Hapres | Journal of Sustainability Research

sustainability.hapres.com

Article

Investigating the Application of a Transportation Energy Consumption Prediction Model for Urban Planning Scenarios in Machine Learning and Shapley Additive Explanations Method

Shideh Shams Amiri, Maya Mueller, Simi Hoque *

Drexel University, Philadelphia, PA 19104, USA

* Correspondence: Simi Hoque, Email: sth55@drexel.edu.

ABSTRACT

Accurate forecasts of future energy usage are an important step towards reaching carbon mitigation commitments for city policymakers. Beyond identifying sources of emission intensity for a region, the forecast mechanism must be capable of compensating for gaps in available data and of accounting for the uncertainties behind the dynamics of an urban system. By considering a range of possible scenarios, the prediction model can identify recurring sources of high energy consumption and fine-tune areas of priority with incoming data. This paper considers the impact of predicted shifts in demographic and economic trends for the region on transportation energy consumption. The transportation energy use model is formulated from the Delaware Valley Regional Planning Commission (DVRPC) open-source Household Travel Survey (HTS). Based on these data inputs, a Machine Learning (ML) algorithm is implemented in the form of an Extreme Gradient Boosting (XGBoost) model to estimate energy consumption with a corresponding SHapley Additive exPlanations (SHAP) analysis of feature contribution. From this, a synthetic population is produced using the ML outputs and marginal sums with data from the Census Bureau's American Community Survey (ACS) to estimate energy consumption for the region. The results indicate that shifting dominant travel modes and income distribution in accordance with the Enduring Urbanism forecast projections led to a decrease in household transportation energy use. Moreover, additional analysis of the model output demonstrates that changes in energy use depend strongly on geographic area and income group.

KEYWORDS: energy consumption; machine learning; scenario planning; bottom-up modeling; household transportation energy model

Open Access

Received: 09 November 2021

Accepted: 16 February 2022

Published: 03 March 2022

J Sustain Res. 2022;4(1):e220001. https://doi.org/10.20900/jsr20220001

INTRODUCTION

In 2018, the U.S. emitted 6677 million metric tons of CO_2 [1]. City planners and policy makers are urgently seeking a new sustainable city planning paradigm. Given the uncertainties and temporal dynamics of climate change, urban policy makers must identify projects and target areas that minimize energy consumption efficiently under the constraints of a finite budget. It is imperative to consider a range of potential future forecasts and identify the corresponding variables responsible for energy consumption. In this way, the forecasts aid in the reaching of carbon emission benchmarks and in the creation of mitigation policies that integrate knowledge of how the city system will evolve over time.

When predicting energy consumption, most studies utilize either top-down or bottom-up approaches. The top-down methods draw on historical aggregate data sets of energy consumption and top-level indicators of energy use, such as macroeconomic and/or demographic characteristics of the region [2–7]. Most often, the intention behind top-down models is to estimate the impact of various large-scale features on energy consumption at the regional level, but this approach is not as effective for analyzing the sources of energy use at a more disaggregate scale. A bottom-up approach takes sample data from smaller geographic units in the region to develop estimates and can be applied to analyze the spatial variation of energy use [8–10].

Although bottom-up approaches can discern micro-level patterns in energy use, the acquisition of reliable data at the disaggregate scale can be challenging for larger study areas. As top-down and bottom-up models perform at different scales, variation in results between each approach is often inevitable [11]. In Zhang et al. [12], the authors propose a solution to this dilemma by applying an Elastic Net regression model and statistically matched Residential Energy Consumption Survey (RECS) and Public Use Microdata Sample (PUMS) data to develop a synthetic population for the Atlanta Metropolitan region. This technique successfully produced residential energy estimations consistent with top-down estimations for the study area.

With respect to a transportation energy model, Amiri et al. [13] evaluates the performance of various modeling techniques for predicting household transportation energy consumption with Mean Absolute Errors (MAE), Mean Absolute Percentage Errors (MAPE), and Average R^2 scores. Machine learning methods were found to outperform Elastic Net regularization for the transportation energy use model.

Reiter and Marique [14] utilize both top-down and bottom-up approaches to develop forecasts of energy consumption in their simulation of urban renewal for Liège, Belgium. Demographic forecasts are drawn from recent trends in local and national datasets in a top-down approach, whereas energy usage data was developed through transport and building energy models in a bottom-up approach. The authors then constructed six plausible scenarios relating to changes in regional energy policies for existing residential building stock and the increase of school to work travel of residents. The results of the Liège study indicate that housing renovation would be essential in order to reach the city's carbon emission benchmarks, and moreover demonstrate how available energy use prediction models can be integrated into scenario planning contexts.

Previous studies on building or transportation sector energy consumption models tend to focus on model accuracy while neglecting model interpretability. Nevertheless, understanding the decisions behind a model's predictions is a critical factor in the model's practical applicability, as the validity of the prediction cannot be justified by solely relying on model accuracy.

Explainable Artificial Intelligence (XAI) offers the tools and methods to reach interpretable machine learning predictions without sacrificing model complexity. At the present, there are very few published studies that address the application of XAI in the transportation sector.

The current research has two objectives: to apply XAI on a transportation energy model with a focus on the interpretation of local and global features and to demonstrate how existing bottom-up approaches to predicting energy consumption can augment scenario planning forecasts. With this, urban planners and policy makers can trust the mechanisms underlying the predictive model and utilize these insights to develop targeted policies to mitigate transportation energy consumption in the region. This study develops a ML prediction model on household transportation energy [13] and produces a synthetic dataset of energy estimations for the study area [12,15]. The base case and scenarios are constructed from available projections of demographic and socioeconomic data for Philadelphia County. The model inputs are altered according to a selection of potential future trends, resulting in corresponding synthetic datasets for each scenario. The study then analyzes how the forecasted shifts in household-level behavior alter the model output and identifies potential geographic areas of interest based on variation in transportation energy use estimates across the study area.

METHODOLOGY

Data and Data Processing

The 2012 open-source Household Travel Survey (HTS) is used to develop a household transportation energy consumption prediction model. The Delaware Valley Regional Planning Commission (DVRPC) conducts household travel surveys for the Delaware Valley region. The 2012 HTS contains data on the daily travel behavior of residents in the Pennsylvania counties of Bucks, Chester, Delaware, Montgomery, and Philadelphia. The HTS also includes New Jersey Delaware Valley counties of Burlington, Camden, Gloucester, and Mercer. The dataset contains demographic and travel information for 5677 households, 13,830 residents, 10,570 vehicles, and 48,646 trips across 10 counties. For every household member, information on travel mode, trip purpose, destination and time of travel was recorded, along with demographics such as age, gender, vehicle availability, and employment status.

In addition to the HTS, the DVRPC's shapefiles of zonal boundaries and zonal data such as employment density and the number of households, schools, and bus stops are applied as model inputs to evaluate neighborhood characteristics. In all, 31 continuous and 12 categorical variables were selected to be included in the analysis. Descriptive statistics for these variables are available from the Supplementary Materials in Tables S1, S2, and S3 for demographics, trip information, and zone characteristics, respectively. In order to reduce collinearities across variables related to energy consumption, a multi-step screening process was performed. Also, variables that are directly used in transportation calculations such as HH travel distance were removed. The data and data processing methods are described comprehensively in.

Household transportation energy (HTE) consumption is the output of the prediction model, and depends on household trip generation, travel mode, fuel type, and trip distance. The HTE is calculated using the method described in Jiang [16] and is based on the DVRPC's reported daily travel patterns.

Equations (1–3) show how transportation energy consumption is measured at the household level. Fuel economy, fuel energy, and energy intensity factors are obtained from the Department of Transportation tables [17].

$$E_i^T = \sum_{\mathrm{m}} \mathrm{E}_{\mathrm{i}}^{\mathrm{m}} \tag{1}$$

$$E_i^m = \sum_j \sum_k (TF_j^m\, i.j.k * \frac{TD^{m_{i.j.k}}}{TO^{m_{i.j.k}}}) * EI^m \tag{2}$$

$$\boldsymbol{EI^m = FU^m * EC^m} \tag{3}$$

i—*i*th Household

j—*j*th Person in the household

k—Purpose

m—Mode

E_i^T—Total household daily transport energy consumption (kWh/HH)

$TF^{m}{}_{i.j.k}$—Trip frequency for person *j* in household *i* for purpose k with mode m (Trips/Day)

$TD^{m}{}_{i.j.k}$—Average trip distance for person *j* in household *i* for purpose k with mode m (Mile/Trip)

$TO^{m}_{i,j,k}$—Trip occupancy for person *j* in household i for purpose k with mode m
EI^{m}—Energy intensity factor for mode m (kWh/mile)
FU^{m}—Fuel economy factor for mode m (L/mile or KWH/mile)
EC^{m}—Energy content factor for mode m (kWh/L)

Transportation Energy Model Development

The model is trained using the explanatory features of categorical and numerical variables from households, trips, and neighborhood characteristics. The target feature of the model is transportation energy consumption by kilowatt-hour (kWh). K-fold cross-validation was used for training the dataset to find the best model parameters and to avoid over-fitting. Grid search algorithms were used for hyper-parameter (learning rates, weights, and thresholds) optimization for the model by searching through subsets of the hyper-parameter space to find the sequence to the lowest cross-validation error.

The main challenge in applying ML algorithms is in finding a method that achieves both high accuracy and interpretability. In a previous study by Amiri et al. [13], the performance of various ML models in predicting household transportation energy consumption was evaluated. Tree-based ensemble models could compute the global explanation to identify influential prediction features; however, the insights are still limited and are not useful in assessing the validity of individual predictions. Gradient boosting decision tree is a popular ML algorithm and has effective applications such as Extreme Gradient Boosting (XGBoost). For this study, the household transportation energy consumption prediction is reformulated as a XGBoost problem along with SHapley Additive exPlanations (SHAP) analysis to gain a better understanding of the model and to assess the validity of individual predictions. The XGBoost output is then integrated with the Census Bureau’s American Community Survey (ACS) marginal sums to produce a synthetic dataset of households.

Transportation Energy Model Interpretability

Training time, accuracy, and interpretability are vital to the energy prediction model. However, interpretability tends to be neglected in many studies, particularly with popular ML models (e.g., XGBoost and Neural network) inherent in “black-box” systems. XGBoost is a commonly used ML model as the algorithm performs with high accuracy, but the decision boundary in a XGBoost problem can be difficult to find if there are many features in the model. Without a clear decision boundary, the model’s rationale for determining one case to be stable or unstable in a power system becomes nearly impossible to interpret. An alternative solution lies in drawing out explanations of individual predictions, which can be achieved with XAI approaches like SHAP.

SHAP is a special case of the Shapley value introduced by Lundberg and Lee [18] and is implemented in the current paper to understand the insight of the ML prediction model. Classic Shapley value estimation derives from coalitional game theory such that a single prediction from the ML model is broken down into the component contributions of each model input, or feature value. The Shapley value for a given feature value represents the marginal contribution of the prediction instance in consideration of all possible coalitions of instances. In this way, the summed feature contributions for the prediction are equivalent to the difference between the predicted value for a given instance and the average predicted value for the dataset. For more detail on Shapley value estimation methods refer to Molnar [19].

SHAP provides a measure of additive feature importance in the form of a linear function of binary variables. Additive feature attribution methods begin with the original prediction model, or the transportation energy model developed by XGBoost (see the previous section) and develops an explanation model of simplified inputs based on the original. Unlike other XAI methods like LIME, SHAP satisfies all desirable properties for additive feature attribution methods [18]. Additionally, Shapley value estimation methods like SHAP guarantee the fair distribution of feature contribution for a given prediction instance.

For the above reasons, the current research implements SHAP to explain the output for the transportation prediction model with a SHAP summary plot for the marginal contribution of each feature on the model output and SHAP partial dependence plots to analyze the interaction effects for feature pairs found to have greater contributions to the prediction in the SHAP summary plot (see the Results and Discussion Section for the Transportation Energy Model).

Synthetic Household Dataset

The 2015 DVRPC HTS household data is sampled from Traffic Analysis Zones (TAZs) within the Delaware Valley region, and the ACS census data is drawn from geographically larger Minor Civil Divisions (MCDs) in order to serve as controls for the synthesized TAZs. The PopGen 2.0 software package is applied for producing the synthetic population. The PopGen algorithm takes in the ML output household samples as a joint probability distribution across categorical features. ACS data is drawn on the scale of Minor Civil Division (MCD) and TAZ numerical totals for each respective subdivision. MCDs are significantly larger than TAZs, such that numerous TAZs are contained in each MCD. Each sample is given a weight according to variable prevalence in the full synthesized population through an Iterative Proportional Updating (IPU) algorithm in the PopGen software. The variables are selected according to which HTS features have corresponded marginal columns in the ACS dataset, and conversely the sample variables are chosen by SHAP for feature importance. The resulting output is a synthetic dataset with the number of household and person samples equivalent to the region's total number of households and persons. Each sample contains information on energy consumption in daily kilowatt-hour (kWh) and the sample location in terms of TAZ.

Household size, household workers, and number of vehicles per household make up the categorical sums per TAZ, and the total number of households and population count per MCD comprise of the region marginal totals. Daily transportation energy consumption in kWh is appended to the sample at the TAZ level. The synthesized dataset of the total household transportation energy consumption per TAZ is aggregated by MCD and for the entire county.

Forecast Scenarios

The DVRPC Greater Philadelphia Future Forces provides a series of potential outcomes for the Delaware Valley Region based on the interaction between population, employment, urban and transportation infrastructure development, income per capita, and other relevant variables. These "what-if" scenarios are exploratory and mostly qualitative possibilities for the region. One of the potential future forecasts for Philadelphia is entitled "Enduring Urbanism" and describes an acceleration of urbanism into 2045. In this scenario, population and employment opportunities become disproportionately concentrated in urban areas. In turn, public transit and pedestrian developments are prioritized and more sustainable travel modes are widely accessible to residents. The corresponding downside is that the flow of new residents leads to gentrification, with lower income households being pushed into now fiscally distressed suburbs.

From the Enduring Urbanism forecast, this study constructs a base case and two scenarios. The first scenario represents a more optimistic interpretation of the Enduring Urbanism future in terms of more walkable urban areas and increased public transit use. Scenario I only focuses on impacts related to transportation energy use for households. Scenario II, in contrast, analyzes the effect on gentrification as described in the forecast. The descriptions of each of the Enduring Urbanism scenarios are described in the following below sections.

Base Case

In the Enduring Urbanism forecast, population increases by 17% from the year 2016 to 2045. The base case assumes a proportional change in population across TAZs for a total of 1.82 million persons residing in Philadelphia County in the year 2045. The Enduring Urbanism forecast has no quantitative predictions on the change in number of households for Philadelphia. As the average household size for Philadelphia County is

2.55 persons per household [20], the study assumes this ratio will stay consistent in the base case and the number of households will also increase by 17% for a total of 713,725 households in the region. For the DVRPC HTS of the transportation model, the study alters the numerical population and household variables of the sample according to these assumptions. However, all other variables remain constant in the energy use estimation such that the added households will have a proportional per capita increase if they are represented as marginal totals per subdivision and kept constant if they are not.

The transportation energy use totals are aggregated according to three types of geographic subdivisions for Philadelphia County: Traffic Analysis Zones (TAZs), Minor Civil Divisions (MCDs), and Public Use Microdata Areas (PUMAs). There are 688 TAZs, 17 MCDs, and 11 PUMAs within Philadelphia County. In order to minimize error, this study refrains from converting between geographic units when altering datasets according to scenario assumptions as the type of subdivision varies across energy usage sample data and the subdivision categories frequently overlap (i.e., do not have a perfect one-to-one correspondence).

Scenario I

Implementation of the 30th street station development

The first scenario analyzes the effects on household transportation energy use given the successful implementation of the 30th Street Station Development in Philadelphia. This development proposes to build a 100-acre neighborhood over the rail yards and enhance the connectivity of the existing 30th Street transit station. The District Plan aims to accommodate 20 to 25 million passenger trips, provide housing for up to 10,000 new residents and support up to 40,000 new jobs. The Technical Report envisions that this development, in addition to the concentration of employment opportunities and education opportunities in urban areas, will result in a shift in primary transportation modes for Philadelphia. The current study assumes that the most significant change in travel modes for the 30th Street Station Development will be within a 30-min commute to the TAZ containing the 30th Station District. Notably, only the radius of a public transit, walking, or biking commute are considered when selecting the study area as the Technical Report predicts a dramatic reduction in the use of personal vehicles and restricted automobile traffic in the inner-city. This region is approximately contained in three MCD neighborhoods: Central, South, and University/Southwest. These 194 TAZs are filtered from the 688 TAZs in Philadelphia County through the TravelTime QGIS application.

The first step in constructing the scenario based on the projected changes involves the altering of the DVRPC HTS for the households sampled from the target area TAZs. The dataset is adjusted to account for the 10,000 incoming residents inhabiting the 30th Station District TAZ and the 40,000 newly employed residents evenly distributed throughout the study area of the three MCDs. For all other variables found significant from the SHAP analysis, a proportionality table is applied to increase the numerical sample values according to the number of new households. As the DVRPC HTS dataset has a normal mean distribution across variables, the categorical samples are not altered. Originally there are 8310 samples for all of Philadelphia County. Out of this, the study filtered 696 samples contained in the target area and appended 157 new samples to account for the increased proportion of different travel modes.

The primary transportation mode per household sample for daily trips (HTS variable name: MODE) is altered in accordance with the predicted increase in transit ridership by 75%. Walking and biking modes increase by 150% compared to 2010. Subsequently, the number of daily household motorized trips (HH_MO_TRIPS), non-motorized trips (HH_NM_TRIPS), daily aggregated model travel time (Model_TravTime), and number of bikes per household (BIKE) change in accordance with the changes in transportation mode. See Supplementary Tables S1–S3 for more information on the mentioned variables.

Scenario II

Disappearance of mixed-income neighborhoods

For the second scenario, the study investigates the effect of gentrification on household transportation energy use. The DVRPC Technical Report predicts middle- and lower-income households will be pushed out of core urban areas for the Delaware Valley region to outlying suburbs. This what-if scenario outlines a problem area for the DVRPC: these inhabitants will be pushed to low-density suburban areas that have dwindling employment opportunities and a neglected transportation infrastructure, and in turn Philadelphia County will face the negative effects of reduced income diversity in its neighborhoods. Formerly mixed-income areas will become high-income, and the currently low-income areas will remain low-income to demonstrate the forecasted income stratification.

According to the census bureau’s definition of an urban geographical area, Philadelphia County is thoroughly urbanized as each TAZ contains at least 1000 people per square mile. As the technical report only referred to gentrification in the geographical context of urban versus rural areas, the study area includes all of Philadelphia County.

We begin the setup of our scenario in terms of PUMA regions, of which there are 11 total PUMAs for Philadelphia County. First, we will identify the income status (low-, mixed-, or high-income) for each PUMA region based on the DVRPC’s 2018 median income data. Currently there is no consensus for what quantifies a neighborhood as a mixed-income one. The study applies the following definition for classifying the income group of Philadelphia County PUMAs such that households are divided into three groups: those below 80%, between 80% to 120%, and above 120% of the Area Middle Income (AMI) [21]. If each of these groups make up to 20%, but no more than 50%, of total households in the region, then the PUMA is classified as a mixed-income area. In this way, a mixed-income PUMA demonstrates that a wide variety of income groups are represented in the neighborhood and no one group has a dominant presence. A high-income PUMA will have over 50% of households that have median incomes greater than 120% of the AMI, and a low-income PUMA will have over 50% of households that have median incomes less than 80% of the AMI.

There are various approaches to modeling displacement due to gentrification, from using lags in housing prices to studying the geographic proximity of neighboring tracts [22]. Under the Enduring Urbanism forecast assumptions, gentrification is primarily caused by the market's unwillingness to meet the demand for mixed-income housing with the negative effects compounded by the lack of new housing developments. From this, the study assumes that there is high local excess demand such that neighboring high-income neighborhoods are essentially "crowding out" former residents of mixed-income households [23]. To implement this, we take the mixed-income PUMAs and change the distribution of the household samples' median incomes to match that of a neighboring high-income PUMA that contains a relatively high number of intersecting TAZs.

For example, PUMA 3209 is high-income and PUMA 3211 is mixed-income due to the distribution of household median incomes in each respective region. As both PUMAs share a border, the distribution of household income in the energy consumption sample is manually altered such that PUMA 3211 has an identical distribution of income groups to high-income PUMA 3209. The post-scenario PUMA 3211 will have 21% of households earning less than 80% of the AMI, 11% earning between 80% and 120% of the AMI, and 68% earnings over 80% of the AMI.

RESULTS AND DISCUSSION

Transportation Energy Model

After creating a list of all the energy consumption-related variables, the study conducted a multi-step screening process and eliminated correlated or redundant variables to minimize collinearities. Variables with little theoretical relevance to transportation energy (e.g., the method of transit payment) or those with little variation (< 5%) were removed from the dataset. Also, variables that are directly used in transportation calculations (e.g., household travel distance) were removed to eliminate redundancies. As a result, 31 continuous and 12 categorical variables were selected to be included in the analysis. All the neighborhood characteristic variables (e.g., population, the total number of households, number of employees, etc.) are standardized, and the categorical variables are converted to binary variables.

The XGBoost model is implemented using the Python 3 scikit-learn function. Grid-search is used to optimize the hyperparameters. The MSE, MAE, and R^2 values are then calculated for both the training and test datasets across the developed XGBoost model. These results are corresponded to models that are trained using the best-obtained hyperparameters.

With respect to transportation energy consumption, the results show that the XGBoost model performs well. The results are summarized in Table 1 for the MSE, MAE, and R^2 of the training and test datasets.

Table 1. Cross-validation results for the XGBoost model.

	Training			**Test**		
	MSE	MAE	R^2	MSE	MAE	R^2
XGBoost	2.045	0.90	0.92	2.90	0.985	0.87

The global and local interpretability afforded by SHAP techniques allows urban planners and engineers to understand and trust the results of their ML-based energy predictions. The study applies a SHAP summary plot for the current study's transportation energy model to understand the importance of model features in terms of the range of their effect over the dataset (see Figure 1). The color of each dot in the SHAP plot indicates how change in the value of a feature affects the change in energy consumption for the household according to the gradient scale provided on the y-axis. The SHAP x-value for a given feature represents the effect that feature has on the transportation energy prediction model. The x-axis units explain the XGBoost model's change of margin output in the unit of log-odds. Overlapping points are scattered vertically, providing the distribution of SHAP values for each feature.

The model features are sorted according to their predictive power such that the highest feature of household motorized trips (HH_MO_TRIPS) is the most important predictor of transportation energy consumption, with household travel time (Model_TravTime) as the second highest predicting feature. From Figure 1, the colors indicate that being in a household with more frequent motorized trips and longer travel times increases the chances of consuming more transportation energy, and households with fewer motorized trips and shorter travel times have a higher chance of using less transportation energy. The number of household non-motorized trips (HH_NM_TRIPS) has a negative impact on the model's prediction of energy use (i.e., a higher number of non-motorized trips increases the chances of a lower transportation energy use prediction in the model). Conversely, the total number of vehicles per household (OP_VEH) increases the prediction. Other important features are the number of total vehicles per household (OP_VEH), household size (HH_SIZE), primary travel mode by private vehicle (MODE_5), number of household workers (HH_WORK), number of bus stops within a TAZ (BUSSTOP) and the household's home-to-work travel type (TOUR_TYPE_1).

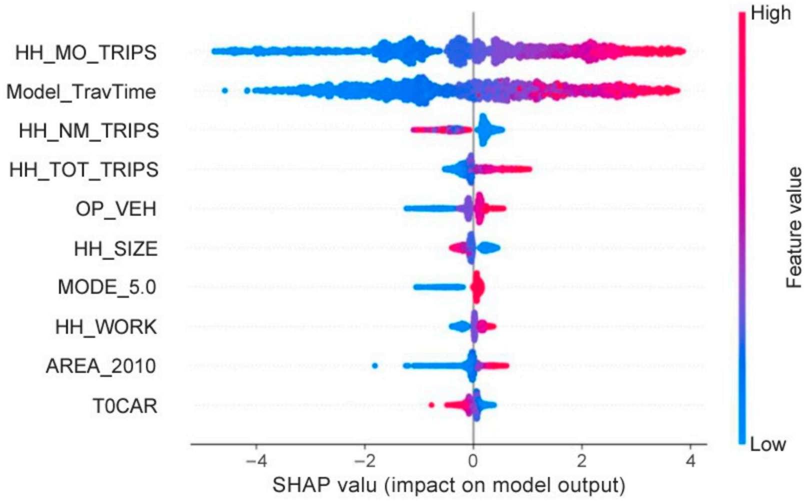

Figure 1. SHAP summary plot for feature importance.

While a SHAP summary plot gives a general overview of each feature, a SHAP partial dependence plot shows the marginal effect on the model prediction in terms of the interaction effects between two features. In this way, partial dependence plots demonstrate whether the relationship between the target feature and another selected feature is linear, uniform, or more complex.

With a SHAP partial dependence plot, we investigate the effect of the number of household motorized trips on the transportation energy use prediction model (Figure 2). The x-axis is the value of the selected feature (HH_MO_TRIPS), and the y-axis is the Shapley value for the household's transportation energy consumption with units in log-odds. The upward slope demonstrates that there is a positive trend between number of motorized trips and energy consumption.

Interaction effects drive the vertical dispersion of Shapley value with the gradient color bar ranging from blue to red for the feature of household travel time (Model_TravTime). Households with a low number of motorized trips that require longer travel times are less likely to have high energy use predictions (as shown by the red data points on the lower right-hand side of Figure 2) when compared to households with a high number of motorized trips and shorter travel times.

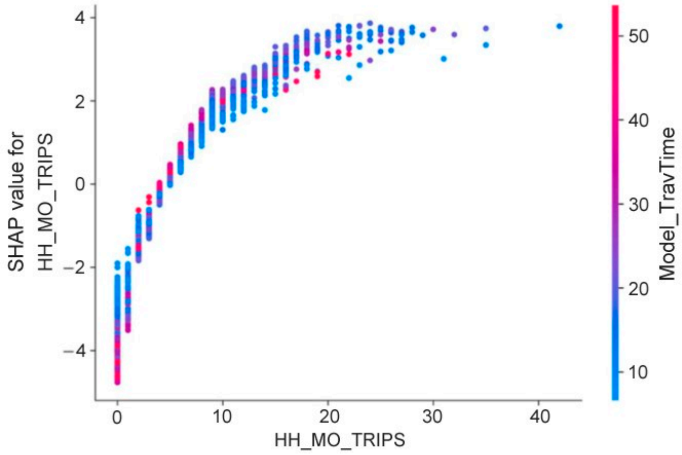

Figure 2. SHAP partial dependence plot demonstrating the marginal effect of the number of household motorized trips (HH_MO_TRIPS) and travel time (Model_TravTime) on the outcome of the transportation energy prediction model.

For further analysis, the study analyzes the marginal effect of the number of household non-motorized trips (HH_NM_TRIPS) and travel time on the model with a SHAP partial dependence plot (Figure 3). Figure 3 demonstrates that for households with two to five non-motorized trips, the predicted outcome of the transportation energy model is lower as the Shapley value for most of the data points are below zero. When there are zero non-motorized trips for the household, longer travel times lead to higher model outputs.

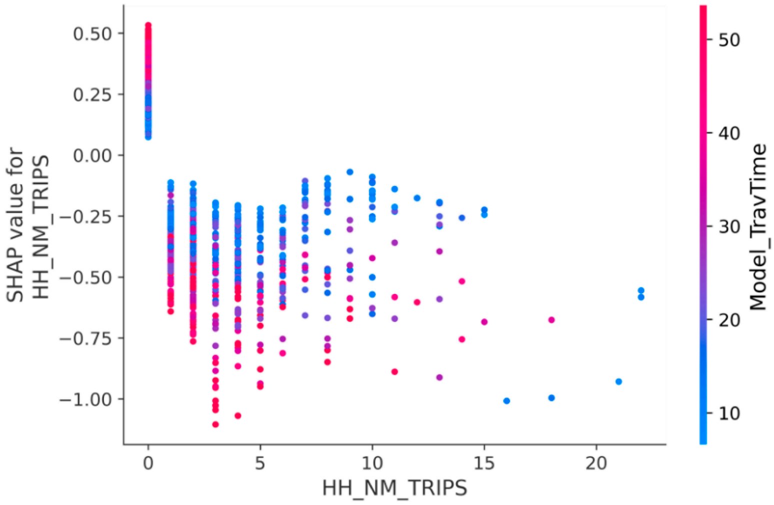

Figure 3. SHAP partial dependence plot demonstrating the marginal effect of the number of household non-motorized trips (HH_NM_TRIPS) and travel time (Model_TravTime) on the outcome of the transportation energy prediction model.

In Figure 4, the upward slope indicates that the higher the number of total trips for the household (HH_TOT_TRIPS), the higher the prediction in energy consumption. The figure suggests there is little interaction between number of total trips and number of operating vehicles (OP_VEH).

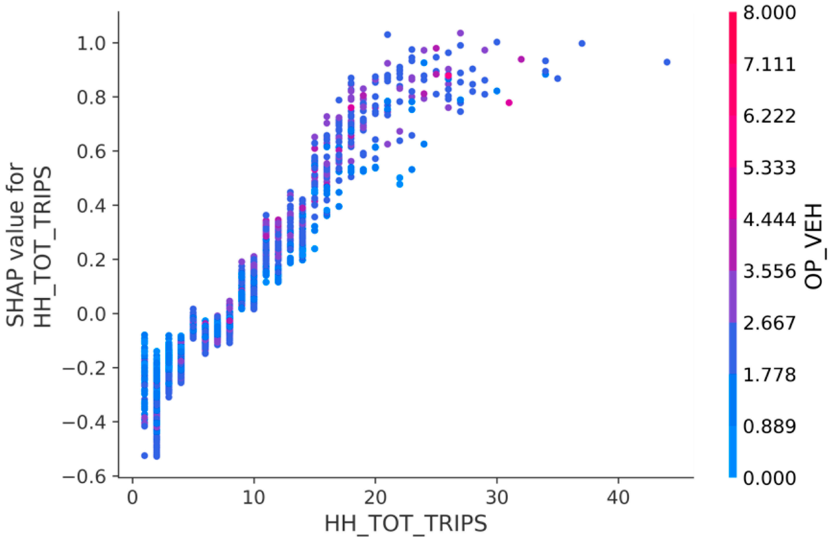

Figure 4. SHAP partial dependence plot demonstrating the marginal effect of the total number of household trips (HH_TOT_TRIPS) and number of operating vehicles per household (OP_VEH) on the outcome of the transportation energy prediction model.

PopGen is applied for producing the synthetic population output of transportation energy consumption. Based on the feature importance results from SHAP and available ACS data, variables are selected at the TAZ and MCD level. Household size, household workers, and number of vehicles per household make up the categorical sums per TAZ, and total number of households and population per MCD comprise the region marginal totals. Daily transportation energy consumption in kWh for both the base case and the scenarios are appended to the sample at the TAZ level. The synthesized dataset of the total household transportation energy consumption per TAZ is aggregated by MCD and for the entire county. Analysis of the PopGen results is continued below in the Scenario Results section for the scenario I study area MCDs (Central, South, University-Southwest) by TAZ. Although the one-way ANOVA test and SHAP method incorporate scenario II data at the TAZ level, descriptive statistics and heat maps of energy consumption are aggregated by MCD and by income status for that scenario due to the difficulty in discerning geographic patterns in energy consumption at such a disaggregate scale.

Scenario Results

Base case

Base case for scenario I

The base case for scenario I assumes an increase in population based on projections for the year 2045 and the successful implementation of the 30th Street Development. Per capita transportation energy use is aggregated over TAZs in terms of daily kWh (M = 78.48, SD = 39.08).

Base case for scenario II

The analysis of scenario II centers on changes in household income distribution for all of Philadelphia County. In this scenario, the base case assumes the population increases proportionally throughout the study area. Base case energy consumption is calculated by daily per capita kWh (M = 50.85, SD = 235.43).

With respect to energy use by MCD, Central is in the highest quantile of per capita energy consumption at 123.48 kWh/day and North consumes the least per capita energy at 10.27 kWh/day.

Scenario I results

A one-way analysis of variance (ANOVA) is calculated to compare the two treatment groups of base case energy use and scenario I energy use. The analysis is significant F(1, 380) = 61.79, p < 0.0001; scenario I households with the forecasted shifts in dominant travel modes consume less transportation energy per capita (M = 40.96, SD = 27.71) than base case households (M = 50.85, SD = 32.36).

Transportation energy use declines by nearly 30% due to the shift in dominant travel modes. Of the three MCDs, Central has the highest levels of energy use for both the base case and scenario I. At the same time, Central undergoes the sharpest decline in energy consumption at approximately −14 million kWh/day, thus contributing to nearly 70% of the change in total energy for the study area. However, when comparing energy consumption across TAZs, the results for Central demonstrate more diverse patterns. A significant majority of TAZs in the 10th percentile of scenario energy consumption are from the Central MCD, yet TAZs in Central are simultaneously prevalent in the top 90th percentile. South also contains TAZs in the top 90th percentile with no presence in the lowest 10th percentile. University-Southwest uses the least amount of transportation energy at 16 kWh per capita and has minimal presence in either extreme of energy consumption for both the base case and scenario I. Table 2 provides a comparison of the base case and scenario I energy outputs for the study area.

Table 2. Base case and scenario I comparison with total daily energy (kWh/day) and per capita energy (kWh/day/person) for study area MCDs (aggregated across TAZs).

MCD	Base Case Energy	Base Case Energy Per Capita	Scenario I Energy	Scenario I Energy Per Capita
Central	19,391,468	123	4,839,702	30
South	7,677,989	53	2,892,468	20
University/Southwest	6,017,934	58	1,686,878	16

The change in energy use from the base case is displayed in Figure 5 with each study area TAZ assigned a value based on an equal quantile legend for daily per capita decrease. The Central MCD contains the highest concentration of TAZs in the 75th quantile of decrease of daily per capita energy use.

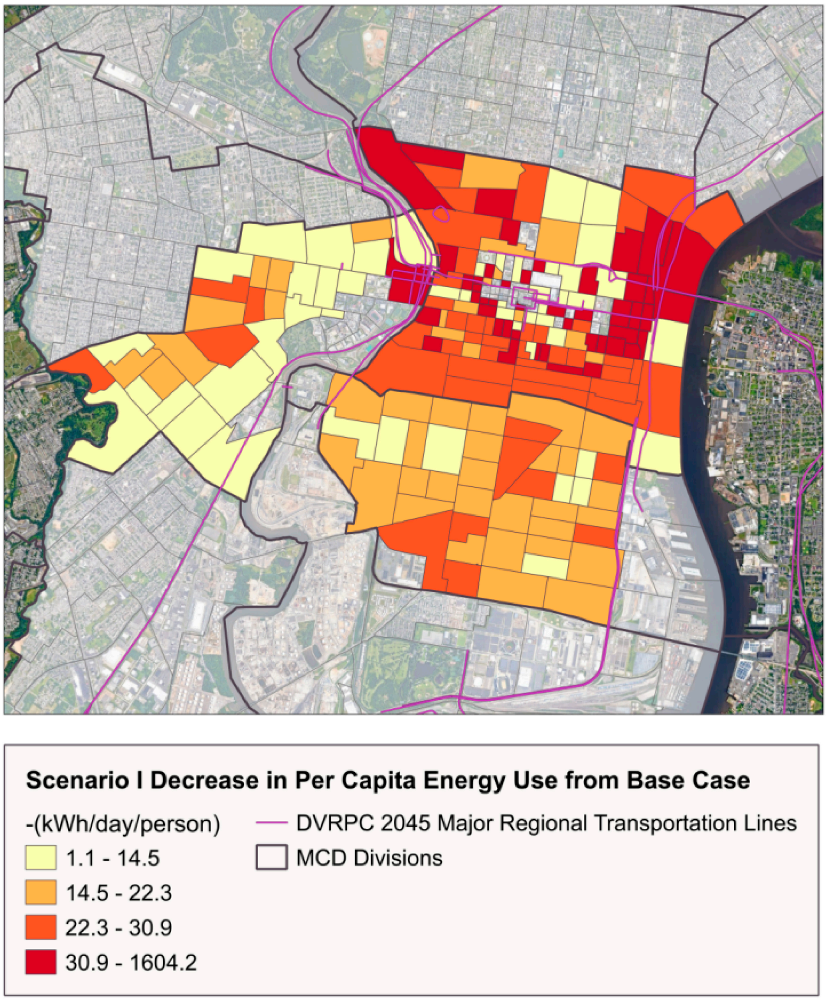
Scenario I Decrease in Per Capita Energy Use from Base Case
-(kWh/day/person)
1.1 - 14.5
14.5 - 22.3
22.3 - 30.9
30.9 - 1604.2
DVRPC 2045 Major Regional Transportation Lines
MCD Divisions

Figure 5. Scenario I change in transportation energy consumption heat map for study area MCDs.

The synthesized output is put through SHAP for the TAZs in the top 90th and lowest 10th quantile of energy consumption. For all three MCDs, the variable of number of motorized trips per household is the strongest explanatory feature of energy consumption, with daily aggregated model travel time the second highest feature. In Central, population is the third most important explanatory feature in SHAP, whereas for South and University-Southwest transportation energy use is better explained by travel mode (i.e., private automobile, van, or truck). Refer to Figures 6–8 for the SHAP plots of Central, South, and University-Southwest, respectively.

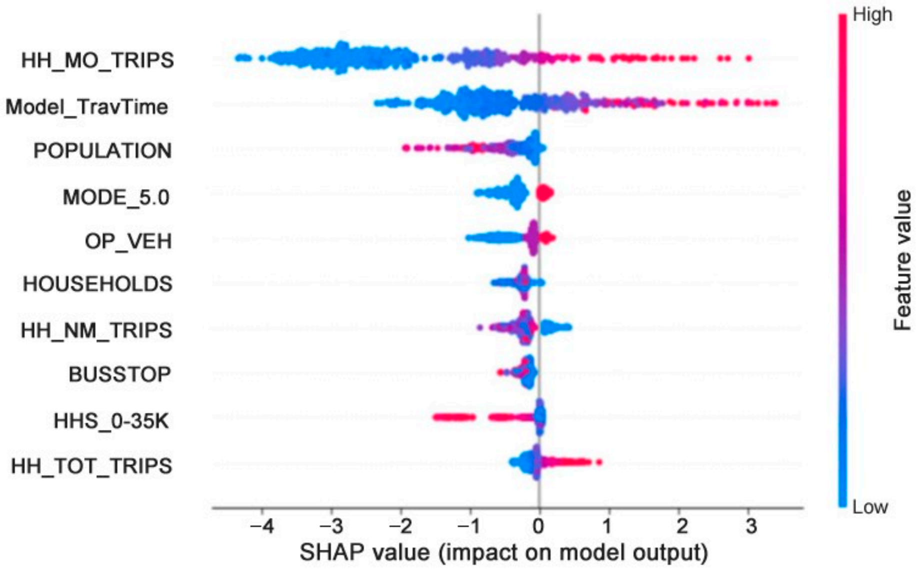

Figure 6. SHAP results for scenario I transportation energy use in the Central MCD.

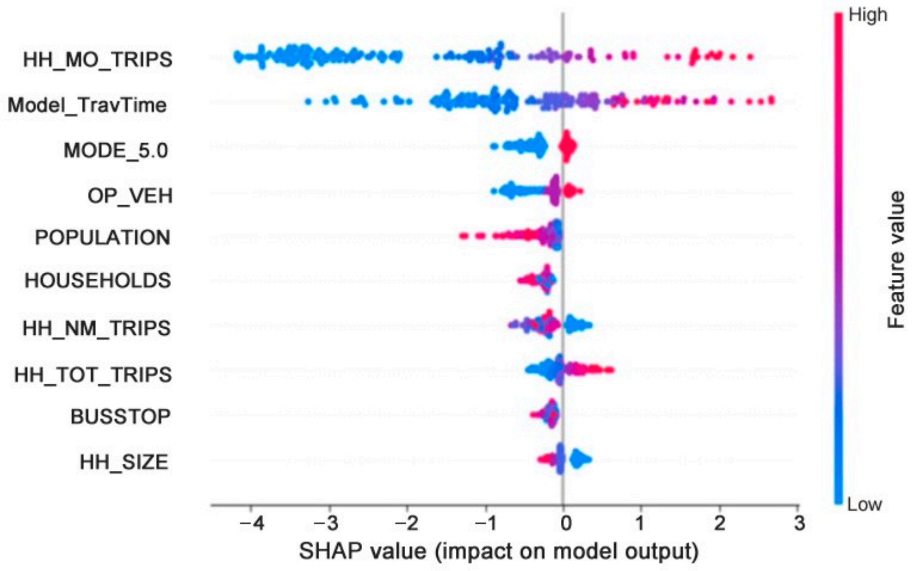

Figure 7. SHAP results for scenario I transportation energy use in the South MCD.

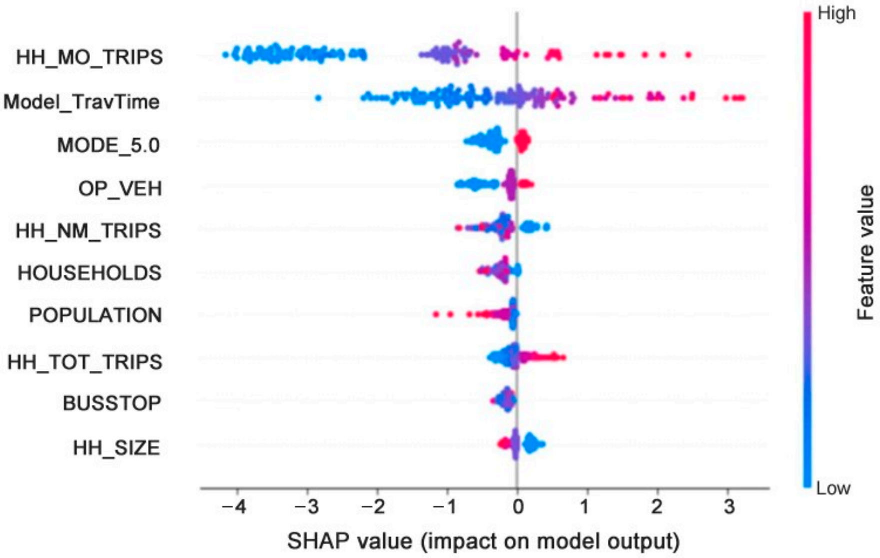

Figure 8. SHAP results for scenario I transportation energy use in the University-Southwest MCD.

Notably, scenario I did not involve a dramatic alteration of the data; only 157 household samples were appended to a dataset originally containing 8310 samples for the county, and the change in population and household count was disproportionately focused on the 30th Street Station Development. Nonetheless, there was a significant decrease in transportation energy consumption for all of Philadelphia County (48 daily kWh per capita in the base case to 34 daily kWh per capita in scenario I). The scenario indicates that the 30th Street Station Development can contribute significantly to meeting Philadelphia's GHG emission reduction benchmarks given the participation of the public in utilizing the new transit developments as primary travel modes over private motorized vehicles.

Scenario II results

Scenario II resulted in an approximate 11% decrease in transportation energy use, with daily per capita consumption at 42 kWh for the county. The one-way ANOVA test shows that the base case and scenario II are significantly different in terms of transportation energy consumption Philadelphia County F(1, 860) = 4.52, p =0.03. Scenario II households consume less transportation energy per capita (M = 146.24, SD = 286.49) than the base case households (M = 51.17, SD = 33.06).

For simplicity, the study refers to high-income PUMAs in scenario II that were previously mixed-income PUMAs in the base case as "HM" to demonstrate that the household income distribution was manually altered in the scenario to represent the gentrification process. "H" represents high-income PUMAs that were not previously mixed-income and "L" represents low-income PUMAs, either for the base case or the scenario.

For both the base case and scenario II, high-income (H) PUMAs have the greatest total consumption and low-income PUMAs the least. However, low-income (L) PUMAs consume the most transportation energy on a per capita basis at nearly 75% of the total. High-income (H) PUMA energy use adds up to approximately 55 million kWh, or 74% of the county total. In the base case and scenario II, both mixed-income (M) and formerly mixed-income (HM) PUMAs consume the least amount of energy per capita, with the formerly mixed-income (HM) group experiencing a slight decrease in scenario II. The results are outlined in Table 3 below.

Table 3. Base case and scenario II transportation energy totals and per capita use with respect to PUMA income status.

PUMA Income Status	Base Case Energy	Base Case Energy Per Capita	PUMA Income Status	Scenario II Energy	Scenario II Energy Per Capita
L	21,886,053	99,031	L	18,617,947	115,482
M	24,943,129	9,980	HM	22,846,325	8962
H	34,682,410	27,689	H	31,283,518	29,638
Totals:	81,511,592	136,700		72,747,790	154,081

Figure 9 shows the decrease in transportation energy for scenario II with TAZs aggregated by MCD rather than PUMA for the sake of comparison with scenario I. Similar to scenario I, the most dramatic drop in energy use is attributable to the Central MCD. The MCDs surrounding Central (University-Southwest, South, and Lower South) are all in the top 50th quantile of change in energy consumption.

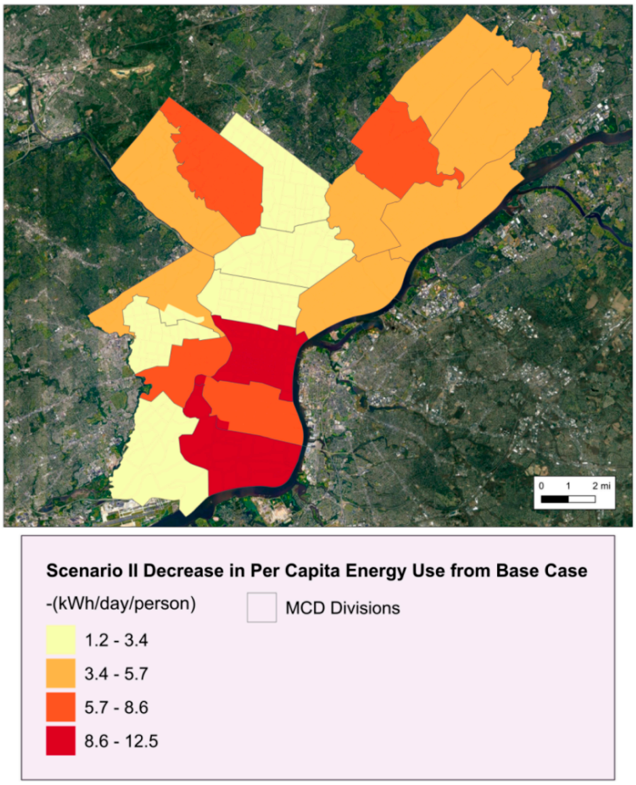

Figure 9. Scenario II change in transportation energy consumption heat map for Philadelphia County.

In order to further investigate the impact of changes in income group on energy consumption, the study includes a secondary one-way ANOVA test for the low- (L), formerly mixed- (HM), and high-(H) income groups energy consumption in scenario II. The results demonstrate significant differences F(2, 483) = 4.84, *p* =0.008. A post-hoc Tukey-Kramer analysis for pairwise comparisons (see Table 4) indicates that the L group consumes more per capita energy than their high-income counterparts. The H and HM groups are not significantly different in terms of energy use, as expected given their matching income distributions and the initial ANOVA test results. Lastly, the L group consumes more per capita energy than the HM group.

Table 4. Tukey-Kramer test for scenario II with treatments as PUMA income status and value as scenario II transportation energy for TAZs in Philadelphia County.

Income Status		Diff	Lower	Upper	*q*-value	*p*-value
H	L	55135.23	6994.18	103276.29	3.81	0.02
H	H^M	4485.34	−46721.91	55692.60	0.29	0.90
L	H^M	59620.58	7084.35	112156.80	3.77	0.02

The SHAP results for scenario II indicate that the number of household motorized trips and daily aggregated model travel time are the strongest explanatory features for transportation energy consumption in Philadelphia County (Figure 10). Household number of non-motorized trips also demonstrates a sizable impact on energy use for the region. The lower income (annual household income <$35K) feature is also present in the model, albeit with the least impact of the 20 SHAP features. Notably, the variable for household income in the SHAP analysis only has two categories (annual household income <$35K and annual household income >$35K), whereas the setup for scenario II involves changes to the dataset at a more detailed numerical scale. Thereby, the analysis of Shapley values does not provide in-depth insight on the role income distribution changes had on energy consumption.

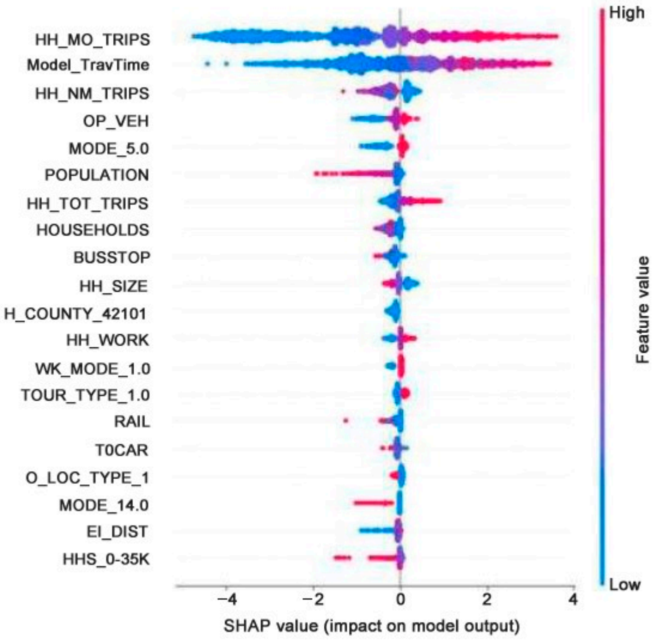

Figure 10. SHAP results for scenario II transportation energy use.

From the above results, gentrification in the Enduring Urbanism forecast leads to a decrease in transportation energy use for the county. In general, low-income TAZs consume more per capita energy, suggesting that the decline in total energy consumption is attributable to the scenario’s reduction of households that earn less than 80% of the AMI. This result is corroborated by previous research that indicates low-income households have greater energy burdens than their higher-income counterparts [24]. It is important to note that mixed-income PUMAs consistently consume less than 1% of the total per capita energy, both before and after the change in income distribution (M and HM). This implies a potential relationship between neighborhoods with greater income diversification and reduced rates of transportation energy use.

CONCLUSION

This study implemented an integrated bottom-up approach using the DVRPC HTS and the ACS datasets to estimate the impact of forecasted trends in household-level behavior on transportation energy consumption for Philadelphia. Through the application of a ML XGBoost algorithm for increased predictive model reliability, the study compensates for the weaknesses in bottom-up energy models and successfully produced a comprehensive synthetic population for a large region. Applying this method as a model for Philadelphia County offers insights on energy consumption patterns in the DVRPC Enduring Urbanism forecast. The results suggest that the proposed 30th Street Station Development and the resulting shift in dominant travel modes will be responsible for the strongest reduction of transportation energy use. Gentrification and the absence of mixed-income neighborhoods will also lead to a decline in transportation energy consumption. The findings demonstrate that mixed-income PUMAs are associated with low transportation energy consumption in the model. The relationship between the income diversification of neighborhoods and reduced transportation energy use warrants further research. Given the Enduring Urbanism forecast is consistent with demographic and socioeconomic trends in 2045, the Central MCD is a critical area of focus for meeting long-range plan energy benchmarks. Central consumes the most total energy across scenarios and demonstrates unique behavior in terms of how population density affects transportation energy use at the TAZ level. Moreover, the current research proposes a methodological template for applying reliable energy models to synthesized data in a scenario-based context. With this approach, future research can apply this method for residential, commercial, and other energy use models in the context of predictive scenario planning. In this way, urban planners and policy makers can target potential problem areas and strategize on sustainability initiatives with a high level of quantitative detail and spatial resolution.

SUPPLEMENTARY MATERIALS

Supplementary Table S1: Descriptive statistics of household demographics.

Supplementary Table S2: Descriptive statistics of trip characteristics.

Supplementary Table S3: Descriptive statistics of zone characteristics.

AUTHOR CONTRIBUTIONS

Shideh Shams Amiri: Conceptualization, Methodology, Data curation, Software, Formal analysis, Visualization, Writing—Original draft preparation, Validation.

Maya Mueller: Writing—Reviewing and Editing, Methodology, Formal analysis, Software, Visualization.

Simi Hoque: Supervision, Conceptualization, Writing—Reviewing and Editing and Funding acquisition.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Bates B, Kundzewicz ZW, Wu SH, Arnell N, Burkett V, Döll P, et al. Climate Change and Water: Intergovernmental Panel on Climate Change. Available from: https://www.ipcc.ch/site/assets/uploads/2018/03/doc13-5.pdf. Accessed 2022 Feb 21.
2. Hirst E. A model of residential energy use. Simulation. 1978;30(3):69-74.
3. Jumbe CB. Cointegration and causality between electricity consumption and GDP: empirical evidence from Malawi. Energy Econ. 2004;26(1):61-8.
4. Lakshmanan T, Han X. Factors underlying transportation CO_2 emissions in the U.S.A.: A decomposition analysis. Transp Res D Transp Environ. 1997;2(1):1-15.
5. Saha GP, Stephenson J. An evaluation of residential energy conservation strategies in New Zealand. Energy. 1980;5(5):445-50.
6. Saidi K, Hammami S. The impact of CO_2 emissions and economic growth on energy consumption in 58 countries. Energy Rep. 2015;1:62-70. doi: 10.1016/j.egyr.2015.01.003
7. Zhang Q. Residential energy consumption in China and its comparison with Japan, Canada, and USA. Energy Build. 2004;36(12):1217-25.
8. Hirst E, Goeltz R, White D. Determination of household energy using 'fingerprints' from energy billing data. Int J Energy Res. 1986;10(4):393-405.
9. Belaïd F, Roubaud D, Galariotis E. Features of residential energy consumption: Evidence from France using an innovative multilevel modelling approach. Energy Policy. 2019;125:277-85. doi: 10.1016/j.enpol.2018.11.007
10. Garikapati VM, Pendyala RM, Morris EA, Mokhtarian PL, McDonald N. Activity patterns, time use, and travel of millennials: a generation in transition? Transp Rev. 2016;36(5):558-84.
11. Wilson D, Swisher J. Exploring the gap: Top-down versus bottom-up analyses of the cost of mitigating global warming. Energy policy. 1993;21(3):249-63.
12. Zhang W, Guhathakurta S, Pendyala R, Garikapati V, Ross C. A Generalizable Method for Estimating Household Energy by Neighborhoods in US Urban Regions. Energy Procedia. 2017;143:859-64. doi: 10.1016/j.egypro.2017.12.774
13. Amiri SS, Mostafavi N, Lee ER, Hoque S. Machine learning approaches for predicting household transportation energy use. City Environ Interact. 2020;7. doi: 10.1016/j.cacint.2020.100044
14. Reiter S, Marique AF. Toward Low Energy Cities: A Case Study of the Urban Area of Liége, Belgium. J Ind Ecol. 2012;16(6):829-38.
15. Konduri KC, You D, Garikapati VM, Pendyala R. Enhanced Synthetic

Population Generator That Accommodates Control Variables at Multiple Geographic Resolutions. Transp Res Rec. 2016;2563(1):40-50.

16. Jiang Y. Does energy follow urban form?: an examination of neighborhoods and transport energy use in Jinan, China [dissertation]. Massachusetts (US): Massachusetts Institute of Technology; 2010.
17. National Transit Database. 2013 Table 19: Transit Operating Statistics Service Supplied and Consumed. Available from: https://www.transit.dot.gov/ntd/data-product/2013-table-19-transit-operating-statistics-service-supplied-and-consumed. Accessed 2022 Feb 21.
18. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Available from: http://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/0/02b26cfa6ecc8cd3c12583d9006de8c2/$FILE/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Accessed 2022 Feb 21.
19. Molnar C. Interpretable machine learning. A Guide for Making Black Box Models Explainable. Available from: https://christophm.github.io/interpretable-ml-book/. Accessed 2022 Feb 21.
20. 2015–2019 American Community Survey 5-Year Estimates [Internet]. Available from: https://www.census.gov/data/developers/data-sets/acs-5year.html. Accessed 2022 Feb 21.
21. Kneebone E, Reid C, Holmes N. Spatial Context: The Geography of Mixed-Income Neighborhoods. Available from: https://case.edu/socialwork/nimc/sites/case.edu.nimc/files/2019-04/Spatial%20Context%20Kneebone%20Reid%20Holmes..pdf. Accessed 2022 Feb 21.
22. Cohen M, Pettit KLS. Guide to Measuring Neighborhood Change to Understand and Prevent Displacement. Available from: https://www.urbanwaterslearningnetwork.org/wp-content/uploads/2019/04/guide_to_measuring_neighborhood_change_to_understand_and_prevent_displacement.pdf. Accessed 2022 Feb 21.
23. Piazzesi M, Schneider M, Stroebel J. Segmented Housing Search. Am Econ Rev. 2020;110(3):720-59.
24. Drehobl A, Ross L, Ayala R. How High Are Household Energy Burdens? An Assessment of National and Metropolitan Energy Burdens across the United States. Available from: https://www.energy.gov/sites/default/files/2021-12/ACEEE%2C%20Household%20Enegy%20Burdens.pdf. Accessed 2022 Feb 21.

How to cite this article:
Amiri SS, Mueller M, Hoque S. Investigating the Application of a Transportation Energy Consumption Prediction Model for Urban Planning Scenarios in Machine Learning and Shapley Additive Explanations Method. J Sustain Res. 2022;4(1):e220001. https://doi.org/10.20900/jsr20220001