects of data aggregation on model outcomes (Guerra et al., 2012; Liu et al., 2023; Yang et al., 2021). It is still uncertain whether findings from rail stations can be applied to other public transit catchment areas, such as Bus Rapid Transit (BRT) systems. Second, previous research regarding public transit catchment areas, with an emphasis on data aggregation methods, often assumes a linear link between land-use factors and mode share and uses inferential approaches (Guerra et al., 2012; Yang et al., 2021). However, the association between BE and commuting mode share may exhibit complexity, and the assumption of linearity could introduce bias into the results (Abdollahpour et al., 2024; Xiao & Wei, 2023). Additionally, few studies, if any, take a comprehensive approach to

investigating the impact of buffer sizes and shapes on the three primary commuting modes (active, public transit, and private vehicles), comparing the results across different sizes and shapes at a national level. The majority of existing studies focus on direct transit ridership models and overlook multi-modal investigations. Such an approach would enhance the generalizability of the findings.

To fill these research gaps, the current study employs inferential and machine-learning approaches to comprehensively examine the presence and seriousness of data aggregation bias in BE-mode-share models, and how data aggregation bias is sensitive to the type of (rail vs. Bus Rapid Transit (BRT)) as well as data analysis approaches. The analysis is based on approximately 2,800 fixed-guideway transit stations, comprising 2,180 rail and 620 BRT stations, distributed across 34 metropolitan areas in the United States. This study utilizes a variety of datasets, including the American Community Survey, Smart Location Database, Longitudinal Employer-Household Dynamics (LEHD), Census/TIGER 2019, and the Bureau of Economic Analysis, among others. This research aims to address these research questions: (i) Does the data aggregation process impact the output of BE-mode share models around public transit stations? (ii) Which buffer sizes and shapes are more robust to data aggregation around rail and BRT stations? (iii) Which BE features are (not) sensitive to data aggregation around public transit stations? The current study offers several key contributions. First, it expands our understanding of how data aggregation biases affect BE and mode share models for three main commuting modes, including active, public transit, and private vehicles across rail and BRT station areas at the national level. Second, it identifies the optimal sizes and shapes of transit catchment areas for rail and BRT stations to address data aggregation bias and guide mode share planning. Third, the study highlights key predictors of commuting mode share around transit stations, regardless of data aggregation effects. Lastly, it provides empirical insights into the sensitivity of different data analysis approaches (inferential vs. machine learning) in revealing data aggregation bias in mode share analysis.

The structure of this article is as follows: Section two presents a comprehensive review of relevant literature pertaining to the impact of data aggregation on results of travel behavior studies, while section three outlines our data extraction process and modeling methodologies employed. Section four presents modeling results and offers insights into the interpretation of scale and size effects. Finally, the concluding section discusses the findings, offers policy recommendations, identifies limitations, proposes directions for future studies.

2 Literature review

2.1 Data aggregation and mode share

Focusing on TODs as a strategy to address sprawl development has been emphasized over the past three decades (Feudo, 2014; Liu et al., 2022). Examining the impact of the BE on travel behavior in the concept of TOD has gained significant attention in both academic research and practical applications (Jamme et al., 2019). Most studies explore the relationship between land-use characteristics at the station level and transit ridership, often relying on direct ridership models (Loo et al., 2010; Zhang et al., 2023). Compared to traditional four-step models, direct ridership models offer advantages such as lower cost, ease of interpretation, and faster application (Cardozo et al., 2012; Liu et al., 2023).

However, focusing on mode share models to capture multi-modal demand rather than solely emphasizing transit ridership can provides a comprehensive view of transportation demand by considering the proportion of travelers using different modes (Pan & Sharifi,

2024; Yang et al., 2019) around public transit stations. In contrast, ridership-based studies reflect only public transit demand without accounting for other modes like driving, walking, or cycling. Understanding travel distribution across various modes is crucial for effective transportation planning and policymaking in the concept of TODs, as different land-use characteristics influence travel behavior differently (e.g., residential or employment density effects on public transit use vs. private vehicle use) (Tao et al., 2023).

It should be noted that the mode share-BE model, like any other modeling approach around public transit stations, can be affected by data aggregation bias. This issue aligns with the concept of the modifiable areal unit problem (MAUP), which refers to the bias resulting from different aggregation approaches with different shapes and scales of the unit of analysis (Openshaw, 1984). Past studies have limitedly tested on mode share models and found evidence for MAUP (Gao et al., 2021; Yang et al., 2021).

Variations in results often arise from spatial data aggregation methods and the level of spatial resolution (Clark & Scott, 2014; Oliver et al., 2007). The scale of aggregation plays a crucial role in shaping the outcomes of mode share-BE relationships, as larger units tend to smooth out variations, while smaller units capture more localized patterns (Gao et al., 2022; Li et al., 2022; Luo et al., 2025). Additionally, the configuration of spatial units significantly influences results, as different zoning methods—such as block groups, circular buffers, network-based buffers, or Thiessen polygons—can divide the same area differently, leading to variations in variable relationships (Mitra & Buliung, 2012; Yang et al., 2021). Moreover, artificial boundaries in spatial units may exclude relevant data or group dissimilar areas together, further affecting model outputs.

Existing studies on the impact of data aggregation on travel behavior highlight the effects of scale, size, and method across various travel behavior contexts (Duan et al., 2023; Pani et al., 2019; Zhou et al., 2022). For example, (Zhang & Kukadia, 2005) examined the presence of scale and zone effects on mode choice in the U.S. using five circular buffers around household locations and three different administrative boundaries. Similarly, Yang et al. (2019) confirmed the influence of scale and shape effects on different travel purposes based on seven circular buffer sizes around home and work locations. Lastly, Gao et al. (2021) investigated the effects of data aggregation in terms of method, size, and shape on dockless bike-sharing usage in Shenzhen, China, finding that most BE variables were sensitive to data aggregation.

2.2 Review of the existing findings

The existing literature examining the impact of data aggregation approaches on the link between BE features and travel behavior has predominantly utilized the "5D" variables defined by Ewing and Cervero (2010). These variables include density, diversity, design, destination accessibility, and distance to transit, serving as input variables to quantify BE features around transit stations (Clark & Scott, 2014; Laviolette et al., 2022; Li et al., 2020).

Most existing studies investigating the influence of data aggregation on travel behavior around public transit stations have concentrated on rail stations (Guerra et al., 2012; Liu et al., 2023; Yang et al., 2021). Other transit station systems, such as BRT, have been relatively overlooked. It is important to acknowledge that the findings on the impact of data aggregation on BE-mode share models around rail stations may not be applicable to other kinds of public transit system, such as BRT.

Furthermore, a considerable portion of existing research assumes a linear relationship for BE-travel behavior studies (Hong et al., 2014; Kuby et al., 2004; Wu et al., 2021).

These studies typically employ regression techniques such as ordinary least squares, negative binomial, mixed effects, and logistic regression. However, some studies on travel behavior have indicated that the relationship between BE and travel behavior may exhibit greater complexity, suggesting the need to examine non-linear assumptions (Ding et al., 2018). Empirical findings have further revealed the nonlinear link between BE and mode share around transit stations (Abdollahpour et al., 2024).

In recent years, studies using machine learning methods have also found strong evidence for the impact of data aggregation on nonentity associations and the contribution of predictors (Zhang et al., 2023). For example, Laviolette et al. (2022) demonstrated that BE features such as population density, entropy, and design elements are related to car ownership (a proxy for travel behavior) across different buffer sizes.

Moreover, the existing studies have concentrated on exploring the influence of size in the context of data aggregation between 200m and 1200m (Yang et al., 2019). The influence of various shapes of the unit of analysis has been overlooked. Studies commonly employ a circular buffer method when delineating buffer zones, often neglecting the real activity space and overlapping areas of station coverage (Gutiérrez et al., 2011; Wu et al., 2021). However, some researchers have demonstrated that accounting for overlapping areas in defining public transit catchment areas can alter both the shape of catchment areas and the output of the models (Sun et al., 2016).

2.3 Research contribution

In summary, several gaps in the current literature warrant further exploration. First, the majority of existing studies focus on direct transit ridership models and overlook multi-modal investigations in the impact of data aggregation around public transit stations. Second, most studies predominantly focus on rail stations, neglecting other public transit types, such as Bus Rapid Transit (BRT), when investigating the effect of data aggregation on BE and travel behavior models. Third, existing research often operates under the assumption of linearity, with limited investigations into nonlinearity and the extent of bias introduced by different data analysis approaches. Finally, inconsistencies in previous studies exist regarding the optimal size and shape of catchment areas and their key contributors to travel behavior. This study addresses these gaps by adopting a comprehensive approach that considers various data aggregation scales and defines catchment areas using methods such as circular buffers, network buffers, Thiessen polygons, and administrative boundaries. We explore the impacts of scale (e.g., different buffer sizes) and zone (e.g., administrative boundaries) while comparing different data analysis approaches (e.g., machine learning vs. inferential methods) in relation to data aggregation bias.

3 Method

This section provides a detailed account of the steps taken to collect and prepare the data and describes the analytical techniques used.

3.1 Study area and data extraction methods

We gathered data from 2,794 fixed guideway transit stations, which included 2,174 light and heavy rail stations as well as 620 BRT stations, distributed across 34 metropolitan statistical areas in the United States (Fig. A1-1 in the Appendix A1). We

considered several steps to confirm the station types. Initially, we accessed the National Transit-Oriented Development (TOD) database to gather longitude and latitude coordinates of rail and BRT stations. Following this, we compared the station types with the most recent information from the National Transit Database (NTD) and used Google Maps to validate the station locations, ensuring the reliability of the location data.

In this study, variables were measured at nineteen different levels of analysis to reflect various data aggregation processes in terms of size, shape, and methods. We selected these methods and sizes based primarily on existing literature (Eom et al., 2019; Farber & Marino, 2017; Nasri & Zhang, 2014). The nineteen units of analysis include six circular buffers, six Thiessen polygons, and six network buffers (each at 200, 400, 600, 800, 1000, and 1200 meters), as well as one administrative unit (US Census block group). The six selected buffer sizes are based on common buffer sizes in the existing literature on BE–travel behavior around public transit stations (Abdollahpour et al., 2024; Guerra et al., 2012; Liu et al., 2023; Yang et al., 2021). Once the catchment areas were defined, we computed an area-weighted value for variables, factoring in the proportional representation of each block group within the corresponding buffer zone. More details regarding the data extraction are provided in Appendix A2.

3.2 Variables and measurement

We consider the commuting mode share as the outcome variable for this study because it reflects the proportion of travelers using different modes of transportation, offering insights into travel behavior and preferences (Cui et al., 2020). Mode share also can provide a deeper understanding of how effectively transport infrastructures are being utilized (Henao et al., 2015). Moreover, understanding commuting mode share at an aggregate level can help highlight the broader influence of urban planning decisions on population-level patterns, supporting more equitable and efficient infrastructure investments (Braun et al., 2016). Accordingly, to establish the primary mode share of commuting around public transit stations, we utilized the 5-year estimates from the American Community Survey (2015-2019). Specifically, we examined three outcome variables: (i) the proportion of individuals commuting via active transportation modes (cycling and walking), (ii) the proportion of commuters using public transit, and (3) the proportion of individuals commuting by private vehicles.

The explanatory variables were computed in the same way as the outcome variables. Around stations, we included the D variables from the Smart Location Database along with socio-economic factors, such as income, race, gender, age, education, and car ownership. In addition, we define polycentricity and population density as regional level urban form in our analysis. Table 1 provides an explanation and descriptive statistics of variables used at 800-meter circular buffer as an example. A full list of descriptive statistics for all levels of aggregation can be found in Appendix A3.

Table 1. Variables, explanations, data sources and descriptive statistics

Variables	Description	800-meter circular buffer	
		BRT	**Rail**
		Mean (SD)	Mean (SD)
Outcome variables			
Private vehicles[1]	Percent private vehicles commuting mode share within catchment areas	66.07 (23.1)	48.93 (27.07)
Public transit[1]	Percent transit commuting mode share within catchment areas	12.85 (13)	26.25 (19.71)
Active modes[1]	Percent active (walking and cycling) commuting mode share within catchment areas	12.57 (13)	13.47 (13.42)
Explanatory variables – station area level			
Residential density[1]	Number of households within catchment areas	9.23 (11.8)	17.30 (18.6)
Employment density[2]	Number of jobs within catchment areas	39.02 (88.5)	70.19 (200.9)
Job per household[3]	The number of jobs per each household	12.75 (32)	16.04 (62.72)
Destination accessibility[4]	Distance from CBD (km)	12.4 (17.1)	9.3 (8.6)
Land use diversity[3]	Job housing land use diversity	0.61 (0.13)	0.58 (0.13)
Block size[4]	Average block size (Acre)	376.5 (670)	418.9 (1154)
Intersection density[3]	Intersection density with respect to the multi-modal intersections	144.7 (79.5)	168.90 (94.7)
Road density[3]	The length of road (km)	25.6 (8.7)	29.30 (9.06)
Station density[5]	Number of public transit stations within the catchment areas (all types of stations)	18.63 (6.24)	20.05 (8.03)
Job accessibility by private vehicle[6]	Total jobs reachable within a 45-minute drive by private vehicles	304629 (615601)	803871 (107310)
Job accessibility by active transit[6]	Total jobs reachable within a 45-minute transit and walking commute	167064 (156206)	319295 (236107)
Control variables: socio-economic			
Transit service[6]	Aggregate transit service frequency, afternoon peak period (number of trips or departures per hour) within catchment areas	45.2 (61.69)	76.89 (95.87)
Income[1]	Percent low-income workers within the catchment areas	0.25 (0.09)	0.24 (0.08)
Car ownership[1]	Percent household without car within the catchment areas	0.20 (0.17)	0.29 (0.22)
Bachelor and higher degree[1]	Percent of the population having Bachelor and higher degree within the catchment areas	40.05 (20.9)	46.12 (22.41)
Population 65 and over[1]	Percent population over 65 within the catchment areas	13.07 (5.94)	12.91 (5.72)
Population under 20	Percent population under 20 within the catchment areas	19.14 (8.58)	18.15 (8.32)
Race[1]	White = 1; others = 0	1=50% 0=50%	1=55% 0=45%
Gender [1]	Male = 1; Female = 0	1=63% 0=37%	1=65% 0=35%
Urban form: regional level		BRT Mean (SD)	Rail Mean (SD)
Population density[7]	Number of people per square Kilometer	348 (342)	603 (376)
Polycentric development[4]	Population shares of a region's centers to its total population	0.5 (0.1)	0.48 (0.12)

Data source: American Community Survey 2019 (5-year estimates), [2] Data source: The Longitudinal Employer-Household Dynamics (LEHD) 2019, [3] Data source: Smart Location Database (SLD) 2019, [4] Data source: Calculated by using Census/TIGER 2019, [5] Data source: Points of Interests (POIs) 2019, [6] Data source: General Transit Feed Specification (GTFS) 2019, [7] Data source: The Bureau of Economic Analysis (BEA) 2019.

3.3 Analytical methods

We employ two widely used machine-learning algorithms in BE-travel behavior studies, namely Random Forest (RF) and Extreme Gradient Boosting Trees (XGBT) (Abdollahpour et al., 2025; Aghaabbasi & Chalermpong, 2023; Eldafrawi et al., 2023), alongside two commonly used inferential approaches, multiple linear regression (MLR) and Hierarchical Linear Modeling (HLM) (Park et al., 2018; Renne et al., 2016; Tian et al., 2023), to analyze data and demonstrate how aggregation approaches can affect the model results. By using these four methods, we can compare machine learning and inferential approaches, ensuring robust findings and exploring both predictive accuracy and explanatory insights across these widely used techniques.

To compare and evaluate model performance, we consider a few metrics to select the best model for each outcome according to rail vs BRT catchment areas. Models are evaluated based on R-Squared (R^2), Root Mean Squared Error (RMSE), and Median Absolute Error (MAE). Accordingly, R^2 indicates the proportion of variance explained by the model, RMSE measures the magnitude of prediction errors with emphasis on large deviations, and MAE provides a robust measure of typical prediction errors, less sensitive to outliers. These metrics together offer a comprehensive evaluation of model fit and accuracy, suitable for comparing linear regression and XGBoost models across rail and BRT catchment areas. All of these models and analyses are conducted in RStudio and Python.

3.3.1 Inferential approaches

Inferential statistical approaches include regression analysis (Makar & Rubin, 2018). Multiple linear regression (MLR) helps understand the association between variables. MLR is one of the straightforward classical regression models used to predict the outcome variable based on some explanatory variables

Another commonly used regression model in travel behavior and BE studies is hierarchical linear modeling (HLM). Our data exhibits a hierarchical structure, with station areas "nested" within metropolitan areas. Thus HLM is a well-suited statistical approach for analyzing nested data (Hox et al., 2017). HLM accommodates dependence among observations, addressing issues such as biased standard errors and inefficient regression coefficients that may arise in single-level statistical methods. The estimation method for the HLM model in our study is restricted maximum likelihood (REML) to provide unbiased estimates of variance parameters. For inferential models, we include all built environment variables at the station and regional levels, as well as socio-economic features as control variables. RStudio was used to run inferential models.

3.3.2 Machine learning approaches

For this study, we used two ensemble algorithms, such as the Random Forest (RF) and Extreme Gradient Boosting Trees (XGBT) to analyze data. The RF model is an ensemble learning technique that creates numerous decision trees throughout the training process. Each “tree in the forest” operates independently, and the final prediction is made by averaging or voting the predictions of all the individual trees (Alpaydin, 2020; Breiman, 2001). On the other hand, XGBT employs a machine learning algorithm that operates on a gradient boosting framework and incorporates regularization methods, as well as parallelization, to effectively analyze the data (Chen & Guestrin, 2016). Unlike RF, which builds independent trees, XGBT builds trees in a sequential manner, with each

new tree aimed at addressing the errors made by the preceding trees (Wade & Glynn, 2020).

XGBT and RF models have the capability to assess the relative importance of predictors, reflecting the contribution of predictors in reducing model prediction errors. Additionally, both the RF and XGBT models enable the creation of partial dependency plots, which provide important insights into the relationship between the BE and mode share. These plots demonstrate how variations in particular predictor variables influence the outcome, while holding other factors constant.

Additionally, it should be noted that before building the machine learning and inferential models, the variance inflation factor (VIF) was calculated for the predictors to identify potential multicollinearity. Recognizing highly correlated predictors can help address potential biases in the models' outcomes, such as when producing partial dependency plots in machine learning approaches (Molnar, 2020). Predictors with a high VIF value (greater than 10) were removed. Appendix A4 shows more details regarding the VIF analysis of each rail and BRT station model separately.

4 Results

This section presents our findings on the impact of data aggregation methods, as well as the size and shape of station areas, on the outcomes of BE-mode-share models. We focus on three primary aspects: (i) the sensitivity of model performance to different data aggregation methods; (ii) the sensitivity of the contribution of BE factors to mode share models, evaluated through relative importance and significance levels; and (iii) the sensitivity of the relationship between predictors and outcomes, considering the sign, direction, effective range and estimated elasticity of these predictors.

4.1 Optimal data aggregation method, size, shape

We compared the R^2, MAE, and RMSE across OLS, HLM, RF, and XGBT models within different units of analysis (one block group level, 6 circular buffers, 6 network buffers, and 6 Thiessen polygons) for active travel, public transit, and private vehicle mode shares separately (Figure 1). To save space, we stay with one inferential model (HLM due to its suitability based on the data structure) and one machine learning model (XGBT due to better performance than RF model). The complete model performance details can be found in Appendix 5.

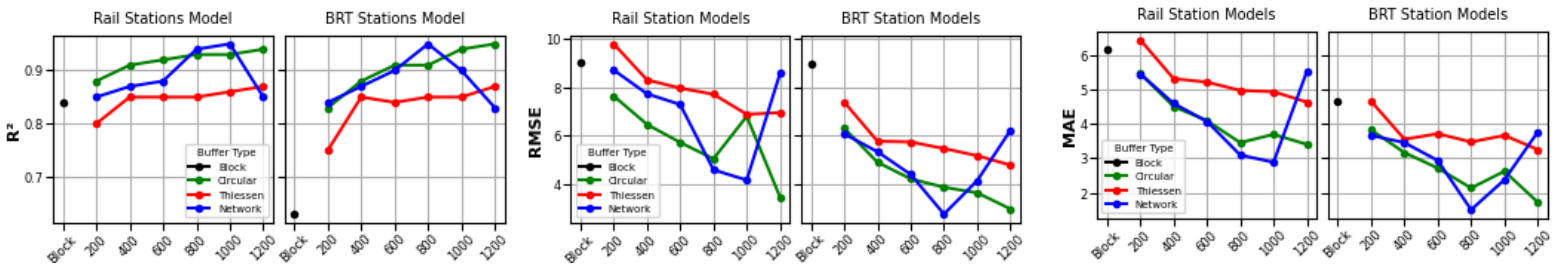
The results indicate that both machine learning and inferential models' performances are sensitive to data aggregation, with machine learning models showing greater sensitivity. On average, R^2 changes by approximately 0.1 more for machine learning models compared to inferential models, while RMSE and MAE changes by 3 and 2 units, respectively, more than in the inferential results (Figure 1-A to A-F).

Furthermore, the results illustrate that while block group-based models have the lowest performance, network buffer-based models outperform other data aggregation methods in predictive power for active, public transit, and private vehicle mode share models for both rail and BRT systems. Following network buffers, circular buffers show better performance, followed by Thiessen polygons, and administrative boundaries. For instance, Figures 1-C and 1- D show that the public transit mode share around BRT stations has the best model performance for 800-meter network buffers (R^2: 0.7 and 0.95 for HLM XGB respectively), followed by 1000-meter circular buffers (R^2: 0.68 and 0.92

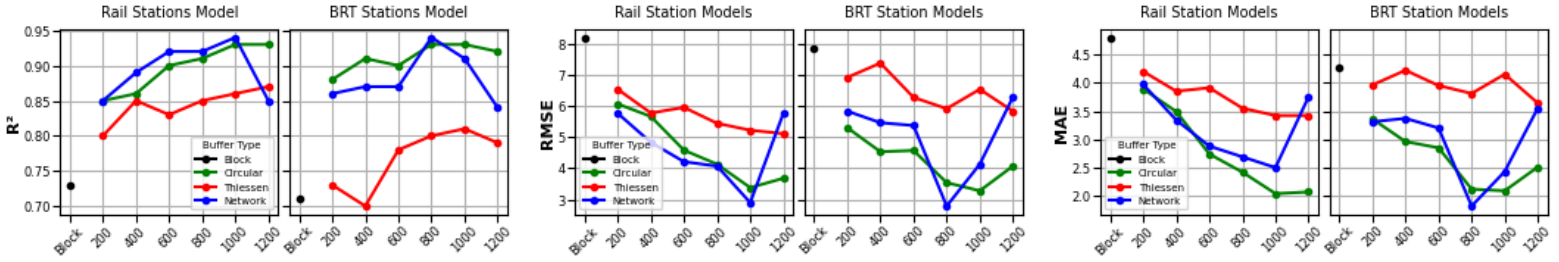
for HLM XGB respectively), 1200-meter Thiessen polygons (R^2: 0.82 and 0.55 for HLM XGB respectively), and block group level (R^2: 0.42 and 0.55 HLM XGB respectively).

Notably, the results highlight that certain network buffers' sizes, tailored to different commuting modes, demonstrate the best model performance and effectively capture the BE effects on mode share across all modeling approaches. Active mode and public transit mode share models around rail stations show the best performance with a 1000-meter network buffer, while for BRT stations, an 800-meter network buffer is optimal (Figures 1-A to 1-D). For private vehicle mode share models, the 1200-meter network buffer demonstrates greater predictive capability than those employing other aggregation methods (Figures 1-E and 1-F).

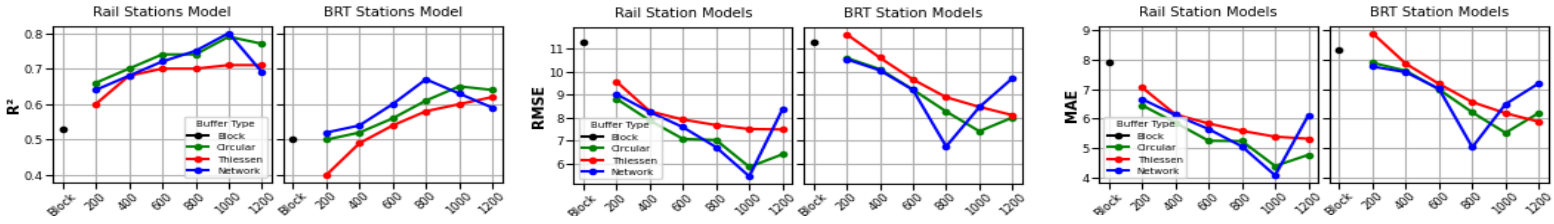
The findings also show that model performance variability, as calculated by the standard deviation of R^2, RMSE and MAE, is more pronounced with certain data aggregation methods and buffer sizes. Thiessen polygons (for BRT's active mode; Figures 1-A and 1-B) and circular buffers (for BRT's public transit and private vehicles; Figures 1-C to 1-F) exhibit the most variability.



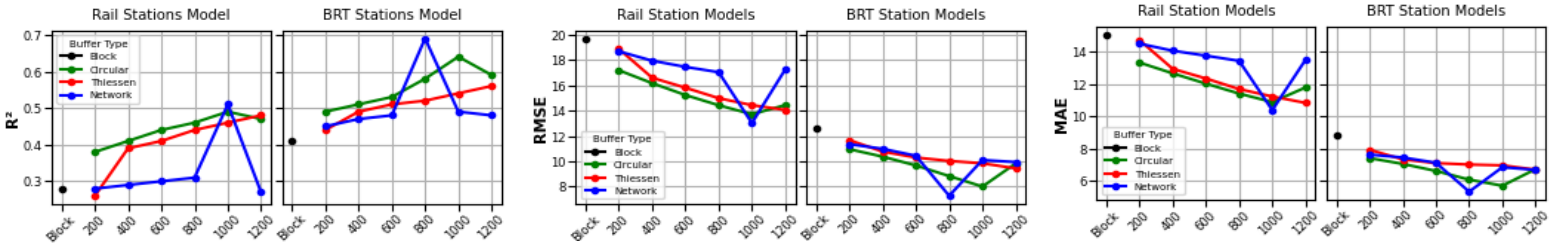
1-A. Active modes of commuting based on XGBT model outputs



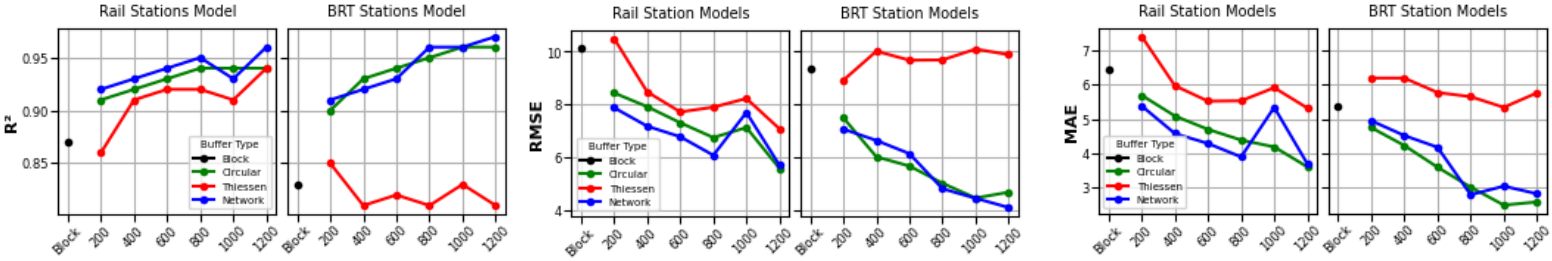
1-B. Active modes of commuting based on HLM model outputs



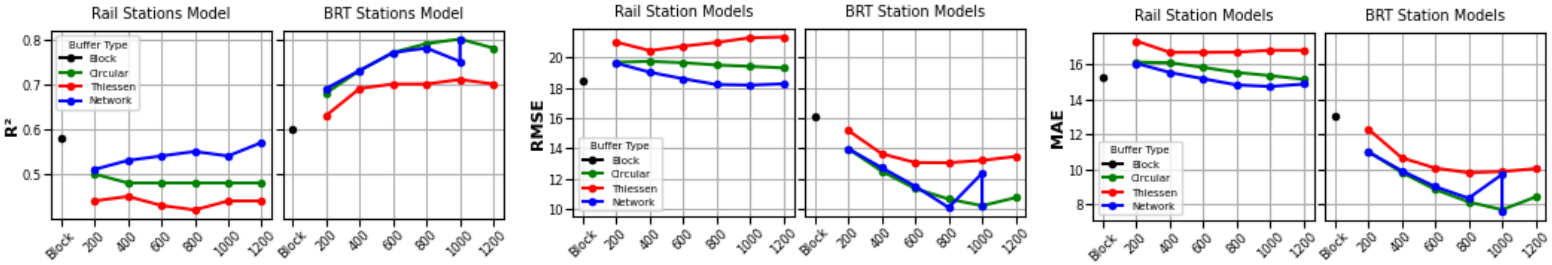
1-C. Public transit commuting based on XGBT model outputs



1-D. Public transit commuting based on the HLM model outputs



1-E. Private vehicle commuting based on XGBT model outputs



1-F. Private vehicle commuting based on HLM model outputs

Figure 1. Model performance of commuting mode shares based on inferential and machine learning approaches around rail and BRT stations

4.2 Sensitive BE features to data aggregation

To investigate the sensitivity of BE features to data aggregation in mode share models, we focus on the concepts of relative importance and nonlinearity effects within machine models. Also, we compare the significance levels, the extent of each factor's influence, termed as elasticity, and the signs and direction of BE features in inferential models.

4.2.1 ML models output

4.2.1.1 Relative importance

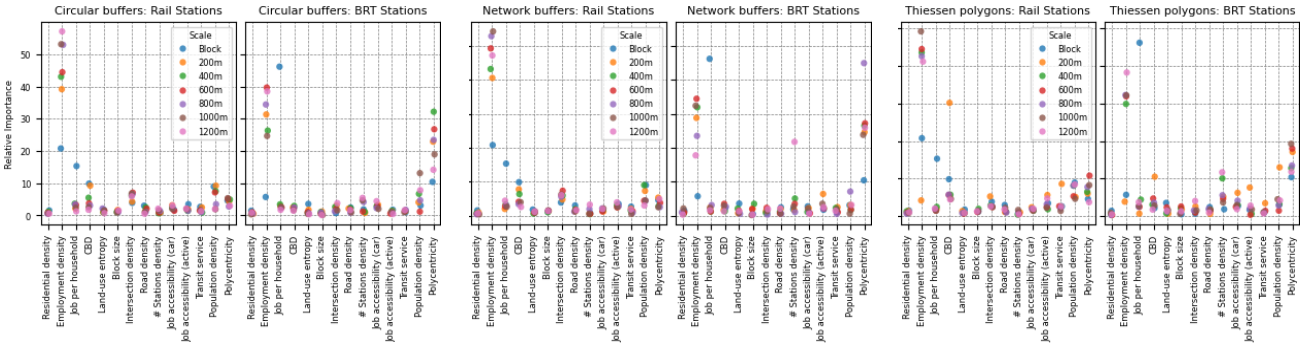
We compared the relative importance and ranking of predictors for rail and BRT stations areas. Figure 2 shows the variation in the relative importance of different BE features across multiple spatial scales. The vertical position of each dot represents the relative importance of that variable, with higher values indicating a stronger contribution in predicting the outcome. Variables with closely clustered dots across different scales

suggest more stability in their relative importance, meaning that the aggregation method or spatial scale does not affect their contribution. On the other hand, a more dispersed pattern of dots indicates greater sensitivity to changes in scale, meaning that the variable's importance fluctuates based on the buffer size. More details regarding relative importance can be found in Appendix 6.

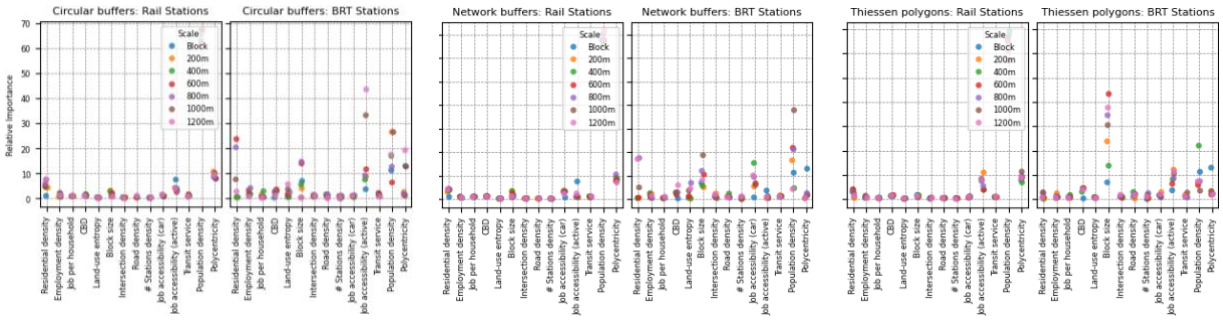
The results highlight that certain BE features consistently rank high in terms of relative importance, such as employment density, job per household, intersection density, residential density, distance from CBD, job accessibility (active), and regional polycentricity and population density (Figure 2 and Appendix A6). However, the impact of these features varies with the data aggregation method and scale. For example, employment density's importance for rail stations' active mode shifts from 20% to 55%, while for BRT stations, residential density changes from 5% to 25%, block size from 10% to 40%, and polycentricity from 15% to 35%.

Furthermore, we used the average variance of relative importance for BE features to compare different buffer sizes (Appendix A6, and Table A6-13). For rail stations, the 1000-meter buffer shows the lowest variance for active (5.41) and public transit mode (0.27), followed by the 800-meter buffer (active: 19.68; public transit: 0.30). For rail stations' private vehicle mode, the 1200-meter buffer has the lowest variance (1.37), followed by the 1000-meter buffer (1.87). For BRT stations, the 800-meter buffer has the lowest variance for active mode (1.36) and public transit (5.97), followed by the 600-meter buffer (active: 1.41; public transit: 9.30). For private vehicles, the 1200-meter buffer has the lowest variance (70.49), followed by the 400-meter buffer (71.65). The findings also show that relative importance sensitivity is less pronounced with network buffer-based models, based on the average variance of relative importance within each method (Figures 2-A to 2-C and Tables A6-14). For instance, for rail and BRT station area's active modes, the average variance of network buffers is 11.15, followed by 13.5 in circular buffers and 27.9 in Thiessen polygons.

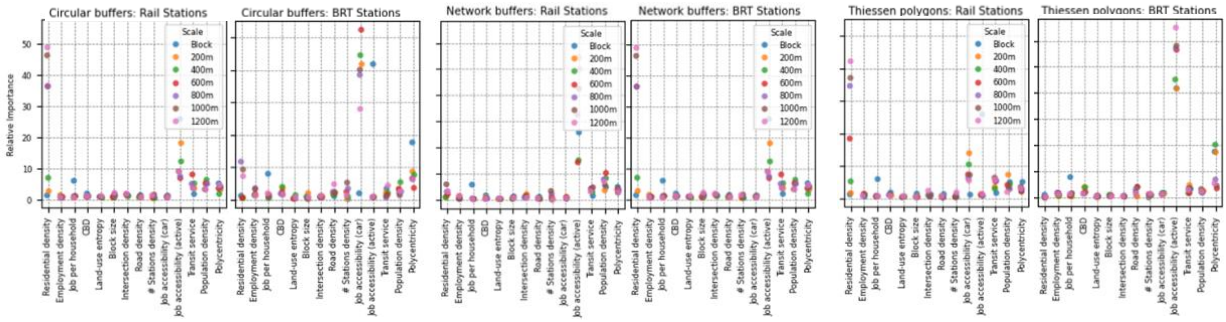
Figure 3 illustrates the collective contribution of predictors across three main categories: local (station) built environment features, socio-economic factors, and regional urban form. The findings suggest that these categories' overall contribution remains stable regardless of data aggregation. For instance, local-level BE features consistently play a dominant role in influencing active modes for both rail and BRT stations, contributing an average of 65% for rail and 55% for BRT across methods and scales (Figures 3-A). Similarly, the results indicate that regional-level urban form contributes notably more to rail stations' public transit mode share and remains stable regardless of the data aggregation method used (Figures 3-C).



2-A. Aggregation method impact on the results of relative importance of predictors in Active modes models

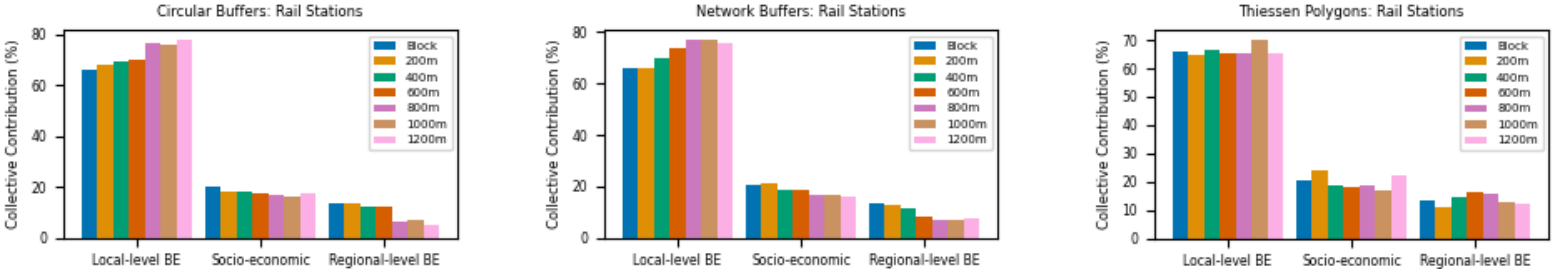


2-B. Aggregation method impact on the results of relative importance of predictors in public transit models

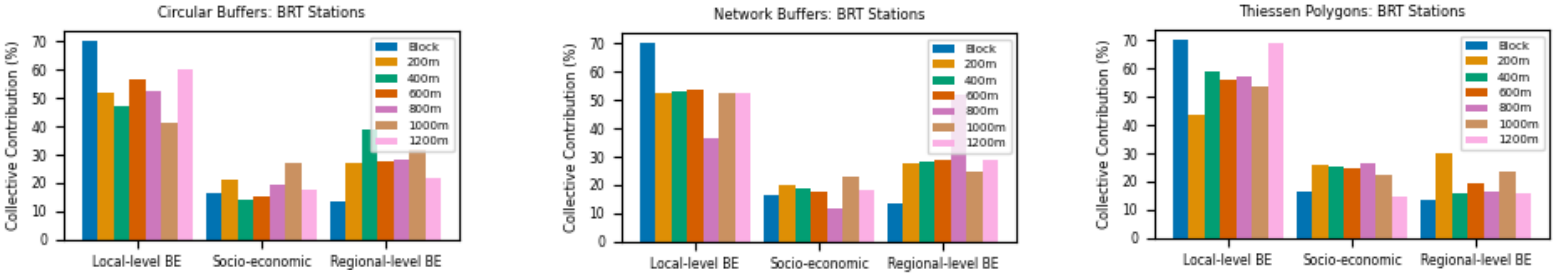


2-C. Aggregation method impact on the results of relative importance of predictors in private vehicles models

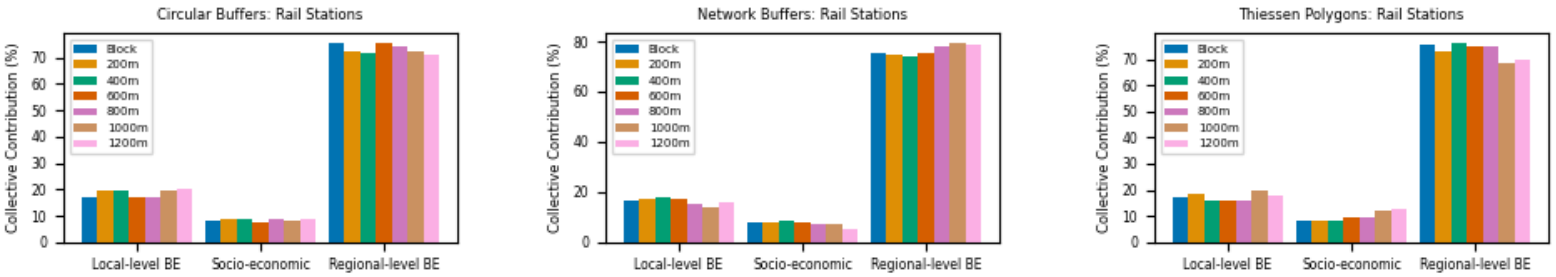
Figure 2. Aggregation method impact on the contribution of predictors in commuting models as relative importance (left side: Circular buffers; the middle: Network buffers; right side: Thiessen polygons)



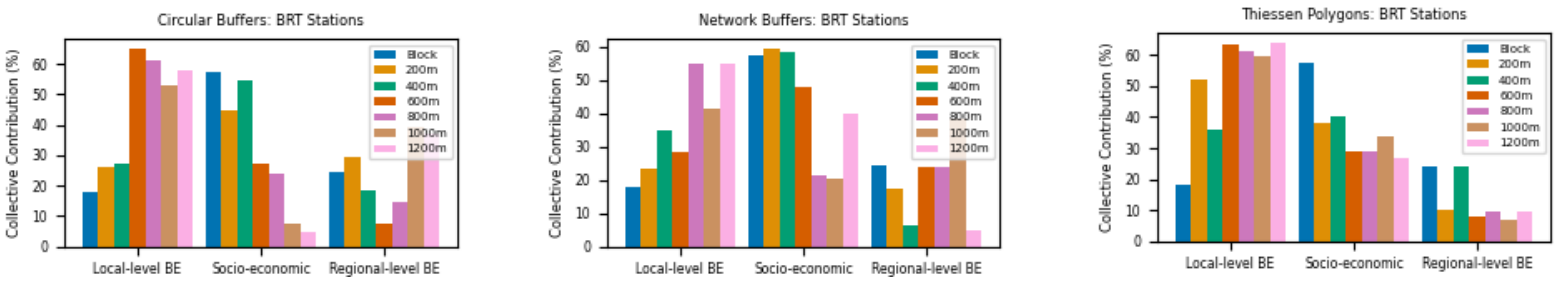
3-A. Collective contribution of three main categories of predictors for active modes of commuting at Rail stations



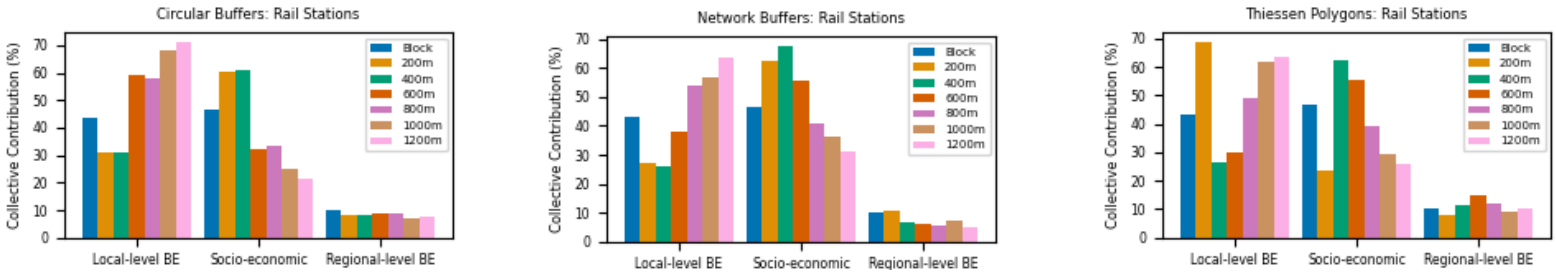
3-B. Collective contribution of three main categories of predictors for active modes of commuting at BRT stations



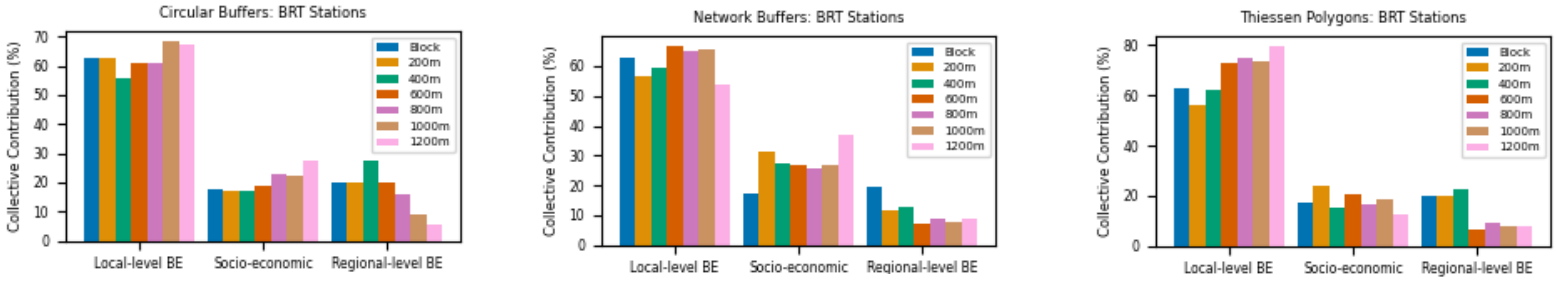
3-C. Collective contribution of three main categories of predictors for public transit mode at rail stations



3-D. Collective contribution of three main categories of predictors for public transit mode at BRT stations



3-E. Collective contribution of three main categories of predictors for private vehicle mode at rail stations



3-F. Collective contribution of three main categories of predictors for private vehicle mode at BRT stations

Figure 3. Collective contribution of predictors in commuting mode share models across different data aggregation methods

4.2.1.2. Nonlinearity effects

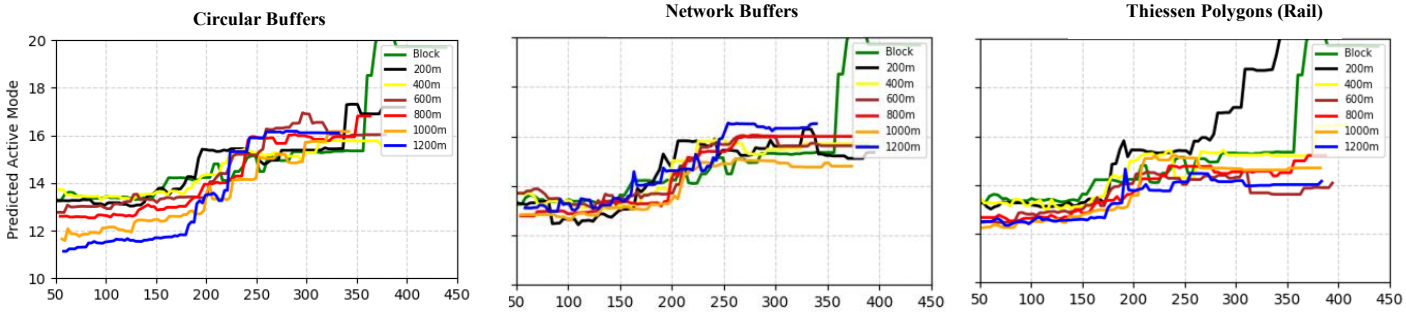
In this section, we examine the impacts of data aggregation on the threshold and nonlinearity effects of BE's features on mode share. Within the limited space, we focused on the key BE features partial dependency plots. The rest of the analysis can be found in Appendix A7.

The results illustrate that nearly all BE features continue to impact mode shares, regardless of the data aggregation methods used, in terms of shape and size. For instance, there is a positive link between intersection density and active modes for both rail and BRT stations across all four different data aggregation methods (Fig. 4a-b).

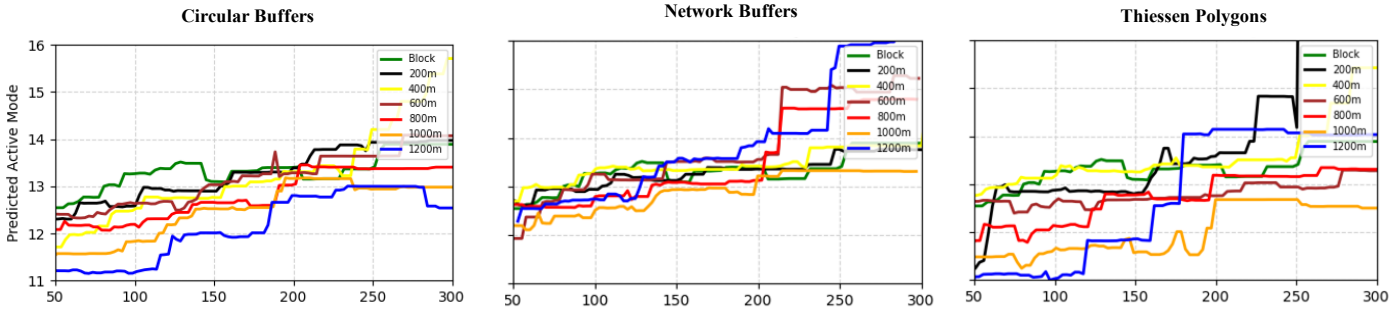
Additionally, the results underscore that the effective range of certain built environment features at the station level, such as employment density (Fig. 4c-d), transit service frequency, and station density, on mode share can vary according to the method, size, and shape of data aggregation around both station types. For instance, the effective range of employment density on active mode share is 0-100 at the block group level, 0-200 at the 800-meter circular buffer, 0-180 at the 800-meter network buffer, and 0-100 at the 800-meter Thiessen polygon around rail stations (Fig. 4c).

However, the findings also underscore that the impact of some BE features and urban form, including population density at the regional level (Fig. 4e), on commuting mode share is less sensitive to the size, shape, and method of data aggregation. For instance, the effective range of the impact of distance from the CBD on private vehicle mode share is 5 km and 20 km for BRT and rail stations respectively across all four different methods of

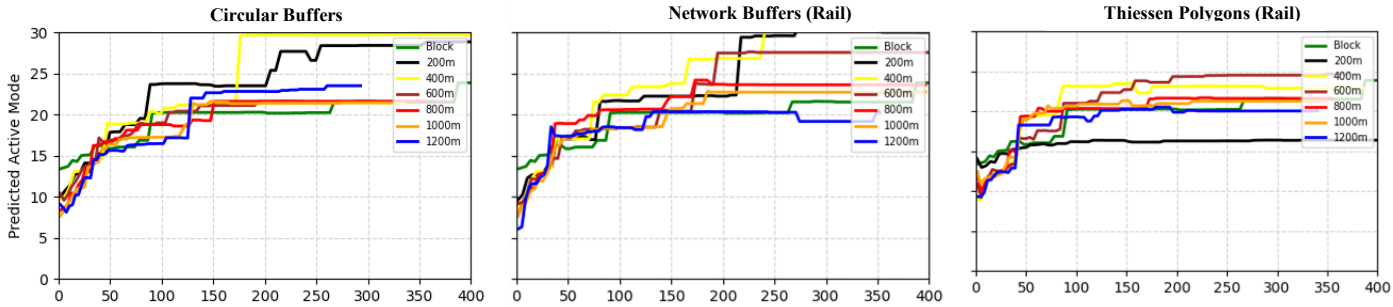
data aggregation. Additionally, the effective range of population density at regional level on private vehicle mode share is 0-600 people per square kilometer for rail catchment areas across all different models.



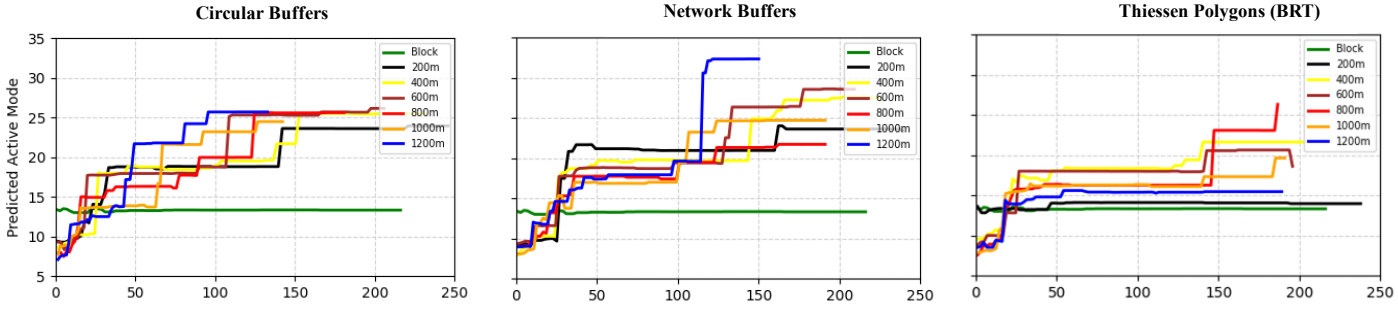
4-A. Intersection density (number of intersections in catchment area) around Rail stations for active modes



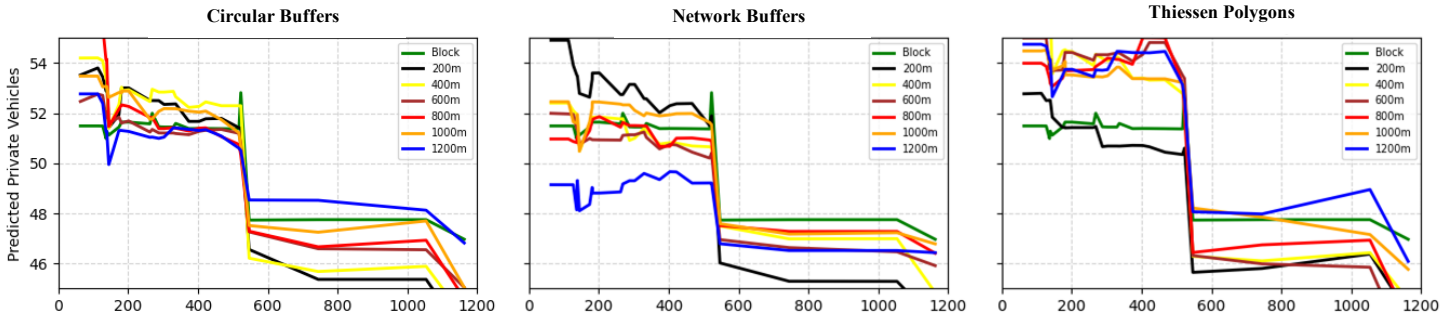
4-B. Intersection density (number of intersections in catchment area) around BRT stations for active modes



4-C. Employment density (number of jobs in catchment area) around rail stations for active modes



4-D. Employment density (number of jobs in catchment area) around BRT stations for active modes



4-E. Population density at regional level (number of populations in square kilometer) around Rail stations for Private vehicles

Figure 4. Nonlinearity association between BEs and mode share around rail and BRT station

4.2.2 Inferential models output

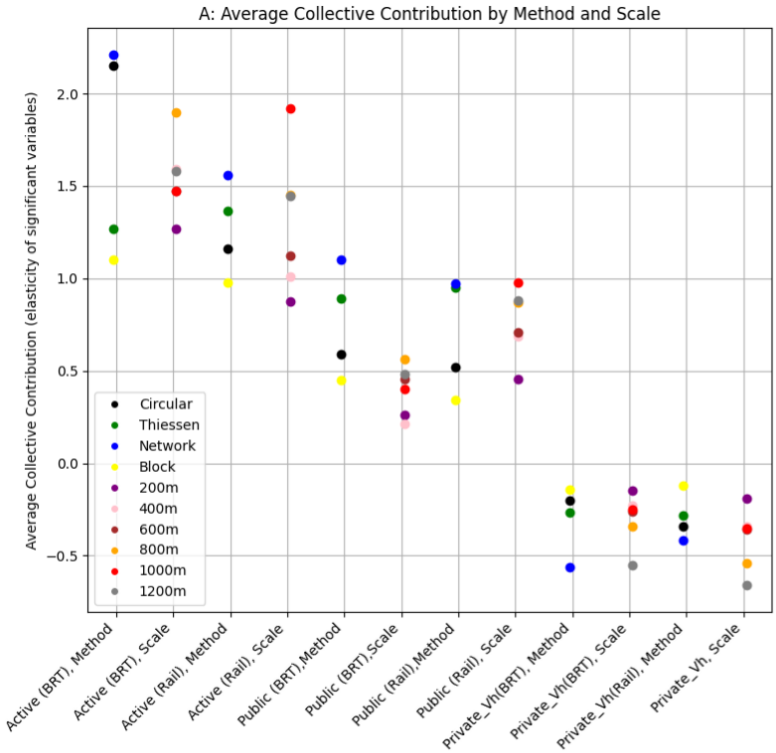
4.2.2.1 Changes in significant variables, and contribution of features

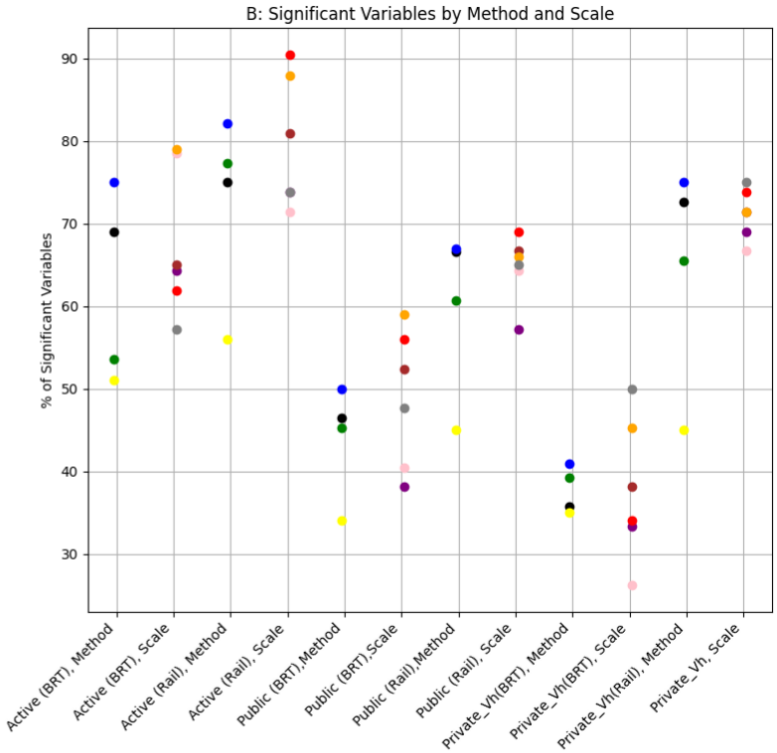
In this section, we present the results, highlighting changes in significant variables, and contributions of BE features at different scales and Methods. The complete results of inferential approaches can be found in Appendix 8.

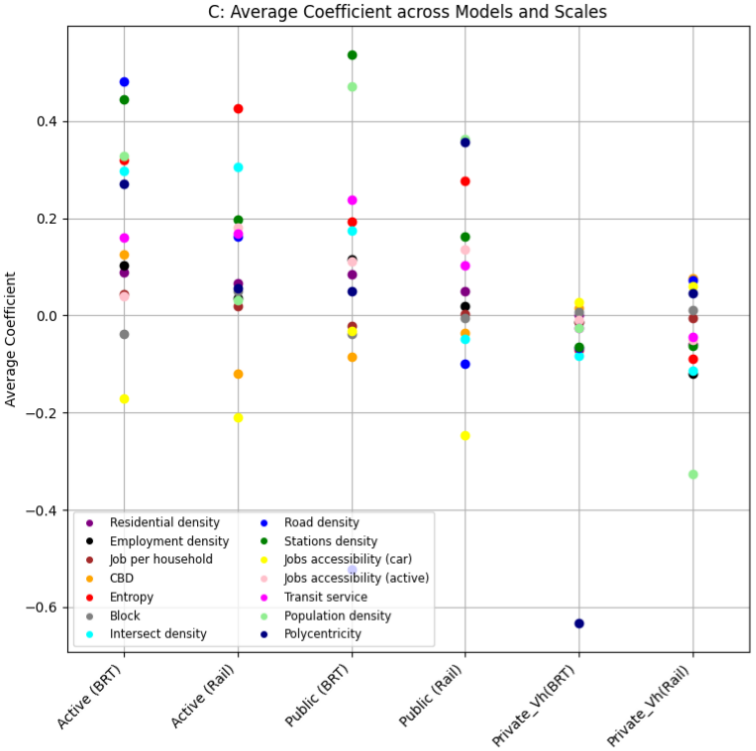
The inferential model results suggest that data aggregation has a moderate impact on the analysis (Figure 5). Specifically, the collective contribution of BE features, the number of significant variables, and the coefficient variance across models and scales remain relatively consistent. For instance, the average collective contribution (estimated elasticity) of BE features for rail's active mode ranges from 0.98 (block group) to 1.2 (circular buffers), 1.4 (Thiessen polygons) and 1.55 (network buffers) (Figure 5: A).

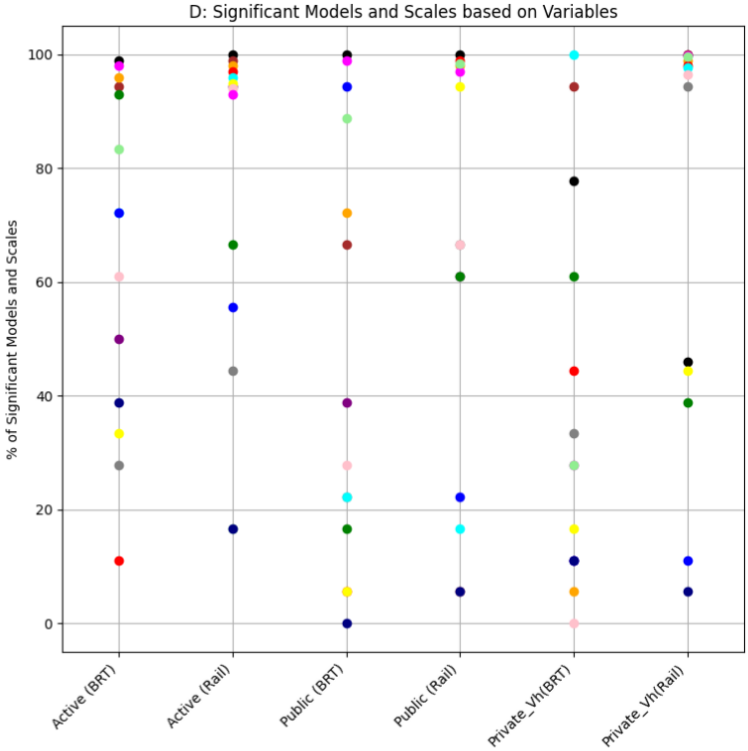
The results also highlighted that specific scales and methods capture more significant BE (built environment) features, collectively contributing more and showing lower variance in terms of significant coefficients of BE variables (Figure 5: A, B, E). For active and public transit modes, the optimal scale was 800 meters for BRT and 1000 meters for rail stations, while 1200 meters worked well for both rail and BRT stations. For example, in BRT public transit models, the proportion of significant variables within the 800-meter buffer was around 70%, nearly 5% higher than other scales. Additionally, network buffer methods demonstrated greater consistency. For instance, the variance of significant coefficients for rail's private vehicle models was 0.1 using the network method, compared to a variance of approximately 0.2 for circular and Thiessen methods (Figure 5: E).

Another key finding is that some BE features, such as employment density, jobs per household, and job accessibility (active), are less sensitive to data aggregation methods and scales with respect to estimated elasticity and significance. For instance, intersection density was significant across nearly all of rail's active mode models, regardless of scale or method, with an elasticity of 0.3—above the average (Figure 5: C and D).


A: Average Collective Contribution by Method and Scale
Average Collective Contribution (elasticity of significant variables)
Circular
Thiessen
Network
Block
200m
400m
600m
800m
1000m
1200m
Active (BRT), Method
Active (BRT), Scale
Active (Rail), Method
Active (Rail), Scale
Public (BRT),Method
Public (BRT),Scale
Public (Rail),Method
Public (Rail), Scale
Private_Vh(BRT), Method
Private_Vh(BRT), Scale
Private_Vh(Rail), Method
Private_Vh, Scale


B: Significant Variables by Method and Scale
% of Significant Variables
Active (BRT), Method
Active (BRT), Scale
Active (Rail), Method
Active (Rail), Scale
Public (BRT),Method
Public (BRT),Scale
Public (Rail),Method
Public (Rail), Scale
Private_Vh(BRT), Method
Private_Vh(BRT), Scale
Private_Vh(Rail), Method
Private_Vh, Scale


C: Average Coefficient across Models and Scales
Average Coefficient
Residential density
Employment density
Job per household
CBD
Entropy
Block
Intersect density
Road density
Stations density
Jobs accessibility (car)
Jobs accessibility (active)
Transit service
Population density
Polycentricity
Active (BRT)
Active (Rail)
Public (BRT)
Public (Rail)
Private_Vh(BRT)
Private_Vh(Rail)


D: Significant Models and Scales based on Variables
% of Significant Models and Scales
Active (BRT)
Active (Rail)
Public (BRT)
Public (Rail)
Private_Vh(BRT)
Private_Vh(Rail)

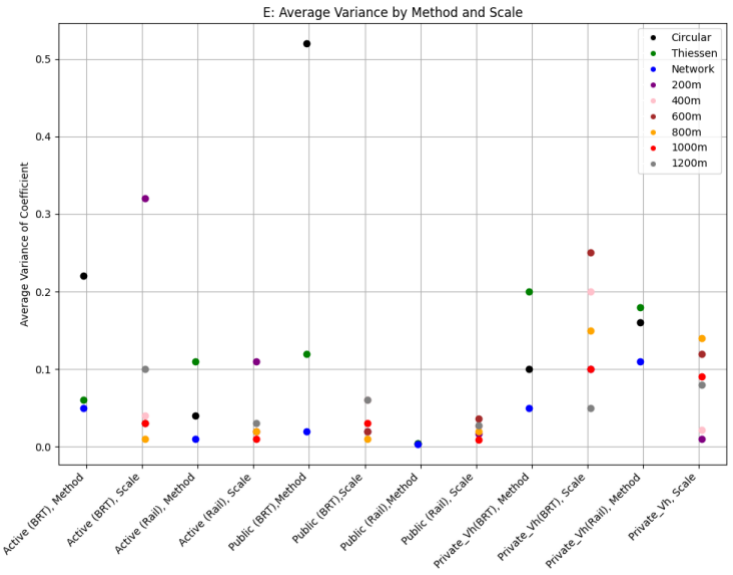


Figure 5. Sensitivity of BE-Mode share models based on HLM model's output

5 Discussion

This study contributes to the literature by examining the impact of data aggregation methods, as well as the size and shape of catchment areas, on BE-mode-share model outcomes for rail and BRT stations. Using both inferential and machine learning approaches, it offers practical insights into the optimal size and shape of catchment areas for rail versus BRT stations, identifies key built environment features that remain robust across different data aggregation methods, and highlights the influence of data analysis techniques in mitigating aggregation bias. Tables 2 and 3 provide a summary of our findings.

Table 2. Summary of best model performance and most robust scale to data aggregation

Mode		Data aggregation type	Data analysis methods	Scale
Active and public transit	BRT	Network buffer based	XGBOOST and HLM	800-meter
	Rail	Network buffer based	XGBOOST and HLM	100-meter
Private vehicles	BRT	Network buffer based	XGBOOST and HLM	1200-meter
	Rail	Network buffer based	XGBOOST and HLM	1200-meter

Table 3. Summary of data aggregation impact on mode share models around transit stations

Metrics	Sensitive to data aggregation (method and size)	Sensitive to station type
Variable relative importance	**Yes**, the relative importance of majority of variables changes depending on the data aggregation	**Yes**, the relative importance of variables also differs between station types (rail vs. BRT)
Collective contribution of local and regional BEs	**No**, the collective contribution of local and regional BE features remains stable regardless of aggregation method or size.	**Yes**, the collective contribution of local and regional built environment factors may vary between station types.
Sign of coefficients	**Yes**, the sign of the coefficients is sensitive to data aggregation.	**No**, the signs of the coefficients are less sensitive to station types.
Effective range and coefficient size	**Yes**, for the majority of local-level BE features. No, for regional-level urban form.	**Yes**, the effective range and coefficient size are sensitive to station types.
Data analysis	**Yes**, both inferential and machine learning approaches are sensitive to data aggregation.	**No**, data analysis is not sensitive to station types.

Note: The following variables are less sensitive to data aggregation in mode share planning around rail and BRT stations: employment density, jobs per household, intersection density, residential density, distance to CBD, job accessibility (active), and regional population density.

5.1 Summary of findings

This study shows that the data aggregation process can impact BE-commuting mode share model outputs within rail and BRT catchment areas. However, machine learning approaches are more sensitive to data aggregation for both rail and BRT stations' models. This finding aligns with previous studies in the field (Barri et al., 2022; Chen et al., 2021; Cheng et al., 2020). The reason behind this may be that machine learning approaches relax the linearity assumption and can capture intricate nonlinear associations between outcomes and predictors (Liu et al., 2023).

Additionally, data aggregation affects all three main mode share models—active, public transit, and private vehicles—though the degree of impact varies. Models with better performance tend to be more stable and robust to data aggregation effects. The results also show that network-buffer-based models outperform others, demonstrating greater consistency in terms of relative importance, significant variables, and the estimated elasticity of BE features. This finding is comparable with previous studies (Oliver et al., 2007; Wu et al., 2021; Yang et al., 2021). The reason behind this finding may be that circular buffers may fail to consider how the existing road network restricts access to destinations, thus not addressing the concept of accessibility effectively and failing to represent the real features of catchment areas (Oliver et al., 2007). In addition, while Thiessen polygons help tackle the overlapping of catchment areas, they can create clusters of small polygons, especially where public transit stations are close to each other, leading to potential inaccuracies in variable calculations (Yang et al., 2021). Conversely, network buffers account for the actual travel routes people may take, providing a more accurate representation of accessible areas for individuals using the existing road and path networks to determine reachable areas within a specified distance.

Also, for active and public transit modes, the optimal scale was 800 meters for BRT and 1000 meters for rail stations, while 1200 meters worked well for both rail and BRT stations' private vehicle models in mitigating data aggregation bias. These findings are supported by consistent predictor contributions in both machine learning and inferential

models, superior model performance in both approaches, and a higher number of significant BE features identified by the inferential models. This finding is comparable to Guerra et al. 2012, who suggested that for transit ridership in the United States, the size of catchment areas has little impact on a model's predictive power (e.g., a 0.25-mile radius explains U.S. transit ridership variation as well as a 0.5-mile or 0.75-mile radius). The inconsistency may be attributed to the method of data aggregation (network buffer and Thiessen polygon vs. circular buffers), the lack of distinction between BRT and rail catchment areas and the method of data analysis (inferential vs machine learning), as well as differences in the nature of the outcome variables (transit ridership vs. commuting mode share models).

Furthermore, the key BE features in commuting mode share models—employment density, jobs per household, intersection density, residential density, distance from the CBD, job accessibility (active), and population density at the regional level—demonstrate consistency in both machine learning and inferential models. These features maintain relative importance across different scales in the ML models and exhibit significant effects with more consistent estimated elasticity in the inferential models. Furthermore, the collective contribution of local (station) BE features to active and private vehicle mode share models is not sensitive to data aggregation for both rail and BRT models and dominates socio-economic features and regional urban form. This may be because local features, such as street connectivity and proximity to destinations, are critical for determining active travel and vehicle use. Similarly, the collective contribution of regional urban form in predicting public transit mode share for rail catchment areas is not sensitive to data aggregation and dominates other predictors. This might be due to the broader impact of regional accessibility and land use patterns on public transit use.

Our findings are comparable to those of Tao and Cao (Tao & Cao, 2023), who applied a nonlinear approach, and Næss (Næss, 2011) who used a linear approach, both concluding that regional urban form has a larger impact on travel behavior (using private vehicles) compared to local urban features. Additionally, our results align with Ewing and Cervero (Ewing & Cervero, 2010) who concluded that active travel (e.g., walking) is more related to the local built environment than the regional urban form. The importance of local built environment features for active travel may be because local urban form, such as the presence of facilities, infrastructure, and destinations around transit stations, is crucial for walking and cycling, regardless of the data aggregation process.

In addition, the results highlight that the overall impact (direction) of BE features on mode share models is not sensitive to data aggregation. This is evident from the stability in the direction of BE's impact (sign of coefficient in inferential models) and the consistent non-linear effect of BE features across different data aggregations in ML models. However, the effective range of some local BE features, such as stations density, is sensitive to the data aggregation process, which is comparable with Gu et al. (2024) study in which they found that the land use features may impact TOD sites differently at different buffer sizes.

5.2 Implications for research and practice

The findings of this study have important implications for urban and transportation planning, particularly in TOD planning. First, the results suggest that urban and transportation planners should account for the sensitivity of BE-mode share model outcomes around public transit stations, regardless of the data analysis approach used. Planners should extract data using multiple methods and across various scales, then run

different models to ensure robust results. This comprehensive approach will support more informed policy development.

Also, the study highlights the importance of analyzing different travel behavior outcomes (active transport, public transit, private vehicles). Practitioners can apply this insight by integrating these models into regional transportation planning to ensure a balanced approach to mobility. For instance, rather than focusing solely on increasing transit ridership, policies could also prioritize shifting private vehicle use to more sustainable modes, such as walking, cycling, or shared mobility services. This could involve supporting bike-sharing programs, improving pedestrian infrastructure, or implementing congestion pricing to reduce car use around transit stations.

In addition, network buffers are recommended over circular buffers, Thiessen polygons, and administrative boundaries (e.g., block groups) for defining rail and BRT catchment areas in TOD planning, as they are less sensitive to data aggregation. Alos, practitioners involved in infrastructure design can use network-buffer based models to identify the most critical factors (e.g., accessibility, residential density, intersection density) in a given area and direct infrastructure investments accordingly. For instance, they can focus on enhancing the connectivity and accessibility of the transportation network (e.g., improving road networks, adding bike lanes, upgrading transit stations) within the key buffer zones around transit stations to optimize their performance.

Urban and transportation planners should also consider the optimal size of catchment areas based on the type of transportation system (rail vs. BRT) and the study's purpose. For example, the optimal scale was 800 meters for BRT and 1000 meters for rail stations in active and public transit modes, while 1200 meters worked well for both rail and BRT stations. Additionally, policymakers and planners can use the identified optimal buffer scales to define catchment areas around transit stations more accurately according to the type of stations. This can improve infrastructure investment and service provision by ensuring that transportation facilities like bus stops, bike-sharing stations, or park-and-ride lots are appropriately located to maximize transit use and reduce car dependency.

Finally, the results suggest that urban and transportation planners should consider employment density, jobs per household, intersection density, residential density, distance from the CBD, job accessibility (active), and regional population density as key predictors for mode share planning within public transit catchment areas, regardless of data aggregation. Additionally, focusing on land use features as a group, e.g., at the local (station) level, may be more effective for guiding land use development, as these features are less sensitive to data aggregation.

5.3 Limitations and future directions

We acknowledge that this study has some limitations. First, we considered administrative boundaries, circular buffers, network buffers, and Thiessen polygons for data aggregation. However, these methods may not best represent public transit catchment areas, and exploring other methods could provide new insights. Secondly, we also acknowledge the limitations of census data and do not include micro variables that affect people's travel behavior, such as street width, greenness, and the number of parking spaces. Another limitation of this study is the inability to distinguish between suburban rail stations and metro/subway stations due to data constraints. While these station types may have different effects on travel behavior, the available datasets do not provide sufficient granularity to make this distinction.

To better understand the impact of data aggregation methods, size, and shape on travel behaviors around public transit stations, future research could explore new data

aggregation methods, such as square-based and hexagon-based zones. Additionally, including more transit types, such as buses, suburban rail stations, and examining more outcomes could also be beneficial. Also, future research should extend this analysis to disaggregate modeling as well, particularly in examining how data aggregation impacts individual or household-level mode choice decisions. Investigating these effects could provide deeper insights into behavioral variations.

6 Conclusion

This study highlights the significant impact of data aggregation on BE-mode share models across 2,794 rail and BRT stations in the United States, utilizing both inferential and machine learning methods. The findings demonstrate that different aggregation methods and buffer sizes influence model outcomes. Network buffers provide more stable results compared to circular buffers, Thiessen polygons, and administrative boundaries. The study also identifies optimal buffer sizes—800 meters for BRT and 1000 meters for rail stations for active and public transit modes, and 1200 meters for private vehicle mode share—helping to minimize sensitivity to aggregation effects. Additionally, key BE features such as employment density, jobs per household, intersection density, and regional population density remain robust across different aggregation approaches.

In summary, this study emphasizes the importance of the data aggregation process in commuting mode share studies for rail and BRT station areas. Urban planners, researchers, and practitioners should be mindful of aggregation bias when studying BE-mode share around public transit stations and applying the results in practice. They should test multiple buffer sizes and aggregation methods for sensitivity to effectively tailor land-use intervention policies and promote active modes of transportation.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback.

Author contribution

The authors confirm their contribution to the paper as follows: conceptualization, data curation, formal analysis, methodology, software, visualization, writing–original draft: Seyed Sajjad Abdollahpour; investigation, writing–review & editing: Huyen T.K. Le; investigation, writing–review & editing: Ralph Buehler; investigation, supervision, validation, writing–review & editing: Steve Hankey

Appendices

Appendices available as a supplementary file at https://doi.org/10.5198/jtlu.2025.2676.

References

Abdollahpour, S. S., Buehler, R., Le, H. T., Nasri, A., & Hankey, S. (2024). Built environment's nonlinear effects on mode shares around BRT and rail stations. *Transportation Research Part D: Transport and Environment*, *129*, 104143.

Abdollahpour, S. S., Le, H. T., & Hankey, S. (2025). Changes in the predictors of transit ridership in post-COVID-19 US metropolitan areas. *Travel Behavior and Society*, *40*, 101002.

Aghaabbasi, M., & Chalermpong, S. (2023). Machine learning techniques for evaluating the nonlinear link between built-environment characteristics and travel behaviors: A systematic review. *Travel Behavior and Society*, *33*, e100640-e100640.

Akbari, P., & Bafarasat, A. Z. (2024). Exploring energy efficiency in historical urban fabrics for energy-conscious planning of new urban developments. *Journal of Urban Planning and Development*, *150*(2), 04024011.

Alpaydin, E. (2020). *Introduction to machine learning*. Cambridge, MA: MIT press.

Babapourdijojin, M., Corazza, M. V., & Gentile, G. (2024). Systematic analysis of commuting behavior in Italy using K-means clustering and spatial analysis: Towards inclusive and sustainable urban transport solutions. *Future Transportation*, *4*(4), 1430–1456.

Barri, E. Y., Farber, S., Jahanshahi, H., & Beyazit, E. (2022). Understanding transit ridership in an equity context through a comparison of statistical and machine learning algorithms. *Journal of Transport Geography, 105*, 103482.

Boarnet, M. G., & Crane, R. (2001). *Travel by design: The influence of urban form on travel*. Oxford, UK: Oxford University Press.

Braun, L. M., Rodriguez, D. A., Cole-Hunter, T., Ambros, A., Donaire-Gonzalez, D., Jerrett, M., …, & de Nazelle, A. (2016). Short-term planning and policy interventions to promote cycling in urban centers: Findings from a commute mode choice analysis in Barcelona, Spain. *Transportation Research Part A: Policy and Practice*, *89*, 164–183.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, *34*, 548–558.

Chen, E., Ye, Z., & Wu, H. (2021). Nonlinear effects of built environment on intermodal transit trips considering spatial heterogeneity. Transportation Research Part D: Transport and Environment, 90, 102677.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785-794).

Cheng, L., De Vos, J., Zhao, P., Yang, M., & Witlox, F. (2020). Examining non-linear built environment effects on elderly's walking: A random forest approach. *Transportation research part D: transport and environment, 88,* 102552.

Clark, A., & Scott, D. (2014). Understanding the impact of the modifiable areal unit problem on the relationship between active travel and the built environment. *Urban Studies*, *51*(2), 284–299.

Cui, B., Boisjoly, G., Miranda-Moreno, L., & El-Geneidy, A. (2020). Accessibility matters: Exploring the determinants of public transport mode share across income groups in Canadian cities. *Transportation Research Part D: Transport and Environment*, *80*, 102276.

De Vos, J., Cheng, L., Kamruzzaman, M., & Witlox, F. (2021). The indirect effect of the built environment on travel mode choice: A focus on recent movers. *Journal of Transport Geography*, *91*, 102983.

Ding, C., Cao, X. J., & Næss, P. (2018). Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. *Transportation Research Part A: Policy and Practice*, *110*, 107–117.

Duan, Y., Yuan, C., Mao, X., Zhao, J., & Ma, N. (2023). Influence of the built environment on taxi travel demand based on the optimal spatial analysis unit. *PLoS one*, *18*(10), e0292363.

Eldafrawi, M., Varghese, K., Afsari, M., Babapourdijojin, M., & Gentile, G. (2023). Predictive analytics for road traffic accidents: Exploring severity through conformal prediction. Paper presented at the 2024 TRB Annual Meeting, January 7–11, Washington DC, USA.

Eom, J. K., Choi, J., Park, M. S., & Heo, T.-Y. (2019). Exploring the catchment area of an urban railway station by using transit card data: Case study in Seoul. *Cities*, *95*, 102364.

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, *76*(3), 265–294.

Farber, S., & Marino, M. G. (2017). Transit accessibility, land development and socioeconomic priority: A typology of planned station catchment areas in the Greater Toronto and Hamilton Area. *Journal of Transport and Land Use*, *10*(1), 879–902.

Feudo, F. L. (2014). How to build an alternative to sprawl and auto-centric development model through a TOD scenario for the North-Pas-de-Calais region? Lessons from an integrated transportation-land use modelling. *Transportation Research Procedia*, *4*, 154–177.

Gao, F., Li, S., Tan, Z., Wu, Z., Zhang, X., Huang, G., & Huang, Z. (2021). Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment factors. *International Journal of Geographical Information Science*, *35*(9), 1905–1925.

Gao, F., Tang, J., & Li, Z. (2022). Effects of spatial units and travel modes on urban commuting demand modeling. *Transportation*, *49*(6), 1549–1575.

Gehrke, S. R., & Clifton, K. J. (2014). Operationalizing land use diversity at varying geographic scales and its connection to mode choice: Evidence from Portland, Oregon. *Transportation Research Record*, *2453*(1), 128–136.

Gu, X., Lin, S., & Wang, C. (2024). Integrated impact of urban mixed land use on TOD ridership: A multi-radius comparative analysis. *Journal of Transport and Land Use*, *17*(1), 457–481.

Guerra, E., Cervero, R., & Tischler, D. (2012). Half-mile circle: Does it best represent transit station catchments? *Transportation Research Record*, *2276*(1), 101–109.

Gutiérrez, J., Cardozo, O. D., & García-Palomares, J. C. (2011). Transit ridership forecasting at station level: An approach based on distance-decay weighted regression. *Journal of Transport Geography*, *19*(6), 1081–1092.

Henao, A., Piatkowski, D., Luckey, K. S., Nordback, K., Marshall, W. E., & Krizek, K. J. (2015). Sustainable transportation infrastructure investments and mode share changes: A 20-year background of Boulder, Colorado. *Transport Policy*, *37*, 64–71.

Hong, J., Shen, Q., & Zhang, L. (2014). How do built-environment factors affect travel behavior? A spatial analysis at different geographic scales. *Transportation*, *41*, 419–440.

Hox, J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications.* Routledge.

Jamme, H.-T., Rodriguez, J., Bahl, D., & Banerjee, T. (2019). A twenty-five-year biography of the TOD concept: From design to policy, planning, and implementation. *Journal of Planning Education and Research*, *39*(4), 409–428.

Jian, W., Liu, X., Liu, H., Hu, Y., & Gao, L. (2023). The impacts of the multiscale built environment on commuting mode choice: Spatial heterogeneity, moderating effects, and implications for demand estimation. *Journal of Advanced Transportation*, *2023*(1), 9346631.

Khalil, M. A., & Fatmi, M. R. (2025). How effective are discrete-continuous multi-task learning compared to single-output models? Insights from travel mode and departure time analysis. *Expert Systems with Applications*, 127002.

Kuby, M., Barranda, A., & Upchurch, C. (2004). Factors influencing light-rail station boardings in the United States. *Transportation Research Part A: Policy and Practice*, *38*(3), 223–247.

Laviolette, J., Morency, C., & Waygood, E. (2022). A kilometer or a mile? Does buffer size matter when it comes to car ownership? *Journal of Transport Geography*, *104*, 103456.

Li, S., Lyu, D., Huang, G., Zhang, X., Gao, F., Chen, Y., & Liu, X. (2020). Spatially varying impacts of built environment factors on rail transit ridership at station level: A case study in Guangzhou, China. *Journal of Transport Geography*, *82*, 102631.

Li, T., Zhang, M., Jiang, H., & Jing, P. (2022). Understanding the modifiable areal unit problem and identifying appropriate spatial units while studying the influence of the built environment on the traffic system state. *Journal of Advanced Transportation*, *2022*(1), 8288248.

Li, Z., Tang, J., Ji, Y., Liang, X., Hu, L., & Hu, C. (2025). Relationship between the built environment and metro usage patterns: A motif-based perspective. *Tunneling and Underground Space Technology*, *159*, 106488.

Liu, X., Chen, X., Tian, M., & De Vos, J. (2023). Effects of buffer size on associations between the built environment and metro ridership: A machine learning-based sensitive analysis. *Journal of Transport Geography*, *113*, 103730.

Liu, Y., Nath, N., Murayama, A., & Manabe, R. (2022). Transit-oriented development with urban sprawl? Four phases of urban growth and policy intervention in Tokyo. *Land Use Policy*, *112*, 105854.

Loo, B. P., Chen, C., & Chan, E. T. (2010). Rail-based transit-oriented development: Lessons from New York City and Hong Kong. *Landscape and Urban Planning*, *97*(3), 202–212.

Luo, C., Hu, Y., & Wang, F. (2025). A big data approach to mitigating the MAUP in measuring excess commuting. *Computational Urban Science*, *5*(1), 14.

Mitra, R., & Buliung, R. N. (2012). Built environment correlates of active school transportation: Neighborhood and the modifiable areal unit problem. *Journal of Transport Geography*, *20*(1), 51–61.

Molnar, C. (2020). *Interpretable machine learning. A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/

Næss, P. (2011). 'New urbanism' or metropolitan-level centralization? A comparison of the influences of metropolitan-level and neighborhood-level urban form characteristics on travel behavior. *Journal of Transport and Land Use*, *4*(1), 25–44.

Nakshi, P., & Debnath, A. K. (2021). Impact of built environment on mode choice to major destinations in Dhaka. *Transportation Research Record*, *2675*(4), 281–296.

Nasri, A., & Zhang, L. (2014). The analysis of transit-oriented development (TOD) in Washington, DC and Baltimore metropolitan areas. *Transport Policy*, *32*, 172–179.

Nasri, A., & Zhang, L. (2019). Multi-level urban form and commuting mode share in rail station areas across the United States: A seemingly unrelated regression approach. *Transport Policy*, *81*, 311–319.

Oliver, L. N., Schuurman, N., & Hall, A. W. (2007). Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. International Journal of Health Geographics, 6, 1–11.

Openshaw, S. (1984). *The modifiable areal unit problem. Concepts and techniques in modern geography.* Norwich: Geo Books.

Pan, Q., & Sharifi, S. (2024). Third step of four step modeling (mode choice models). *Transportation Land Use Modeling and Policy (TLUMP)*. Retrieved from https://open.umn.edu/opentextbooks/textbooks/transportation-land-use-modeling-and-policy-tlump

Pani, A., Sahu, P. K., Chandra, A., & Sarkar, A. K. (2019). Assessing the extent of modifiable areal unit problem in modelling freight (trip) generation: Relationship between zone design and model estimation results. *Journal of Transport Geography*, *80*, 102524.

Papadakis, D. M., Savvides, A., Michael, A., & Michopoulos, A. (2024). Advancing sustainable urban mobility: Insights from best practices and case studies. *Fuel Communications*, *20*, 100125.

Park, K., Ewing, R., Scheer, B. C., & Tian, G. (2018). The impacts of built environment characteristics of rail station areas on household travel behavior. *Cities*, *74*, 277–283.

Renne, J. L., Hamidi, S., & Ewing, R. (2016). Transit commuting, the network accessibility effect, and the built environment in station areas across the United States. *Research in Transportation Economics*, *60*, 35–43.

Sun, L.-S., Wang, S.-W., Yao, L.-Y., Rong, J., & Ma, J.-M. (2016). Estimation of transit ridership based on spatial analysis and precise land use data. *Transportation Letters*, *8*(3), 140–147.

Tao, T., & Cao, J. (2023). Exploring nonlinear and collective influences of regional and local built environment characteristics on travel distances by mode. *Journal of Transport Geography*, *109*, 103599.

Tao, T., Wu, X., Cao, J., Fan, Y., Das, K., & Ramaswami, A. (2023). Exploring the nonlinear relationship between the built environment and active travel in the twin cities. *Journal of Planning Education and Research*, *43*(3), 637–652.

Tian, G., Kalantari, H. A., & Ewing, R. (2023). Are older adults living in compact development more active? Evidence from 36 diverse regions of the United States. *Computational Urban Science*, *3*(1), 10.

Wade, C., & Glynn, K. (2020). *Hands-on gradient boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd.

Wey, W.-M., & Huang, J.-Y. (2018). Urban sustainable transportation planning strategies for livable City's quality of life. *Habitat International*, *82*, 9–27.

Wu, X., Lu, Y., Gong, Y., Kang, Y., Yang, L., & Gou, Z. (2021). The impacts of the built environment on bicycle-metro transfer trips: A new method to delineate metro catchment area based on people's actual cycling space. *Journal of Transport Geography*, *97*, 103215.

Xiao, W., & Wei, Y. D. (2023). Assess the non-linear relationship between built environment and active travel around light-rail transit stations. *Applied Geography*, *151*, 102862.

Yang, H., Li, X., Li, C., Huo, J., & Liu, Y. (2021). How do different treatments of catchment area affect the station level demand modeling of urban rail transit? *Journal of Advanced Transportation*, *2021*, 1–19.

Yang, L., Hu, L., & Wang, Z. (2019). The built environment and trip chaining behavior revisited: The joint effects of the modifiable areal unit problem and tour purpose. *Urban Studies*, *56*(4), 795–817.

Yang, W., & Chang, J. S. (2025). A quasi-experimental study of light rail transit on jobs-housing balance by regional typology: A case study of South Korea. *Journal of Transport Geography*, *124*, 104173.

Yin, Z., Li, W., Li, C., & Zheng, Y. (2025). The relationship between accessibility and land prices: A focus on accessibility to transit in the 15-min city. *Travel Behavior and Society*, *38*, 100914.

Zhang, M., & Kukadia, N. (2005). Metrics of urban form and the modifiable areal unit problem. *Transportation Research Record*, *1902*(1), 71–79.

Zhang, S., Li, Z., & Liu, Z. (2023). Examining built environment effects on metro ridership at station-to-station level considering circle heterogeneity: A case study from Xi'an, China. *Journal of Advanced Transportation*, *2023*.

Zhou, X., Sun, C., Niu, X., & Shi, C. (2022). The modifiable areal unit problem in the relationship between jobs–housing balance and commuting distance through big and traditional data. *Travel Behavior and Society*, *26*, 270–278.