The Journal of Transport and Land Use jtlu.org
Vol. 17 No. 1 [2024] pp. 881–901

JTLU

Understanding household VMT generation: A comparative analysis with traditional statistical models and a machine-learning approach

Guang Tian (corresponding author)
University of New Orleans
gtian@uno.edu

Bob Danton
University of New Orleans
jfdanton@my.uno.edu

Bin Li
Louisiana State University
bli@lsu.edu

Vijaya (VJ) Gopu
University of Louisiana at Lafayette and Louisiana Transportation Research Center
v.gopu@la.gov

Julius A. Codjoe
University of Louisiana at Lafayette and Louisiana Transportation Research Center
julius.codjoe@la.gov

Abstract: Planners and policy makers have long been interested in predicting people's travel behaviors, including the number of vehicle miles traveled (VMT) they generate. Reducing VMT has come to be seen as a key strategy for lowering greenhouse gas emissions and mitigating their associated health effects, as well as for increasing sustainability and equity within communities and on a global scale. Emerging machine learning methods such as boosted regression trees (BRT) allow for the identification of comparative influences of different factors as well as their nonlinear and threshold effects on travel outcomes, but studies comparing the results of these methods with traditional statistical regressions have been scarce. This study is the first to compare these methods' indications of the impacts of land-use patterns on VMT generation using a large multiregional dataset. The results indicate that the two methods perform similarly in predicting whether households generate any VMT and in predicting the number of VMT generated by those households that generate any. The results of the BRT model validate the statistical models' indications that built environmental variables contribute significantly to VMT production, while further allowing for the identification of nonlinear impacts. Key thresholds and nonlinear effects of land-use variables on household VMT generation are identified from the BRT model. Both models indicate that land-use patterns that are denser, more diverse, and have increased access to transit result in reductions of vehicular trips and overall VMT, while the BRT model provides effective thresholds for these variables useful for developing planning solutions.

Keywords: VMT, travel behavior, land use, built environment, machine learning, nonlinear

Article history:
Received: May 16, 2024
Received in revised form: September 25, 2024
Accepted: November 5, 2024
Available online: December 13, 2024

1 Introduction

Over the past few decades, there has been a shift in transportation system assessment from "how fast vehicles move" to "how well people's travel needs are met," from

https://doi.org/10.5198/jtlu.2024.2531
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.

"speed" to "mobility, accessibility, sustainability, and livability," and from level of service (LOS) to vehicle miles traveled (VMT) (Lee & Handy, 2018). VMT, generally measured per capita, has become recognized as an important gauge of transportation systems, measuring both the economic growth associated with increased vehicular travel as well as the negative effects thereof, such as greenhouse gas emissions and traffic fatalities. As metropolitan planning organizations (MPOs) and governments of all scales have become increasingly aware of the pressures of global climate change, reduction of greenhouse gas emissions has emerged as a top planning priority, and VMT as an important metric of emission-producing activity. Transportation accounts for 27% of all U.S. greenhouse gas emissions, more than any other contributor (U.S. Environmental Protection Agency, 2022). Overall VMT in the U.S. has been growing over the last 50 years, and long-term forecasting indicates it will increase at an average annual rate of 0.7% between 2019 and 2049 (Federal Highway Administration, 2021). Reduction of VMT via planning, policy, and development has therefore come to be seen as a key method available to planners and governments for reducing carbon emissions, slowing processes of climate change, and improving public health. Understanding the effects of various attributes of the built environment on total VMT generation has, as a result, become recognized as key to successful planning for more sustainable futures.

This study aims to provide planners and policymakers with an indication of land-use patterns' effects on VMT and provide an examination of how various models can be used in combination to aid in identifying key factors to target in planning processes aimed at limiting the negative externalities of automobile use. The objectives are: 1) to estimate household VMT models using both traditional statistical methods and a machine learning approach and 2) to evaluate and compare the prediction power, influential variables, and interpretive values indicated by both statistical models and machine learning models.

These objectives were carried out using a travel database consisting of over 1,000,000 trips generated by over 100,000 households in 36 U.S. metropolitan areas. It contains 18 consistently defined and calculated demographic, built environmental, and regional variables. Statistical models (multilevel logistic and linear regressions) and a machine learning model (boosted regression trees (BRT)) were employed to estimate household VMT using R (a popular open-source software for statistical computing and graphics). Their prediction performances were then compared with evaluation metrics (e.g., AUC, r^2, and RMSE), and K-fold cross-validation was used to compare model performance. The study results indicate that both models are similarly effective in demonstrating significant impacts of land-use variables on vehicle trip and VMT generation and provide strong evidence that compact and diverse development patterns can be effective tools to limit VMT and its negative environmental and social effects.

2 Literature review

2.1 Traditional statistical models

Most of the studies produced examining relationships between VMT and the built environment have used traditional statistical and econometric models to deduce and quantify these relationships, including ordinary least squares (OLS), analysis of variance, linear and logistic regression, structural equation modeling, and various forms of spatial analysis. The built environment is generally quantified via measures of so-called "D" variables, namely density (of population, employment, etc.), diversity (of land uses), design (of intersections, road networks, blocks, etc.), destination accessibility, and distance to transit (Cervero & Kockelman, 1997; Ewing & Cervero, 2001; Ewing & Cervero, 2010).

The results of these studies have been mixed; however, the sheer number of studies carried out has facilitated the production of meta-analyses which allow for easier interpretation of trends across the literature. Ewing and Cervero have produced two, both of which found significant combined weighted elasticities of "D" variables on VMT (Ewing & Cervero, 2001; Ewing & Cervero, 2010). Salon et. al. noted issues regarding differences in effects' scales between studies due to differing local contexts, as well as an inability to deduce causal associations from many study designs, but nevertheless found that the impacts of built environment factors on VMT were significant and salient across studies in their meta-analysis (Salon et al., 2012). These studies all support the conclusion that improved access to destinations, denser neighborhoods with good street connectivity and diverse land uses can effectively lower VMT. Stevens' meta-analysis likewise found that people "tend to drive less when D-variables change in the direction of compact development," but questioned the significance levels of those impacts (Stevens, 2017).

Most existing research has been limited to localized geographic contexts, presenting issues in terms of external validity. However, a handful of multi-region studies have been produced. Cervero and Murakami analyzed 370 U.S. urbanized areas, finding a net elasticity of population density on VMT of -0.381 after accounting for certain traffic-*inducing* effects of density, and moderate elasticities for other built environment variables (Cervero & Murakami, 2010). Other multiregional studies also found significant effects of built environment variables on VMT across regions and support arguments that compact, mixed-use developments with good connectivity could lead to VMT reduction (Ewing, et al., 2015; Nasri & Zhang, 2012; Zhang et al., 2012). Nasri and Zhang later focused on the macro-level built environment, or overall urban form of metropolitan regions, concluding that effective land-use policy decisions should be made at the metropolitan (rather than local or neighborhood) scale (Nasri & Zhang, 2014).

Newer trends in the literature among studies using traditional statistical methods have addressed the need to control for self-selection, as well as the development of longitudinal studies. Zhang and Zhang identified a "second wave" of built environment-VMT studies since 2010 which have addressed residential self-selection, whereby neighborhood and housing preferences of individuals and households may confound statistical impacts of built environment-travel outcome correlations (Zhang & Zhang, 2020). Recent studies have just begun to address the need for longitudinal analysis to explore changes in VMT over time, in attempts to better understand causal relationships between built environment changes and changes in VMT (Ewing et al., 2014; Ihlanfeldt, 2020).

2.2 Machine learning approaches

Most recently, studies on the relationships between the built environment and travel behavior have begun to turn to emerging machine learning (ML) models to better model non-linear relationships between built environment variables and travel outcomes. Unlike traditional statistical models, which usually have assumed functional form (e.g., linear model), machine learning algorithms approximate the unknown relationship between the response and predictors through complicated functions without assuming any particular functional forms (Breiman, 2001). The fundamental question to address is whether the same effects of the built environment on travel behavior can be expected in different urban contexts. For example, can the same VMT reduction be expected from doubling population density in low density neighborhoods as in high-density neighborhoods? The idea that land-use variables may have differing impacts within different ranges is known as nonlinear or threshold effects (Galster, 2018). Since 2018, these studies have used new techniques including boosted regression trees (BRT), gradient boosting decision trees

(GBDT), generalized additive mixed models (GAMM) and random forest (RF) algorithms to model non-linear and threshold effects of built environment variables on a wide range of transportation outcomes, including vehicular and pedestrian crashes (Ding, Chen, & Jiao, 2018; Rui et al., 2022) active transit (Lui et al., 2021; Tao et al., 2020; Wali et al., 2021; Xiao & Wei, 2023), transit ridership (Barri et al., 2022; Ding et al., 2019; Du et al., 2022; Gan et al., 2020) and others. However, as of the time of writing only a few of these studies have dealt explicitly with the connections between the built environment and VMT. Two studies carried out in Norway both used GBDT and found significant non-linear and threshold effects of both density and distance (to local, second order, and city centers) on VMT (Ding, Cao, & Næss, 2018; Tao & Næss, 2022). Another used GBDT to explore the relationships between built environment variables and *electric* vehicle miles traveled (eVMT), finding strong effects from electric vehicle infrastructure variables as well as significant influence from distance to the nearest business district (Hu et al., 2021).

2.3 Comparative studies

In recent years there has been an emerging literature in the transportation field on the comparative analysis of ML and traditional methods in predicting various types of travel outcomes, beginning with a 2003 study by Xie et al., which compared decision tree and neural network models to a logit model and found slightly superior performances of those models in predicting mode choice (Xie et al., 2003). Zhao et al. recently noted that while several subsequent studies similarly focused on questions of mode choice have found superior predictive capabilities of ML methods over traditional models, there are distinct limitations in many of those studies in terms of behavioral analysis. They determined that in analyzing mode choice, there seems to be a trade-off between predictive accuracy (the ability of a model to predict a dependent variable based on given inputs) and behavioral soundness (e.g., correctly determining relative variable importance and directions of association between variables) between ML and traditional models (Zhao et al., 2020). Three recent studies have used recently released data on ride-hailing from Chicago to compare model performance in predicting and interpreting a variety of ride hailing and ride splitting outcomes (Li, 2022; Xu et al., 2021; Yan et al., 2020). These studies indicate that while ML models may outperform traditional models under some circumstances and provide important benefits in their ability to model nonlinear effects of socioeconomics and land-use patterns on travel outcomes, they in many cases lack interpretability and thus suggest the opportunity for the use of multiple models to confirm patterns and provide cross validation. Barri et. al. (2022) compared 8 models in their ability to predict travel patterns and transit trip generation among low-income populations in the Toronto, Canada area and similarly indicated trade-offs between improved model fit from ML methods and decreased interpretability. Their study went on to plot the spatial distributions of model differences to indicate potential differences in policy implications dependent on model choice and suggested potential benefits of adopting a mixed-model approach (Barri et al., 2022).

Studies comparing ML methods to traditional econometric and regression-based methods of predicting VMT outcomes in particular from built environment variables remain scarce. A notable exception is a recent study by Li & Kockelman, which compared several ML and traditional statistical methods to analyze model efficacy for determining VMT from built environment variables for a travel dataset from the Dallas-Ft. Worth region (Li & Kockelman, 2022). They compared the results of 10 different ML models to traditional models of OLS, multinomial logit, negative binomial, and spatial auto-regressive analysis. For each model, they calculated both model fit statistics (r^2 and

RMSE) and run-time calculations for a variety of independent variables and travel outcomes. They found that in modeling VMT and other continuous response variables (such as vehicle ownership) that traditional econometric methods worked reasonably well, but that GBDT proved the most effective. Overall, their results indicated that household size had a significant impact (~40%) on VMT, with the OLS model indicating those effects at 1.11 times higher than the ML models; similarly, they found that the OLS model indicated the significance of influence of the number of household workers at around 2 times more than the results of the ML methods, and the results of impacts of worker density showed very large differences between OLS and ML methods. Predictably, the ML models were able to identify certain threshold effects of these variables, whereas the OLS method indicated only linear relationships or certain nonlinear relationships that can be transferred to linear relationships.

While studies aiming to model and predict VMT have proliferated in the transportation literature for quite some time, the application of ML methods to the problem is still in its infancy. Very few studies have done so and those that have are severely limited in external validity due to their execution in single cities, particularly in Norway. No published studies to date have utilized a large multiregional data set alongside these new methodologies to model VMT generation. Furthermore, the prediction performance of ML on VMT has not been tested and systemically evaluated against traditional statistical models, except in one study which is limited in external validity by using data from a single metropolitan area. This study aims to fill the gap by comparing the results of a ML model to traditional regression modeling in predicting VMT generation by using a large, multiregional dataset comprising of detailed household data and consistently defined "D" variables from 36 diverse US metropolitan areas.

3 Methodology

3.1 Data and variables

Household travel survey data, compiled from daily travel diaries, is widely used by metropolitan planning organizations (MPOs) to model and forecast travel demand, and is an indispensable resource for scholars and planners studying travel behavior. To allow for accurate calculation of built environmental variables where people live, we contacted MPOs to assemble survey data from regions able to provide precise household XY locations. This proved a limiting factor in the scope of the dataset, as many MPOs were unwilling to provide precise geocodes due to confidentiality concerns and were thus excluded from the dataset. The resulting data (Table 1) contains 36 U.S. metropolitan regions, out of over 100 MPOs that were contacted, diverse in geographical location, urban form, and size. Additional GIS data including population and employment at the traffic analysis zone (TAZ) level, street networks, transit stops, travel time "skims," and parcel-level land use were collected from MPOs and county assessors for the same or adjacent years to the survey data for each region to allow for computation of built environment variables.

While the time span of these surveys is over 11 years, this is still a cross-sectional study due to the structure of the data. For a given region, we only have data for a single point in time; furthermore, for any given region the built environmental data and household travel data (including demographics) were matched to the same year. The reason for the range is that MPOs do not coordinate nationally in terms of what years they produce household travel surveys. Thus, despite the range of the data collection years across regions, this dataset is still distinctly cross-sectional.

Table 1. Combined data set from 36 regional household surveys

Regions	Survey date	Households	Trips		Regions	Survey date	Households	Trips
Albany, NY	2009	1,453	12,618		Minneapolis–St. Paul, MN-WI	2010	8,234	79,236
Atlanta, GA	2011	9,575	93,681		Orlando, FL	2009	866	7,315
Austin, TX	2005	1,448	14,265		Palm Beach, FL	2009	944	7,166
Boston, MA	2011	7,826	86,915		Phoenix, AZ	2008	4,314	37,811
Burlington, NC	2009	607	5,111		Portland, OR	2011	4,508	47,551
Charleston, SC	2009	243	2,098		Provo, UT	2012	1,464	19,255
Dallas, TX	2009	2,869	27,066		Richmond, VA	2009	612	5,123
Denver, CO	2010	5,551	55,056		Rochester, NY	2011	3,439	23,145
Detroit, MI	2005	939	14,690		Sacramento, CA	2012	1,766	16,048
Eugene, OR	2009	1,674	16,563		Salem, OR	2010	1,668	16,231
Greensboro, NC	2009	2,023	17,561		Salt Lake City, UT	2012	3,490	44,565
Hampton Roads–Norfolk, VA	2009	1,957	16,495		San Antonio, TX	2007	1,563	14,952
Houston, TX	2008	5,276	59,552		San Diego, CA	2012	1,478	15,248
Indianapolis, IN	2009	3,777	37,473		Seattle, WA	2014	4,965	47,877
Kansas City, MO	2004	3,022	31,779		Springfield, MA	2011	850	8,456
Los Angeles, CA	2012	13,605	130,422		Syracuse, NY	2009	654	5,752
Madison, WI	2009	138	1,316		Tampa, FL	2009	2,259	17,538
Miami, FL	2009	1,433	11,580		Winston-Salem, NC	2009	1,459	12,168
Total							107,949	1,059,678

To our knowledge, this represents the largest collection of household travel survey data ever assembled aside from the National Household Travel Survey (NHTS) and presents several advantages over the NHTS in terms of its use in analyzing VMT and other travel outcomes. First, it contains much larger samples for individual regions. Second, precise household XY geocodes enable accurate and consistent calculation of built environments at household locations. The NHTS, in contrast, geocodes households only at the census tract level, a significantly less precise level of geographic accuracy.

Table 2 presents definitions, descriptive statistics, and sample sizes for the variables in the study. Trip distance (used to compute total household VMT) was either provided or calculated with GIS from trip coordinates in the survey. Household travel survey methodology has been standardized in recent decades since the first NHTS in 1969, and it is very consistent in how trip distance and VMT are collected and estimated across MPOs. Household data including age, employment status, household size, and household income were likewise available from the survey data, allowing us to control for sociodemographic influences at the household level. Vehicle ownership was excluded as an independent variable because it was endogenous.

The variables calculated cover all built environmental "D" variables: density, diversity, design, distance to transit, and destination accessibility. They were computed around household locations, within a one-mile network distance buffer. The choice of buffer width was made based on an earlier finding that one-mile buffers offered superior predictive power over smaller buffer widths in predicting household travel outcomes (Ewing, et al., 2015). Destination accessibility (percent of regional employment within 30 minutes by automobile and transit) was calculated using TAZ data and travel time skims received from MPOs. For a given TAZ, we received data regarding the number of jobs within the TAZ and the travel time from the TAZ to each other TAZ. Employment was summed for every TAZ accessible from the household's TAZ within 30 minutes by the given mode and divided by the total jobs within the MPO. Regional control variables are

also included: metropolitan population, regional climate data from the National Centers for Environmental Information, and average gas prices, all for the same years as the survey data for each area.

18 total independent variables are available to predict household VMT generation, all consistently defined and calculated across each region, allowing for maximized external validity of the results. To check for multi-collinearity issues, variance inflation factor (VIF) tests were applied to the independent variables in the final models, and all VIF values were smaller than 5.0, indicating a low correlation between any given pair of independent variables.

Table 2. Variable definitions and descriptive statistics

Variable	Description	N	Mean	S.D.
Dependent variables				
anyvmt	any household VMT (1=yes, 0=no)	107,949	0.94	0.24
lnvmt	natural log of total household VMT (for households with any VMT)	100,676	3.10	1.06
Independent variables – household				
household size	household size	107,949	2.51	1.36
children	number of children (age<16) in the household	107,949	0.40	0.84
workers	number of household members employed	107,949	1.23	0.89
income	real household income (in 1000s of 2012 dollars)	107,949	78.54	53.50
Independent variables – built environment within a one-mile network buffer around household locations				
activity density	population + employment per square mile in 1000s	107,949	7.52	11.12
job-population balance*	the balance between job and population balance within one mile	107,949	0.63	0.25
Land-use entropy**	mix of different land uses (residential, commercial, and public)	107,949	0.49	0.26
intersection density	number of intersections within one mile	107,949	113.93	74.93
4-way intersection	percentage of 4-way intersections	107,949	26.71	17.49
transit stop density	number of transit stops within a one mile	107,949	26.96	55.61
rail station	rail station within half mile (1=yes, 0=no)	107,949	0.06	0.53
destinations in 30 minutes by auto	percentage of regional employment within 30 minutes by auto	107,949	47.99	30.86
destinations in 30 minutes by transit	percentage of regional employment within 30 minutes by transit	107,949	18.60	23.10
Independent variables – region				
reginal population	population within the region 1000s	36	2554.57	3183.23
gas price	average state gasoline price for the year of household travel data	36	2.91	0.15
low temperature	annual average of low temperature	36	45.52	13.14
high temperature	annual average of high temperature	36	72.82	9.22
precipitation	annual precipitation in inches	36	39.98	17.19

* jobpop = 1 – [ABS (employment – 0.2×population)/(employment + 0.2×population)] (Ewing, et al., 2015)

** entropy = -[residential share×ln (residential share) + commercial share×ln (commercial share) + public share×ln (public share)]/ ln (3), where ln is the natural logarithm of the value in parentheses and the shares are measured in terms of total parcel land areas (Ewing, et al., 2015)

3.2 Two-stage hurdle modeling

The dependent variable is total household VMT generation. Since the distribution of VMT is non-normally distributed and skewed towards higher values, the natural logarithm is used, consistent with previous studies (Ewing, et al., 2015). About 6% of households in the sample did not generate any automobile trips, and thus have zero values for the variable of household VMT. They may choose other modes, generating walking, biking, or transit trips. Households with zero values may be qualitatively

different from others, as choosing whether to participate in the generation of VMT may represent a separate decision-making process from choosing how much to drive for those who do. "The zero or nonzero values of the outcome is the result of a separate decision whether or not to "participate" in the activity. On deciding to participate, the individual decides separately how much to participate, that is, how intensively [to participate]" (Greene, 2018, p. 907). A solution is to estimate two-stage "hurdle" models (Ewing, et al., 2015; Greene, 2018). The first stage categorizes households as either producing VMT or not (dependent variable *anyvmt*), while the second stage estimates how much VMT is produced by households generating any (dependent variable *lnvmt*).

3.3 Statistical modeling

The data structure is nested, with households inside regions. Samples from the same region share regional characteristics, so multilevel modeling was needed to control for spatial effects. Multilevel models were estimated as random intercept models to capture spatial effects, meaning only the intercept could vary across levels, and all regression coefficients were treated as fixed. In the two-stage hurdle modeling, the first stage models if households generated any VMT (a dummy variable, modeled with logistic regression), while the second stage models the amount of VMT generated by households which generated any (a continuous variable, modeled with linear regression). This model has been used in previous studies with similar data structures (Ewing, et al., 2015; Tian & Ewing, 2017). The equations used are:

Stage 1 (multilevel logistic regression):

$$\text{Level 1:}\quad P(y = 1 \mid x_1, \ldots x_l) = 1/(1 + e^{-(\pi_0 + \sum_{i=1}^{l} \pi_i x_i)})$$

$$\text{Level 2:}\quad \pi_0 = \beta_{00} + \sum_{j=1}^{m} \beta_{0j} W_j + r_0$$

$$\pi_i = \beta_{i0}$$

Stage 2 (multilevel linear regression):

$$\text{Level 1: } E(y = 1 \mid x_1, \ldots x_n) = \pi_0 + \sum_{i=1}^{n} \pi_i x_i \text{ or } E(y = 1 \mid x_1, \ldots x_n) = e^{-(\pi_0 + \sum_{i=1}^{n} \pi_i x_i)}$$

$$\text{Level 2:}\quad \pi_0 = \beta_{00} + \sum_{j=1}^{m} \beta_{0j} W_j + r_0$$

$$\pi_i = \beta_{i0}$$

Where: *P* is the probability that the dependent variable equals 1 at Stage 1, *E* is the estimated value of the dependent variable at Stage 2;

π_0 is the intercept of the dependent variable at the level 1, π_i is the coefficient of independent variables at level 1, x_i is the independent variables at level 1;

β_{00} is the intercept at level 2, β_{0j} is the coefficient of independent variables at level 2, W_j is the independent variables at level 2;

and r_0 is the random error component for the deviation of the intercepts.

3.4 Machine learning modeling

Boosted regression trees (BRT) combine the advantages of regression trees and boosting algorithms. Regression trees divide a feature space into nonoverlapping rectangular regions (leaves) and fit a simple model (e.g., a constant) on each leaf. Boosting is an ensemble learning approach, which amplifies the performance of a simple model (base learner) by training and combining the predictions of several models. BRT uses regression trees as base learners, inheriting their advantages. Tree algorithms are nonparametric, do not require distributional assumptions on data, easily manage mixed data types (e.g., quantitative and qualitative inputs, missing input values), and can easily fit nonadditive behaviors (e.g., input variable interaction effects) and nonlinear effects.

BRT uses regression trees as base learners for both regression and classification problems and can be understood as an additive model in which terms (except the intercept term f_0) are regression trees fitted in a forward stagewise manner. For simplicity, the BRT regression model can be represented as:

$$f_B(x) = f_0 + \sum_{b=1}^{B} \sum_{m=1}^{Mb} \beta_{bm} I(x \in R_{bm})$$

Where: f_0 is the model intercept, a constant that minimizes the overall loss function value;

B is the total number of regression trees included in the BRT model ensemble and *b* is the index for each individual tree in the ensemble;

M_b is the tree size (i.e., the number of leaves for the tree);

β_{bm} is the fitted constant for disjoint region R_{bm}.

Although tree models can be visualized with easily interpretable tree diagrams, it is not easy to interpret a BRT ensemble consisting of hundreds of trees in the same fashion. Two tools were implemented to aid in interpretation of this "black box" algorithm.

The measure of relative variable influence is used to identify the input variables most influential in the BRT model. These are computed based on the number of times a variable is chosen for splitting, weighted by the improvement from such splitting, and averaged over all trees in the ensemble (Friedman & Meulman, 2003). The relative influence of each variable is scaled so that the sum adds to 100, with higher numbers indicating larger contributions.

Next, partial dependence plotting (PDP) is used to visualize marginal effects of key variables on the response, after accounting for the effects of all other variables in the model. PDPs serve as low-dimensional visual representations of a prediction function which allow easier interpretation of the relationships between input variables and outcomes across the distribution range of the input variable (Greenwell, 2017). In interpreting PDPs, the trend or shape of the plotted function indicates how the variable affects the response differently across its range, i.e., if the effect of the variable on the response is linear or nonlinear. Tick marks at the top of the diagram indicate the distribution of the predictor in deciles (e.g., 10^{th}, 20^{th}, … percentiles).

The "gbm" function in the "gbm" R package was used to generate the models. Within the realm of machine learning, there are many different types of tree models, including GBM, XGBoost, LightGBM, Random Forest, and others. GBM and XGBoost are both popular boosting algorithms that use weak learners (i.e., trees) to achieve better predictive performance. However, there are a few differences in their algorithm implementation. XGBoost enhances the algorithm by embedding regularization, performing parallelization, utilizing different tree pruning strategies, and the inclusion of in-built cross-validation techniques. GBM and XGBoost often achieve similar prediction performance, while the latter is usually computationally more efficient through parallel computation. LightGBM, like XGBoost, is another distributed high-performance framework that uses trees as base learners in a gradient boosting algorithm. LightGBM supports both parallel and GPU learning. The primary reason we use GBM in this study is that GBM has relatively fewer tuning parameters than LightGBM and XGBoost do, while achieving a similar performance. Furthermore, studies have shown that the tree-based ensemble learning algorithms are often less sensitive to the values of the tuning parameters as compared to other machine learning algorithms (Zhu, 2008). We have run the analysis for this study using both GBM and XGBoost. Based on multiple evaluation

metrics (AUC, Misclassification Rate, R-squared, and RMSE), GBM and XGBoost performed very similarly. As such, we chose GBM to demonstrate the comparison.

3.5 Measure of predicting performance

Both the statistical models and ML models were estimated as two-stage hurdle models. The prediction power of statistical models and ML models were evaluated in terms of their prediction accuracy by commonly used evaluation metrics.

Receiver operating characteristic (ROC) curves, the Areas Under ROC Curves (AUCs) and misclassification rates, all common measures of predictive power of models for binary responses, are appropriate for assessing the Stage 1 models. For ROC curves, the true-positive rate (i.e., sensitivity) is plotted on the y-axis and the false-positive rate (i.e., 1-specificity) is plotted on the x-axis. AUCs indicate the predictive accuracy of logistic models, with values ranging from 0.5 (no predictive power) to 1.0 (perfect prediction). Note that AUC score measures the quality of the model's prediction performance irrespective of what classification threshold is chosen. AUC is particularly useful for imbalanced classification situations. Following Manel et al., models with AUC values between 0.7 and 0.9 are considered "useful" and those with values greater than 0.9 are considered "high accuracy" (Manel et al., 2001). Misclassification rates tell the percentage of observations that were incorrectly predicted by the models. However, in an imbalanced dataset (as in this study, where 94% of *anyvmt* values equal 1), misclassification rate may not be a good measure for predicting performance.

R-squared and root mean square error (RMSE) values, common measures of predictive power of models for numeric responses, are appropriate for assessing the Stage 2 models. R-squared values represent the proportion of the variance for a dependent variable that is explained by the fitted regression model, with values from 0 to 1. The higher the R-squared, the better the model's predicting power. RMSE represents the average difference between model-predicted values and actual observed values. The lower the RMSE, the better the model's predicting performance.

K-fold cross-validation (k=5) was applied to choose the optimal models. The model estimation and validation were computed in R 4.3.1 with the *glmer* function (lme4 package) for multilevel modeling, *gbm* function (gbm package) for BRT, and the *roc* function for performance measures (AUC package).

4 Results

4.1 Traditional statistical models

Table 3 presents the results of the best fitting multilevel two-stage regression models for predicting whether households produced any VMT (dependent variable *anyvmt*), as well as for predicting the amount of VMT generated by households which generated any (dependent variable *lnvmt*). The likelihood of a household generating any VMT rises with increases in the socioeconomic variables of household size, number of household workers, and household income, which are all consistent with findings from previous studies. Larger households with more workers have greater needs for transportation to access employment and other destinations, and household income is an indicator of a household's ability to afford vehicles, fuel, and other costs associated with the production of VMT.

Table 3. Multilevel two-stage regressions of household VMT generation

Categories	Variables	Stage 1: Any household VMT (*anyvmt*)			Stage 2: Household VMT (*lnvmt*)		
		coeff.	*Std. Err.*	*p*-value	coeff.	*Std. Err.*	*p*-value
constant		3.581	0.114	< 0.001	3.181	0.033	< 0.001
Socioeconomics	household size	0.222	0.021	< 0.001	0.133	0.005	< 0.001
	children	—	—	—	0.042	0.005	< 0.001
	household workers	0.366	0.021	< 0.001	0.200	0.004	< 0.001
	household income	0.746	0.024	< 0.001	0.127	0.004	< 0.001
Built environment	activity density	-0.182	0.014	< 0.001	-0.040	0.005	< 0.001
	job-pop balance	—	—	—	—	—	—
	Land-use entropy	-0.200	0.021	< 0.001	-0.066	0.004	< 0.001
	intersection density	-0.142	0.023	< 0.001	-0.090	0.006	< 0.001
	4-way intersection	-0.197	0.019	< 0.001	-0.048	0.004	< 0.001
	transit stop density	-0.093	0.013	< 0.001	-0.014	0.004	< 0.001
	rail station	-0.024	0.010	0.016	—	—	—
	destinations in 30 minutes by auto	—	—	—	-0.085	0.007	< 0.001
	destinations in 30 minutes by transit	-0.474	0.032	< 0.001	-0.079	0.006	< 0.001
Region	regional population	—	—	—	—	—	—
	gas price	-0.479	0.126	< 0.001	—	—	—
	low temperature	0.250	0.120	0.038	—	—	—
	high temperature	-0.195	0.102	0.056	—	—	—
	precipitation	—	—	—	—	—	—
AIC		32254.5					
BIC		32395.6					
Log-Likelihood		-16112.3					
Pseudo-R2:					0.18		
"—" This variable is not statistically significant at 0.1 level							

The likelihood of households generating any VMT (dependent variable *anyvmt*) decreases with increases in activity density, land-use entropy, 4-way intersection density, transit stop density, presence of a nearby rail station, and destination accessibility within 30 minutes by transit, consistent with previous literature. Denser, more diverse, and better-connected areas with increased transit accessibility provide more opportunities for people to access employment and other destinations via modes other than private vehicles, such as transit and active transportation. Unsurprisingly, the likelihood of any VMT production also decreases with regional gas prices, as higher gas prices represent an economic barrier or disincentive to vehicular travel.

The amount of VMT generated by households which generated any (dependent variable *lnvmt*) also increases with increases in the socio-economic variables of household size, number of workers, and household income, as well as with the number of children in the household. The amount of VMT generated by these households decreases with increases in activity density, land-use entropy, intersection density, 4-way intersection density, and destination accessibility within 30 minutes by both auto and transit. As with the results of the first stage hurdle model, these findings are consistent with previous research and indicate that denser, more diverse, better-connected, and more accessible regions reduce vehicular travel by those who reside within them.

4.2 Machine learning approaches

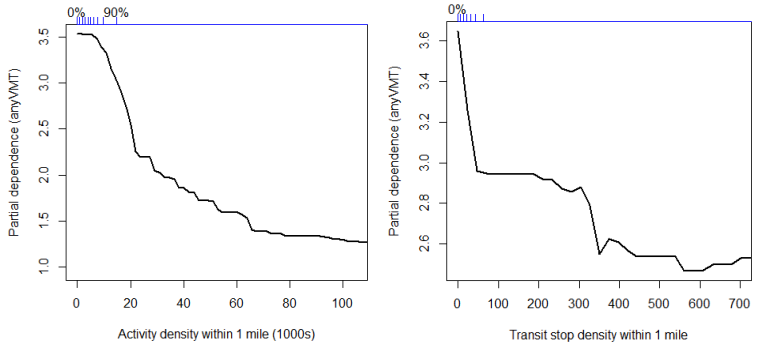
The meta-parameters of BRT models include shrinkage value (or learning rate) and the number of trees in the ensemble. Choosing the appropriate values for the meta-parameters is crucial for the model performance. For example, too many (few) trees in

the ensemble can cause overfitting (underfitting). In this study, a two-dimensional grid search was conducted to find the optimal BRT models which minimize the cross-validation errors. Table 4 presents the results of the best fitting multilevel two-stage BRT models for predicting the likelihood of a household generating any VMT (dependent variable *anyvmt*), as well as for predicting the amount of VMT generated by those households generating any (dependent variable *lnvmt*). The final models were based on shrinkage = 0.05 and number of trees = 5000.

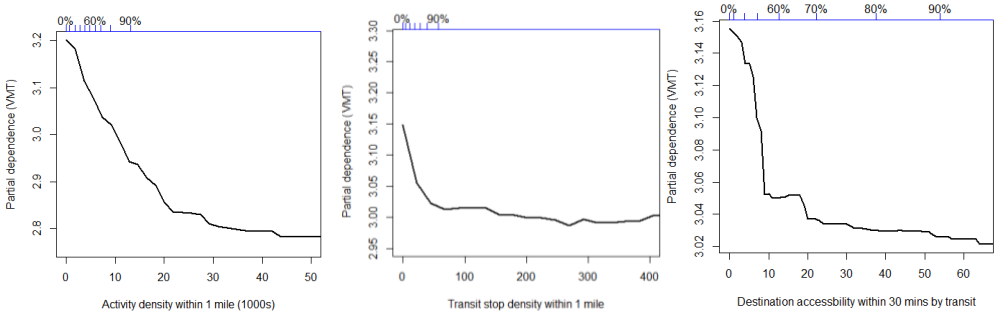
Table 4. Multilevel two stage BRTs of household VMT generation

Categories	Variables	Stage 1: Any household VMT (*anyvmt*)			Stage 2: Household VMT (*lnvmt*)		
		Rank	Relative Importance (%)	Total (%)	Rank	Relative Importance (%)	Total (%)
Socioeconomics	household size	6	4.96	26.00	1	28.31	53.65
	children	18	0.07		14	1.32	
	household workers	9	1.76		2	13.30	
	household income	2	19.21		4	10.72	
Built environment	activity density	1	36.62	71.89	3	11.93	38.08
	job-pop balance	12	0.56		12	1.46	
	Land-use entropy	7	3.96		9	3.15	
	intersection density	4	7.72		8	3.58	
	4-way intersection	8	2.27		13	1.42	
	transit stop density	3	13.69		5	8.41	
	rail station	5	5.98		18	0.41	
	destinations in 30 minutes by auto	14	0.43		10	2.14	
	destinations in 30 minutes by transit	11	0.66		6	5.58	
Region	regional population	15	0.38	2.09	15	0.91	8.25
	gas price	13	0.48		11	1.51	
	low temperature	10	0.69		7	4.44	
	high temperature	16	0.31		16	0.84	
	precipitation	17	0.23		17	0.55	

Unlike regression models, BRT models can determine the relative importance of various independent variables (the percentage of the model's predictive capability which can be attributed to each variable) and rank them, as well as identify non-linear or threshold effects of those variables (ranges at which they are effective in influencing the outcome). In the first stage model (dependent variable *anyvmt*), the combined effects of all built environment variables outweigh the effects of socioeconomic factors, contributing 71.89% of the model's predictive capability, compared to a 26% contribution from socioeconomics and a 2.09% contribution from regional factors. Activity density was the largest single contributor, accounting for 36.62% of the model's predictive capability, followed by household income (19.21%), transit stop density (13.69%), and intersection density (7.72%). In the second-stage model, for determining the amount of VMT generated by households generating any VMT (dependent variable *lnvmt*), the combined influence of socioeconomics is greater than that of the built environment, contributing 53.65% of the model's predictive capability, compared with a 38.08% contribution from the built environmental variables and 8.25% from regional variables. The top-ranking contributions come from household size (28.31%), number of household workers (13.3%), activity density (11.93%), and household income (10.72%), followed by transit stop density (8.41%) and destination accessibility by transit (5.58%).



(a) Any household VMT (*anyvmt*)



(b) Household VMT for households with any VMT (*lnvmt*)

Figure 1. The nonlinear relationship between built environmental variables and household VMT generation

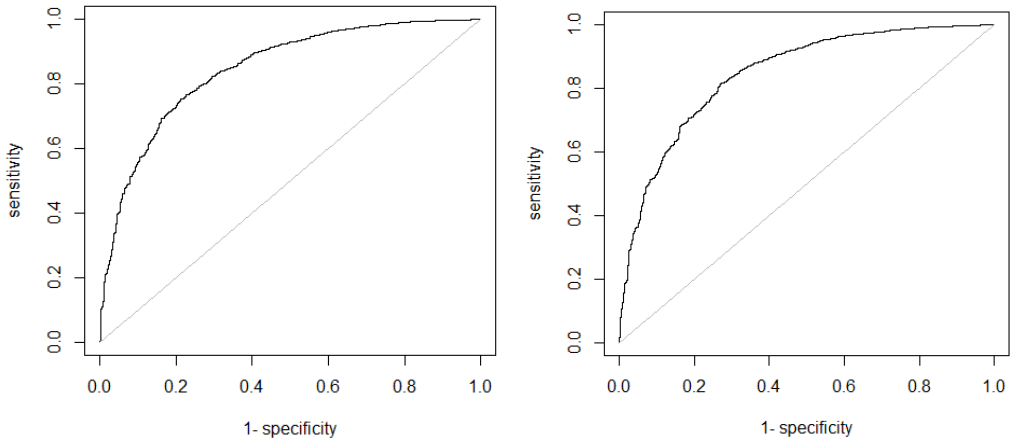
Figure 1a shows PDPs for the top two built environment variables in the first stage model (dependent variable *anyvmt*) after controlling for the effects of the other variables in the model. Both demonstrate non-linear effects on the probability that a household generates any VMT. For activity density, there is a very strong negative association on the probability of generating any VMT across the range from ~0-20,000 (jobs plus population per square mile), a less steep but still significant negative association from ~20,000-65,000, and a diminishing effect above that threshold. The effect of transit stop density exhibits two notable thresholds, with very steep decreases in the probability of any VMT production occurring at ~50 stops within a mile and again at ~300-350 stops.

Figure 1b shows PDPs for the top built environment variables in the second stage model (dependent variable *lnvmt*). Once again, the independent variables show non-linear associations with the amount of VMT produced by these households. The impact of activity density again demonstrates a threshold around 20,000, with a much steeper negative impact in the range below that threshold and a more gradual negative association above. Transit stop density also displays a similar threshold value, ~50 stops, to the Stage 1 model, with a steep negative impact on the amount of VMT produced up to that value and a minimal association above it. Destination accessibility by transit exhibits two thresholds, with a very steep decline in the amount of VMT produced up to a value of ~10% (of regional employment accessible within 30 minutes by transit), little impact between ~10-20%, and a sharp decline around the 20% value, after which the impact diminishes.

4.3 Comparison of predicting performances

To evaluate the predictive performances of both statistical models and ML models, a k-fold cross-validation (k=5) process was applied (Tian et al., 2020). First, the data was divided into five partitions. Models were estimated with one partition (training dataset) while the rest were used to validate the modes (testing dataset). This process was repeated five times until the models were stabilized.

Figure 2 and Table 5 present the measures of the predicting performances of both models. For the first stage models predicting the likelihood of a household generated any VMT (dependent variable *anyvmt*), both the statistical model and ML model have a value of 0.85 for AUC, which is considered a "useful application" and close to "high accuracy". Misclassification rates are also presented for the stage 1 model, but do not indicate much due to the imbalanced distribution between the two categories (0, 1) of the dependent variable (*anyvmt*).



(a) Traditional statistical model, AUC=0.85. (b) Machine learning model, AUC=0.85

Figure 2. Receiver operating characteristic (ROC) curves and the area under the ROC (AUC) statistics for model prediction at Stage 1

For the second stage model predicting the amount of VMT generated by households which generated any (dependent variable *lnvmt*), the ML model has a slightly higher R-squared and lower RMSE than the statistical model. The difference in the values is not significant however, and the results indicate the prediction performance of the BRT model is again similar to that of the statistical model. These results indicate a cross-validation between the two approaches in terms of their ability to predict VMT outcomes from socioeconomic and built environment variables.

Table 5. Measures of predicting performance

	Stage 1 (*anyvmt*)		Stage 2 (*lnvmt*)	
	AUC	Misclassification Rate	R-squared	RMSE
Traditional statistic model (multilevel models)	0.85	0.058	0.20	0.94
Machine learning model (BRT)	0.85	0.057	0.22	0.92

5 Discussion

The regression model results reaffirm previous research that increases in socioeconomic variables (such as household size, number of workers, and income) lead to increases in VMT generation, while change in the built environment in the direction of compact development (i.e., increases in activity density, land-use entropy, intersection density, transit stop density, and destination accessibility) leads to decreases in VMT. The BRT model allows for closer examination of how these variables impact VMT generation as well as their relative levels of influence on the outcome, a question which has long been debated in the planning scholarship. The model indicated that the impacts of land use may outweigh those of socioeconomics in determining whether households generate any VMT, while socioeconomic factors may play a moderately larger role in determining how much VMT is generated by those households generating any. Crucially, the BRT model allows for the identification of effective ranges and thresholds at which built environment variables have the strongest effects on VMT outcomes. The need for models which can uncover these so-called nonlinear effects has long been indicated by planners and statisticians who note that real-world mechanisms do not follow linear patterns (Breiman, 2001; Galster, 2018). Even traditionally modeled studies have indicated confounding effects of the built environment on VMT, such as the possibility of density exhibiting positive or negative effects at different ranges (Cervero & Murakami, 2010). One notable result of the present analysis is that while many previous studies have found ML methods to demonstrate better model fit than traditional statistical methods, this study demonstrated a similarly high level of predictive accuracy between the methods. Due to the lack of studies specifically examining VMT generation with multiple models, this calls for further investigation; it also reinforces previous indications that ML methods may best be used in conjunction with rather than simply as an alternative to more traditional methodologies (Barri et al., 2022). As machine learning models are still relatively recent in terms of application to planning practice and have been understudied in terms of their comparative accuracy in predicting VMT generation in particular, the use of these simple, efficient, and inherently interpretable models as a means of validating associations of impacts of independent variables on travel outcomes derived from ML models continues to be of value for planners.

The high level of predictive accuracy of both models in the study provides strong validation in terms of indicating the significant impact of land-use patterns on VMT generation. Both models indicate that increasing activity density, transit stop density, and destination accessibility should be effective land-use techniques for reducing VMT generation. Land-use diversity and street design are influential in both models as well, particularly in the first stage analysis, although use diversity impacts appear to be slightly less significant compared to other factors in the ML model. While the non-linear effects of these variables observed in the PDPs indicate important thresholds, they do not contradict the direction of association observed in the statistical model. In other words, both models indicate that compact, diverse, and accessible land-use patterns are

associated with decreased production of vehicle trips and VMT. The BRT model indicates that many of these effects begin to dimmish above certain thresholds (such as activity densities above 20,000 and transit stop densities above 50 stops within a mile), indicating that planners should focus efforts on areas which are below these thresholds to obtain more significant VMT reductions. By estimating VMT using two distinct methodologies, this study cross-validates the findings regarding the impact of built environmental variables on travel. At the same time, this study shows that the ability to determine effective ranges and thresholds is an essential benefit of the BRT methodology, whereas the simplicity, ease of use and interpretability of traditional models also has benefits.

Considering the massive contributions personal vehicle usage has on overall greenhouse gas emissions, and the wide-ranging negative effects of those emissions on society, planning for ways to reduce VMT has become an increasingly urgent consideration. Increasing VMT has contributed to issues ranging from traffic congestion, low air quality, pollution, reduced physical activity and personal health, global climate change, and social inequalities. By demonstrating the efficacy of machine learning models in predicting VMT generation, this study gives planners, engineers, and policymakers essential tools to better understand how built environments contribute to VMT generation and how to plan communities in ways that will lead to VMT reduction. In addition to demonstrating predictive capabilities with an equally high accuracy to traditional models, these models enable us to determine the effective ranges and thresholds at which various built environment parameters lead to the minimal possible generation of VMT, allowing for more cost-effective planning goals and applications. This information is crucial as the negative externalities of increased automotive use continue to affect personal, community, and global health outcomes.

As the range of dataset in this study spins over 11 years, we do acknowledge that there were events that could have affected overall VMT generation, such as the 2007–2008 financial crisis. However, we do not think it should significantly skew the findings of this research. Since the study is not examining VMT generation over time (longitudinal trends) but rather impacts on VMT generation at a given place and time, there is no indication that a potential drop in VMT generation during that period due to decreases in household income or other factors would have a significant impact on which variables impacted VMT generation. An analysis of the shifts in amount of VMT generated during that period would be suited to a study with a longitudinal dataset, which we do hope to develop as we update and expand this dataset in the future with future iterations of MPOs' household travel surveys.

6 Conclusion

The regression model results reaffirm previous research that increases in socioeconomic variables (such as household size, number of workers, and income) lead to increases in VMT generation, while change in the built environment in the direction of compact development (i.e., increases in activity density, land-use entropy, intersection density, transit stop density, and destination accessibility) leads to decreases in VMT. The BRT model allows for closer examination of how these variables impact VMT generation as well as their relative levels of influence on the outcome, a question which has long been debated in the planning scholarship. The model indicated that the impacts of land use may outweigh those of socioeconomics in determining whether households generate any VMT, while socioeconomic factors may play a moderately larger role in determining how much VMT is generated by those households generating any. Crucially, the BRT model allows for the identification of effective ranges and thresholds at which

built environment variables have the strongest effects on VMT outcomes. The need for models which can uncover these so-called nonlinear effects has long been indicated by planners and statisticians who note that real-world mechanisms do not follow linear patterns (Breiman, 2001; Galster, 2018). Even traditionally modeled studies have indicated confounding effects of the built environment on VMT, such as the possibility of density exhibiting positive or negative effects at different ranges (Cervero & Murakami, 2010). One notable result of the present analysis is that while many previous studies have found ML methods to demonstrate better model fit than traditional statistical methods, this study demonstrated a similarly high level of predictive accuracy between the methods. Due to the lack of studies specifically examining VMT generation with multiple models, this calls for further investigation; it also reinforces previous indications that ML methods may best be used in conjunction with rather than simply as an alternative to more traditional methodologies (Barri et al., 2022). As machine learning models are still relatively recent in terms of application to planning practice and have been understudied in terms of their comparative accuracy in predicting VMT generation in particular, the use of these simple, efficient, and inherently interpretable models as a means of validating associations of impacts of independent variables on travel outcomes derived from ML models continues to be of value for planners.

The high level of predictive accuracy of both models in the study provides strong validation in terms of indicating the significant impact of land-use patterns on VMT generation. Both models indicate that increasing activity density, transit stop density, and destination accessibility should be effective land-use techniques for reducing VMT generation. Land-use diversity and street design are influential in both models as well, particularly in the first stage analysis, although use diversity impacts appear to be slightly less significant compared to other factors in the ML model. While the non-linear effects of these variables observed in the PDPs indicate important thresholds, they do not contradict the direction of association observed in the statistical model. In other words, both models indicate that compact, diverse, and accessible land-use patterns are associated with decreased production of vehicle trips and VMT. The BRT model indicates that many of these effects begin to dimmish above certain thresholds (such as activity densities above 20,000 and transit stop densities above 50 stops within a mile), indicating that planners should focus efforts on areas which are below these thresholds to obtain more significant VMT reductions. By estimating VMT using two distinct methodologies, this study cross-validates the findings regarding the impact of built environmental variables on travel. At the same time, this study shows that the ability to determine effective ranges and thresholds is an essential benefit of the BRT methodology, whereas the simplicity, ease of use and interpretability of traditional models also has benefits.

Considering the massive contributions personal vehicle usage has on overall greenhouse gas emissions, and the wide-ranging negative effects of those emissions on society, planning for ways to reduce VMT has become an increasingly urgent consideration. Increasing VMT has contributed to issues ranging from traffic congestion, low air quality, pollution, reduced physical activity and personal health, global climate change, and social inequalities. By demonstrating the efficacy of machine learning models in predicting VMT generation, this study gives planners, engineers, and policymakers essential tools to better understand how built environments contribute to VMT generation and how to plan communities in ways that will lead to VMT reduction. In addition to demonstrating predictive capabilities with an equally high accuracy to traditional models, these models enable us to determine the effective ranges and thresholds at which various built environment parameters lead to the minimal possible generation of VMT, allowing for more cost-effective planning goals and applications.

This information is crucial as the negative externalities of increased automotive use continue to affect personal, community, and global health outcomes.

As the range of dataset in this study spins over 11 years, we do acknowledge that there were events that could have affected overall VMT generation, such as the 2007–2008 financial crisis. However, we do not think it should significantly skew the findings of this research. Since the study is not examining VMT generation over time (longitudinal trends) but rather impacts on VMT generation at a given place and time, there is no indication that a potential drop in VMT generation during that period due to decreases in household income or other factors would have a significant impact on which variables impacted VMT generation. An analysis of the shifts in amount of VMT generated during that period would be suited to a study with a longitudinal dataset, which we do hope to develop as we update and expand this dataset in the future with future iterations of MPOs' household travel surveys.

Acknowledgments

This work was financially supported by the Louisiana Transportation Research Center [# 23-4TIRE, 2022-2023].

Author contribution

G. Tian: Conceptualization; G. Tian: Data acquisition; G. Tian & B. Li: Methodology; G. Tian & B. Li: Formal analysis and investigation; G. Tian, B. Danton & B. Li: Writing—original draft preparation. G. Tian, B. Danton, B. Li, V. Gopu & J. Codjoe: Writing—review and editing.

References

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.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, & design. *Transportation Research Part D: Transport and Environment, 2*(3), 199–219.

Cervero, R., & Murakami, J. (2010). Effects of built environments on vehicle miles traveled: Evidence from 370 US urbanized areas. *Environment and Planning A: Economy and Space, 42*, 400–418.

Ding, C., Cao, X., & Liu, C. (2019). How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. *Journal of Transport Geography, 77*, 70–78.

Ding, C., Cao, X., & 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.

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. *Accident Analysis and Prevention, 112*, 116–126.

Du, Q., Zhou, Y., Huang, Y., Wang, Y., & Bai, L. (2022). Spatiotemporal exploration of the non-linear impacts of accessibility on metro ridership. *Journal of Transport Geography, 102*, 103380.

Ewing, R., & Cervero, R. (2001). Travel and the built environment: A synthesis. *Transportation Research Record, 1780*, 87–114.

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

Ewing, R., Hamidi, S., Gallivan, F., Nelson, A. C., & Grace, J. B. (2014). Structural equation models of VMT growth in US urbanized areas. *Urban Studies, 51*(14), 3079–3096.

Ewing, R., Tian, G., Goates, J., Zhang, M., Greenwald, M. J., Joyce, A., . . . Greene, W. (2015). Varying influences of the built environment on household travel in 15 diverse regions of the United States. *Urban Studies, 52*(13), 2330–2348.

Federal Highway Administration. (2021, May). *FHWA forecasts of vehicle miles traveled (VMT): Spring 2021*. Retrieved from https://www.fhwa.dot.gov/policyinformation/tables/vmt/2021_vmt_forecast_sum.pdf

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine, 22*, 1365–1381.

Galster, G. C. (2018). Nonlinear and threshold effects related to neighborhood: Implications for planning and policy. *Journal of Planning Literature, 33*(4), 492–508.

Gan, Z., Yang, M., Feng, T., & Timmermans, H. J. (2020). Examining the relationship between built environment and metro ridership at station-to-station level. *Transportation Research Part D: Transport and Environment, 82*, 102332.

Greene, W. (2018). *Econometric analysis (8th Edition)*. London: Pearson.

Greenwell, B. M. (2017, June). pdp: An R package for constructing partial dependence plots. *The R Journal, 9*(1), 421–436.

Hu, X., Cao, Y., Peng, T., Gao, R., & Dai, G. (2021). Nonlinear influence model of built environment of residential area on electric vehicle miles traveled. *World Electric Vehicle Journal, 12*, 247.

Ihlanfeldt, K. (2020). Vehicle miles traveled and the built environment: New evidence from panel data. *Journal of Transport and Land Use, 13*(1), 23–48.

Lee, A. E., & Handy, S. L. (2018). Leaving level-of-service behind: The implications of a shift to VMT impact metrics. *Research in Transportation Business & Management, 29*, 14–25.

Li, W., & Kockelman, K. M. (2022). How does machine learning compare to conventional econometrics for transport data sets? A test of ML versus MLE. *Growth & Change, 53*(1), 342–376.

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems, 96*, 101845.

Liu, J., Wang, B., & Xiao, L. (2021). Non-linear associations between built environment and active travel for working and shopping: An extreme gradient boosting approach. *Journal of Transport Geography, 92*, 103034.

Manel, S., Williams, H. C., & Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: The need to account for prevalence. *Journal of Applied Ecology, 38*(5), 921–931.

Nasri, A., & Zhang, L. (2012). Impact of metropolitan-level built environment on travel behavior. *Transportation Research Record, 2323*, 75–79.

Nasri, A., & Zhang, L. (2014). Assessing the impact of metropolitan-level, county-level, and local-level built environment on travel behavior: Evidence from 19 U.S. urban areas. *Journal of Urban Planning and Development, 141*(3), 04014031.

Rui, A., Tong, Z., Ding, Y., Tan, B., Wu, Z., Xiong, Q., & Liu, Y. (2022). Examining non-linear built environment effects on injurious traffic collisions: A gradient boosting decision tree analysis. *Journal of Transport & Health, 24*(5), 101296.

Salon, D., Boarnet, M. G., Handy, S., Spears, S., & Tal, G. (2012). How do local actions affect VMT? A critical review of the empirical evidence. *Transportation Research Part D: Transport and Environment, 17*(7), 495–508.

Stevens, M. R. (2017). Does compact development make people drive less? *Journal of the American Planning Association, 83*(1), 7–18.

Tao, T., & Næss, P. (2022). Exploring nonlinear built environment effects on driving with a mixed-methods approach. *Transportation Research Part D: Transport and Environment, 111*, 103443.

Tao, T., Wu, X., Cao, J., Fan, Y., Das, K., & Ramaswami, A. (2020). 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., & Ewing, R. (2017). A walk trip generation model for Portland, OR. *Transportation Research Part D: Transport and Environment, 52*, 340–353.

Tian, G., Park, K., Ewing, R., Watten, M., & Walters, J. (2020). Traffic generated by mixed-use developments—A follow-up 31-region study. *Transportation Research Part D: Transport and Environment*, 78 (2020), 102205.

U.S. Environmental Protection Agency. (2022, May 19). *Carbon pollution from transportation*. Retrieved from https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation

Wali, B., Frank, L. D., Chapman, J. E., & Fox, E. H. (2021). Developing policy thresholds for objectively measured environmental features to support active travel. *Transportation Research Part D: Transport and Environment, 90*, 102678.

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.

Xie, C., Lu, J., & Parkany, E. (2003). Work travel mode choice modeling with data mining: Decision trees and neural networks. *Transportation Research Record, 1854*(1), 50–61.

Xu, Y., Yan, X., Liu, X., & Zhao, X. (2021). Identifying key factors associated with ridesplitting adoption rate. *Transportation Research Part A: Policy and Practice*, *144*, 170–188.

Yan, X., Liu, X., & Zhao, X. (2020). Using machine learning for direct demand modeling of ridesourcing services. *Journal of Transport Geography*, *83*, 102661.

Zhang, L., Hong, J., Nasri, A., & Shen, Q. (2012). How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in US cities. *Journal of Transport and Land Use, 5*(3), 40–52.

Zhang, M., & Zhang, W. (2020). When context meets self-selection: The built environment–Travel connection revisited. *Journal of Planning Education and Research, 40*(3), 304–319.

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behavior and Society, 20*, 22–35.

Zhu, M. (2008). Kernels and ensembles: Perspectives on statistical learning. *The American Statistician, 62*(2), 97–109.