


Applying model-based recursive partitioning to improve pedestrian exposure models to support transportation safety analyses

Jakob C. Wiegand^{1*}, Vikash V. Gayah¹

¹The Pennsylvania State University, the United States of America 

*Corresponding author: jakob.wiegand@psu.edu

Guest editor: **Nikiforos Stamatiadis**, University of Kentucky, the United States of America

Reviewers: **Mariusz Kiec**, Cracow University of Technology, Poland
Mike Sewell, Gresham Smith, the United States of America

Received: 25 May 2024; Accepted: 18 November 2024; Published: 17 January 2025

Abstract: Pedestrians are among the most vulnerable road users in urban areas, and their safety is a growing concern for transportation planners and engineers. Pedestrians are at disproportionately high risk for injuries or fatalities in crashes with motor vehicles, highlighting the critical need to address their safety. To address the dangers urban pedestrians face, the relationship between pedestrian crashes and their contributing factors must first be understood. One way to do this is to use statistical models relating pedestrian crash frequency with quantifiable contributing factors, such as land use, demographics, and roadway characteristics. Perhaps the most important of these factors is pedestrian exposure, which is often difficult to obtain because pedestrian volumes are not as widely available as vehicle volumes. Since pedestrian volumes are not available across an entire network, they are often estimated using statistical models—for example, negative binomial (NB) regression—rather than being directly observed. These models are typically a ‘one-size-fits-all’ approach, applying the same model to estimate pedestrian exposure across the entire network. However, relationships between pedestrian exposure and explanatory features—such as population, infrastructure design, and land use context—might differ significantly with respect to the context of an individual location. To address this issue, this paper proposes a model-based recursive partitioning (MBRP) algorithm to develop pedestrian exposure models. The MBRP approach combines traditional statistical methods (e.g. NB regression) with recursive data partitioning techniques commonly found in tree-based machine learning methods. This innovative approach yields a collection of exposure models stratified according to selected input variables with unique relationships between explanatory variables and exposure. The proposed method was tested on pedestrian exposure data from North Carolina significantly improved predictions of pedestrian volumes by approximately 10%. Therefore, the MBRP algorithm presents a promising tool for advancing pedestrian safety analyses in practical applications.

Keywords: exposure model, model based recursive partitioning, negative binomial regression, pedestrians, pedestrian safety

1 Introduction

Pedestrian safety is a topic of growing concern in the United States, especially in urban areas where 84% of pedestrian fatalities in traffic crashes occur (NHTSA, 2023b). Widely considered to be the most vulnerable roadway users, pedestrians are disproportionately represented in fatality statistics, and their representation is growing at an alarming rate. Data from the Fatality Analysis Reporting System (FARS) indicates that 7 388 pedestrians were killed in traffic crashes in 2021, about 19% of all traffic fatalities in the United States. As the greatest pedestrian fatality total since 1981, that pedestrian fatality total also represents a 12.5% increase from the previous year, following a trend in yearly increases dating back to 2013 when pedestrian fatalities accounted for 3% less of all traffic fatalities (NHTSA, 2023a). To address the growing concern for pedestrian safety, we must first understand the relationship between factors that contribute to pedestrian crashes and crash outcomes.

Often these relationships are quantified through statistical models that relate frequency of pedestrian crashes with observable factors that contribute to the crashes. Many different models have been used to investigate the relationships between pedestrian crashes and their contributing factors – most often including combinations of variables such as land use, socio-demographic information, and roadway characteristics. But perhaps the most important predictor of pedestrian safety outcomes is pedestrian exposure. Unfortunately, pedestrian exposure can be difficult to obtain because pedestrian volumes are not as widely available across a roadway network as vehicular volumes. This is because network-scale pedestrian counts are resource-intensive due to high costs associated with labor for manual counts and prices charged for automated counts. Some pedestrian volume data is more widely available from fitness-tracking sources such as Strava, but data from fitness-trackers is flawed in that they relay a self-reporting sample and the trackers are subject to error due to poor GPS reception, which is especially an issue in urban environments (Lee & Sener, 2020). Accurate volumes can also be difficult to obtain and even impractical to use due to highly variable daily volumes, short trips which may not be observed, and difficulties in detecting individuals (Lagerwey et al., 2015). Instead of directly using counts directly, pedestrian volumes are usually estimated through statistical models.

Several studies have developed pedestrian exposure models to predict the amount of pedestrian activity at a given level based on input variables that reflect the built environment, roadway features and other variables. These have traditionally been developed using log-linear ordinary least squares regression (OLS) or negative binomial (NB) regression (Behnam & Patel, 1977; Griswold et al., 2019; Hankey et al., 2012; Haynes et al., 2010; Lindsey et al., 2006, 2007; Liu & Griswold, 2009; Miranda-Moreno & Fernandes, 2011; Pulugurtha & Repaka, 2008; Schneider et al., 2009, 2012). Pedestrian exposure has also been estimated using Tobit models and by modifying NB regression techniques to artificially inflate zero value pedestrian counts, in both cases relating the counts to demographics, land use, and traffic data (Lee et al., 2019). Still other studies suggest using stepwise linear regression to account for spatial variations in independent variables (Hankey & Lindsey, 2016; Hankey et al., 2017; Lu et al., 2018).

While these model types have their own merit, they also have various flaws when applied to estimating pedestrian exposure. OLS regression does not account for the count nature of volume data, and using a log-linear form assumes a logarithmic distribution that may not be observed. Similarly, Tobit models assume a normal distribution for the dependent variable which is not typically observed in pedestrian count data. Stepwise linear regression may result in atheoretical coefficient estimates, which makes interpretation complicated and may limit transferability to other datasets. NB regression is the most appropriate and the most common, because of its ability to account for overdispersion in fluctuating pedestrian volumes and the count nature of the data. Even so, all these models are flawed in that they are typically a ‘one-size-fits-all’ approach in which the same model is used to estimate pedestrian exposure at all locations within a transportation network. However, the relationships between pedestrian exposure and explanatory features—such as population, infrastructure design, and land use context—might differ significantly with respect to the context of an individual location, which may not be known a priori. Incorporating these differences could help improve the exposure model and provide more accurate predictions.

To help address this issue, this paper proposes a model-based recursive partitioning (MBRP) algorithm to develop pedestrian exposure models. The MBRP approach combines traditional statistical methods (e.g.

NB regression) with the recursive data partitioning techniques commonly found in tree-based machine learning methods. The proposed method was tested on pedestrian exposure data obtained in North Carolina and shown to significantly improve predictions of pedestrian volumes by approximately 10%. Therefore, the MBRP algorithm presents a promising tool for advancing pedestrian safety analysis in practical applications.

The remainder of this paper is organized as follows. Section 2 provides a description of the methodology used for NB regression and the proposed MBRP approach. Section 3 describes the dataset used. Section 4 provides an analysis of the results from the MBRP model. Finally, Section 5 contains concluding remarks.

2 Methodology

The goal of this research is to demonstrate the potential of the MBRP algorithm to estimate pedestrian exposure better than traditional regression methods. To do so, exposure models were developed over a training dataset using traditional NB regression methods and the MBRP algorithm. The performance of the MBRP model was judged relative to the NB regression model based on goodness of fit statistics over a separate test dataset and cumulative residual (CURE) plots.

2.1 NB regression

Pedestrian counts are always a non-negative integer and are therefore most appropriately modeled using count models. Though there are many count regression models, NB regression models are used most extensively in research due to their ability to account for overdispersion in the dataset, which is commonly observed in pedestrian count data. This paper's NB regression procedure is adapted from [Hankey et al. \(2012\)](#) and [Lee et al. \(2019\)](#).

NB regression can be described through the following formulation: Let $i = 1, 2, 3, \dots, N$ represent the index of a given location where N is the number of locations in the dataset. In the NB model, the pedestrian count at a location i , takes an exponential form as shown in Equation (1):

$$\lambda_i = E(y_i) = \exp(\beta X_i + \varepsilon_i) \quad (1)$$

where λ_i is the predicted pedestrian count at location i , y_i is the observed pedestrian count at location i , β is the vector of estimated parameters, X_i is the associated

vector of explanatory variable values observed at location i , and ε_i is an error term such that $\exp(\varepsilon_i)$ has a gamma distribution. A maximum likelihood estimation (MLE) method was adopted using the probability distribution for NB regression as presented in Equation (2), and the likelihood function as presented in Equation (3). The MLE method allows for estimation of the coefficient parameter, β , and the overdispersion parameter, α , relying on the gamma function, denoted by $\Gamma(\cdot)$.

$$P(y_i) = \frac{\Gamma\left(\frac{1}{\alpha} + y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right) y_i!} \left[\frac{\alpha \lambda_i}{1 + \alpha \lambda_i}\right]^{y_i} \left[\frac{1}{1 + \alpha \lambda_i}\right]^{\frac{1}{\alpha}} \quad (2)$$

$$L(\alpha, \beta) = \prod_{i=1}^N P(y_i) \quad (3)$$

Selecting parameters which maximize the likelihood function presented in Equation (3) establishes a model expected to best fit the data. Typically, the likelihood value found by maximizing Equation (3) is very small, so frequently the natural log of Equation (3) is optimized in place and is reported as log-likelihood.

2.2 MBRP algorithm

The MBRP algorithm combines traditional statistical modeling, such as NB regression, with recursive data partitioning techniques commonly found in tree-based machine learning methods. In the MBRP algorithm, the root node is a parametric model fitted over the entire dataset. Child nodes are then formed through splitting based on a decision rule, which continues until the terminal node of the tree model is reached. Figure 1 depicts a general form of the tree model ([Kashani & Mohaymany, 2011](#)). The following formulation for the MBRP algorithm draws extensively from [Seibold et al. \(2016\)](#) and descriptions from [Tang & Donnell \(2019\)](#).

Developing a model using the MBRP algorithm is a process that occurs in three steps:

1. Fit a parametric model to all observations in the dataset;
2. Test coefficient stability over the splitting variable; and
3. Determine the optimal cut point of the splitting variable.

In Step 1, the parametric model with parameter vector, θ (represented by the root node in) may be estimated through methods such as OLS or MLE which inform the objective function, $\Psi(\cdot)$ as found in Equation (4). The coefficients are then estimated through a partial score function as shown in Equation (5), where $\psi(\cdot)$ represents the score function, and β is the estimated parameter (Seibold et al., 2016).

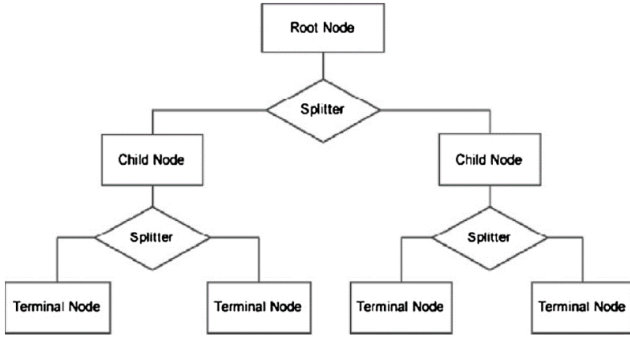


Figure 1 General structure of a tree model (Kashani & Mohaymany, 2011)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^N \Psi((y, X)_i, \theta) \quad (4)$$

$$\sum_{i=1}^N \frac{\partial \Psi((y, X)_i, \theta)}{\partial \beta} = \sum_{i=1}^N \psi_{\beta}((y, X)_i, \theta) \quad (5)$$

In Step 2, a generalized M-fluctuation test is used to test coefficient stability over splitting variables, with the null hypothesis stating that the partial score functions from Equation (5) are independent of partitioning variables. This hypothesis indicates that global estimates of an independent variable are appropriate. Equation (6) shows the functional form of the null hypothesis, where Z is the splitting variable and J is the number of splitting variables (Seibold et al., 2016).

$$H_0^{\beta_j} : \psi_{\beta}((Y, X), \hat{\theta}) \perp Z_j, j = 1, \dots, J \quad (6)$$

The splitting variable is selected based on the greatest correlation with partial score functions, and the optimal cut point of the splitting variable is determined by evaluating segmented objective functions as shown in Equation (7) and choosing the minimum value. In Equation (7), I_b is the set of observations belonging to b under splitting rule (Tang & Donnell, 2019).

$$SCORE = \sum_{b=1}^B \sum_{i \in I_b} \Psi(Y_i, \theta_b) \quad (7)$$

The process outlined above results in child nodes after the splitting rule is followed. The process is then repeated on all child nodes until terminal nodes are reached, where there is no longer coefficient instability as determined in Step 2 of the process. Each of the terminal nodes represents a parametric model, and any observation in the root nodes that meets the splitting criteria to fall into the terminal node can be predicted using the corresponding parametric model. Additional information about the MBRP algorithm can be found in Zeileis & Hornik (2007) and Zeileis et al. (2008).

3 Data

Pedestrian count data analyzed in this study comes from the North Carolina Department of Transportation (NCDOT) and was collected and analyzed for a previous study investigating factors contributing to pedestrian crash outcomes (Gayah et al., 2022). The counts are not a random sample, arising from a variety of convenience samples such as turning movement counts (TMCs) and pedestrian safety studies at particularly dangerous intersections or intersections with high-volumes relative to other intersections in the city. The non-random sample is still considered to be a valid sample because the counts are known to demonstrate a wide range of pedestrian count values, as well as to accurately represent the expected levels of pedestrian activity near the count location. The counts come from the following sources:

- 1 993 counts collected by NCDOT as part of turning movement counts (TMCs) or other similar analyses
- 496 counts collected between 2011 and 2020 by City of Charlotte TMC program as part of a FHWA funded research project
- 19 counts from downtown Raleigh, NC as collected for a NCDOT pedestrian safety study
- 539 counts from Greensboro DOT/Greenville Urban Area Metropolitan Planning Organization (MPO)
- 387 counts provided by Gaston-Cleveland-Lincoln MPO
- 184 counts provided by the City of Durham.

Each observation in the dataset included vehicular volume data, roadway features at the count location, census-level demographic statistics, and land-use statistics for a 0.5-mile radius surrounding the count location. A summary of the variables that were

included in final models is provided in Table 1 and Table 2. The land use mix variable is an aggregate of land uses in the surrounding area adapted from the methodology outlined in Frank et al. (2004) and Gayah et al. (2022). The land use mix value is based on four land use types: high intensity developed, medium intensity developed, low intensity developed, and all other land use classifications combined, where a value of 1.00 represents a perfect balance of all four land uses, and 0.00 represents only a single land use present. Alcohol sales locations are not commonly considered as an indicator variable; however, including this variable in indicator form can be justified through contextualization—the two categories represent the difference between an area with single liquor store and a couple of restaurants, and an area with densely clustered bars expecting high foot traffic and fewer vehicles. Including this variable in this form was found to improve model performance relative to a continuous form of alcohol sales locations.

Of the 3 618 counts provided, the majority were performed at intersections, and therefore intersections were selected as the unit of analysis. Count locations that were not intersections, in non-urban environments, and count durations less than or equal to 2.5 hours or equal to 24 hours were removed from the analysis to limit the analysis to characteristically similar count locations. The remaining 2 430 pedestrian counts ranged from 0 to 14 854 pedestrians over the respective durations, though 75% of counts registered 67 or fewer pedestrians. An 80/20 split was performed over the full 2 430 counts, creating a training dataset (1 942 observations) and a testing dataset (488 observations) such that both the NB and MBRP models would be estimated over the training dataset, and the models' fitness would be assessed using the goodness of fit statistics based on their predictions over the test dataset. The training and test datasets were selected to be representative samples of the full dataset with approximately equal summary statistics, but each dataset was also vetted to ensure representation of outlier counts. Due to the nature of the outlier counts, exact matches for maximum count values could not be achieved, and the training dataset is known to contain a greater proportion of lower count values because of the higher maximum count but otherwise similar summary statistics. Table 3 shows the summary statistics of the final split datasets.

4 Analysis and results

This section describes the estimation results from the NB regression model and the MBRP model, including the assessments of goodness of fit through mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). Additionally, a cumulative residual (CURE) plot was generated with both models plotted against the 95-percent confidence interval as determined for the NB regression model, according to the methodology outlined in Hauer (2015).

The results of the MLE process for the NB regression model and the model generated from the MBRP algorithm are presented in Table 4. In the modeling process, only the variables found to be statistically significant at the 5% significance level were retained. Exceptions were made for some variables found to be insignificant at the 5% level but were deemed theoretically important predictors of pedestrian counts. Additionally, any variables that were found to be inconsistent with theory regarding the sign of the estimated parameter were vetted for their contribution to the model's predictive performance before being considered for removal.

4.1 NB regression

All variables included in the final NB regression model were found to be statistically significant at the 95% confidence level or higher and were found to meet expectations consistent with engineering judgement:

- **Count duration.** In the dataset, the base condition is a 12-hour count. Therefore, it would be expected that a 13-hour count would yield a higher prediction due to a longer period for pedestrians to be observed. The same logic would indicate that a 16-hour count parameter should be greater than the 13-hour count parameter, however, 16-hour counts are typically conducted at two-way stop intersections being considered for signalization. Generally, these intersections have high traffic volumes which make pedestrian access uncomfortable, resulting in fewer pedestrians. Therefore, a smaller or even negative coefficient may be appropriate to describe the effects of a 16-hour count on pedestrians observed.
- **Traffic conditions.** As previously described, pedestrians are less comfortable walking near roadways with higher traffic volumes where they may be more vulnerable to crashes (Miranda-

Table 1 Continuous variables included in final NB and MBRP models

	Mean	Standard deviation	Minimum	Maximum
Land use mix	0.80	0.17	0.10	1.00
Proportion of non-motorized commuters within 0.25 miles of count location	0.07	0.09	0.00	0.55
AADT (veh/day)	20 260	11 133	1 200	77,000
Number of parcels within 0.25 miles of count location (parcel count)	92.98	91.93	2	568

Table 2 Discrete variables included in final NB and MBRP models

	1	0
Count duration = 13 hours (1 indicates yes, 0 indicates no)	74.49%	25.51%
Count duration = 16 hours (1 indicates yes, 0 indicates no)	5.27%	94.73%
Indicator for 0-5 alcohol sales locations within 0.25 miles of count location (1 indicates yes, 0 indicates no)	40.00%	60.00%
Indicator for ≥ 6 alcohol sales locations within 0.25 miles of count location (1 indicates yes, 0 indicates no)	29.29%	70.71%
Indicator for presence of sidewalk (1 indicates present, 0 indicates not present)	69.88%	30.12%
Indicator for presence of crosswalk (1 indicates present, 0 indicates not present)	32.76%	67.24%
Indicator for bus stop located within 0.25 miles of count location (1 indicates present, 0 indicates not present)	58.72%	41.28%
Indicator variable for posted speed limit ≥ 40 mph (1 indicates yes, 0 indicates no)	45.68%	54.32%

Table 3 Summary statistics of pedestrian counts

	Minimum	25th-percentile	Median	Mean	75th-percentile	Maximum
Full dataset	0	4	19	145.7	67.00	14 854
Training dataset	0	4	19	146.4	67.75	14 854
Test dataset	0	4	19	143.0	65.50	9 839

Moreno & Fernandes, 2011). That trend extends also to roadways with higher speed limits as crashes involving pedestrians and vehicles moving at higher speeds are known to result in more severe outcomes for pedestrians (Pulugurtha & Repaka, 2008). In the NB regression model, we find that the coefficients associated with the natural logarithm of AADT and speed limits greater than or equal to 40 miles per hour are negative, indicating a negative impact on pedestrian counts.

- **Pedestrian-friendly infrastructure.** Presence of features such as crosswalks and sidewalks are expected to contribute positively to pedestrian counts. Crosswalks provide priority to pedestrians crossing at intersections, and sidewalks provide a separate right of way for pedestrians to walk safely next to a road (Lee et al., 2019; Lu et al., 2018). These expectations are met by the positive

coefficients associated with presence of a crosswalk or sidewalk in the NB regression model. Bus stops are also associated with greater pedestrian presence due to the access distance between the trip origin/destination and the transit stop (Hankey et al., 2012; Pulugurtha & Repaka, 2008).

- **Land development.** Denser, more varied use of land is expected to result in greater pedestrian counts due to convenience of the walking mode. Increases in parcel count, a measure of urban density, should indicate shorter distances to potential destinations, while land use mix increases suggest a greater variety of residential, commercial, and industrial uses within a small radius (Gayah et al., 2022). Walking becomes more convenient over shorter distances, and therefore more pedestrians are likely to be observed with greater parcel count and land use mix. Additionally,

Table 4 NB regression and MBRP model estimation results

Variable	NB regression	MBRP	
		Parcel count \leq 71 'Low urban density'	Parcel count $>$ 71 'High urban density'
Constant	2.538	3.226	4.926
Count duration = 13 hours (1 indicates yes, 0 indicates no)	0.428	0.747	0.293
Count duration = 16 hours (1 indicates yes, 0 indicates no)	-0.408	<i>0.078</i>	-0.965
Land use mix	1.265	—	—
Indicator for 0–5 alcohol sales locations within 0.25 miles of count location (1 indicates yes, 0 indicates no)	0.172	<i>0.210</i>	0.603
Indicator for \geq 6 alcohol sales locations within 0.25 miles of count location (1 indicates yes, 0 indicates no)	0.680	0.743	1.054
Proportion of non-motorized commuters within 0.25 miles of count location	9.262	8.958	9.009
Natural logarithm of AADT (veh/day)	-0.301	-0.228	-0.260
Indicator for presence of sidewalk (1 indicates present, 0 indicates not present)	0.972	1.046	0.696
Indicator for presence of crosswalk (1 indicates present, 0 indicates not present)	0.576	1.019	<i>-0.089</i>
Indicator for bus stop located within 0.25 miles of count location (1 indicates present, 0 indicates not present)	0.214	0.418	<i>0.008</i>
Natural logarithm of parcel count	0.262	—	—
Indicator variable for posted speed limit \geq 40mph (1 indicates yes, 0 indicates no)	-0.636	-0.746	-0.381
Overdispersion parameter	1.550	2.435	1.226
Akaike Information Criterion (AIC)	17 889	17 857	

Values in italics are not significant at 95% confidence level.

alcohol sales locations are frequently accessed via walking due to the dangers associated with driving under the influence (Gayah et al., 2022).

- **Demographics.** People who reported that they commute by non-motorized means are vastly more likely to be observed in a pedestrian count, and therefore a greater proportion of non-motorized commuters is expected to increase pedestrian counts, as observed by the positive coefficient in the NB regression model (Griswold et al., 2019).

4.2 MBRP model

Under the framework discussed in the methodology section, a MBRP model splitting over the natural logarithm of parcel count was developed. Due to the relatively small dataset, the maximum depth of the tree

structure was limited to ensure that each of the sub-models was trained on a sufficiently large dataset. This also allowed for greater control over variable inclusion based on statistical significance and consistency with theory. The optimal split value of the parcel counts was found to be 71, with 1 197 of the training dataset observations registering parcel counts less than or equal to the optimal split value. Figure 2 depicts the tree structure for the MBRP model developed.

In comparing the estimation results between the NB regression model and the MBRP model, land use mix was found to be statistically insignificant in the MBRP model and was therefore removed from the model. The indicator variables for crosswalks and bus stops were found to be insignificant in the higher urban density model of the MBRP model, though the

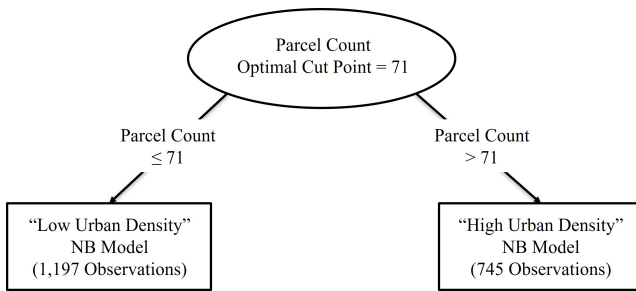


Figure 2 MBRP model tree structure

coefficient for bus stops was consistent with theory, and the contribution from the crosswalk variable was minor enough to overlook given the magnitude of its contribution in the lower urban density model. Aside from these differences, the NB and MBRP model coefficients take the same signs and indicate the same relationships between explanatory variables and pedestrian count outcomes.

The MBRP model allows for additional interpretation of model coefficients between the sub-models, which captures the difference in relationships between the explanatory features and the pedestrian count outcome as they vary with urban density. For example, based on differences in the magnitudes of coefficients presented in the second and third columns of Table 4, we observe that pedestrian counts are more sensitive to the presence of alcohol sales locations in denser urban environments. On the other hand, pedestrians in denser urban environments are less sensitive to the presence of sidewalks, crosswalks, bus stops, and greater speed limits than their counterparts in less dense urban areas. Across all models, the coefficients estimated for the proportion of non-motorized commuters and the natural logarithm of AADT are approximately equal, which indicates that pedestrians everywhere are about as sensitive to both variables, regardless of urban density.

4.3 Model comparison

Following model development, both models' cumulative residuals were evaluated over the testing dataset. The results were plotted on a CURE plot (Figure 3) with the 95% confidence interval from the NB regression plotted to demonstrate how the MBRP model compares to the 'baseline' NB regression model. Overall, we can tell that both models are robust, plotting nearly all observations within the confidence interval. From the CURE plot, we see that for lower count values predicted, the cumulative residuals are about equal across both models. Both models appear

to underpredict observations of approximately 200 pedestrians, though the residuals for the MBRP model are smaller. The trend of smaller residuals can be observed over most of the predicted range for the MBRP model, indicating a generally better fit for the data than the NB regression model.

The models were also tested for goodness of fit by standard error measurements to determine how well the models predict the pedestrian counts relative to the observed pedestrian counts in the test dataset. Table 5 presents the goodness of fit statistics for both models, where the better performance measure is bolded. To better depict the differences between the models, the goodness of fit statistics are presented for arbitrarily determined low-, mid-, and high-count ranges, as well as over the entire test dataset. Due to the nature of MAPE and the presence of zero-count observations, MAPE could not be applied to the low-count range, nor the full test dataset. In other studies, MAPE has been applied to zero-count locations through either log-transformation or by adding a negligibly small value to zero-counts. In this case, neither adjustment was applied to avoid artificially inflating the percentage error of zero-counts, for which any predicted value other than zero would incur immense errors.

Based on the AIC presented in Table 4, we find that the MBRP model is an overall improvement over the traditional NB regression model. But the statistics presented in Table 5 allow for more distinct quantitative results. Over the whole test dataset, the MAE and RMSE both indicate that the MBRP model predicts more accurately than the NB regression model by about 8.0% and 4.4% respectively. This trend is consistently observed across the low- and high-count ranges, where MAE improves by 14.0% and 7.2%, respectively, and the RMSE is improved by 22.2% and 6.3%, respectively. The exception to the trend is in the mid-count range, where the MAE improves by 3.9%, but the RMSE increases by 17.1%. This anomaly can be attributed to a single predicted value with a large residual, which is then highly weighted in the RMSE calculation due to the squared component of the error statistic. The evaluation that the MBRP model outperforms the NB regression model even in the mid-count range is further supported by the 9.0% improvement in the MAPE, which is also observed in the high-count range.

Accounting for evidence provided by the AIC as found in model development, the CURE plots generated

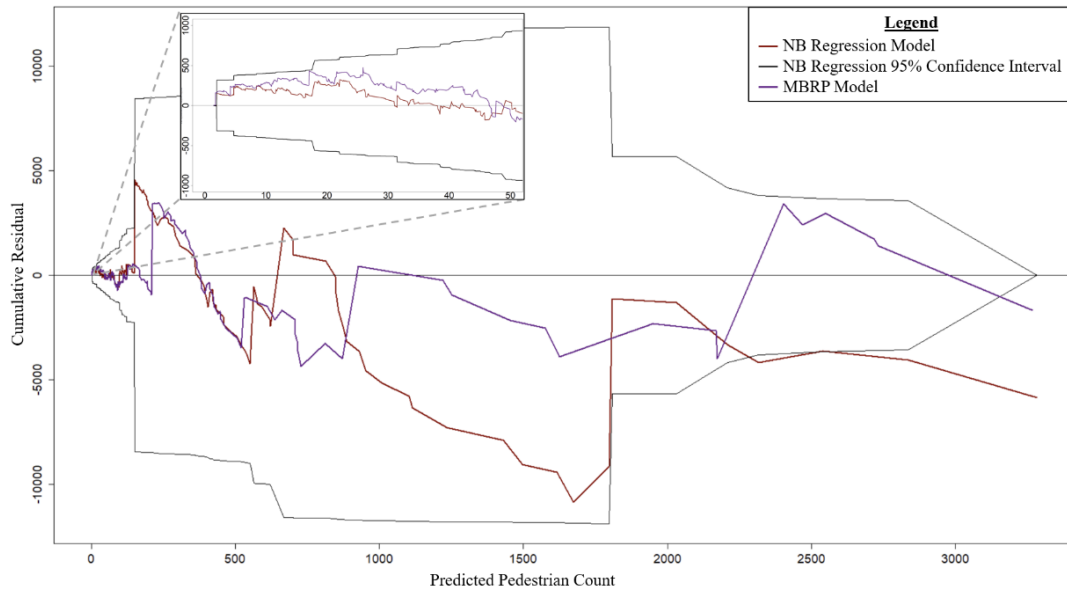


Figure 3 Test data cumulative residual plots

Table 5 Goodness of fit statistics

	Low-Count (0–100) (393 observations)		Mid-Count (101–1000) (82 observations)		High-Count (1000+) (13 observations)		Full Test Dataset (488 observations)	
	NB	MBRP	NB	MBRP	NB	MBRP	NB	MBRP
MAE	47.40	40.78	256.10	245.99	2116.33	1964.48	137.58	126.51
MAPE	n/a	n/a	118%	109%	61%	55%	n/a	n/a
RMSE	125.24	97.49	409.42	479.26	3058.23	2865.33	538.47	514.75

Bolded values indicate best performing model.

for both models, and the goodness of fit statistics as calculated over the test dataset, the MBRP model improves upon the traditional NB regression model in predicting pedestrian count values. Though error evaluations differ in quantity, based on trends in the MAE, RMSE, and MAPE indicate that using MBRP to model pedestrian count data can improve prediction accuracy by approximately 10% compared to traditional NB regression models.

5 Conclusions

This study investigates a possible improvement in pedestrian exposure modeling, by augmenting traditional NB regression models with a tree-based machine learning component, the MBRP algorithm. It was hypothesized that the inclusion of the MBRP algorithm would not only improve predictive accuracy but would allow for additional interpretation of the relationships between explanatory variables and pedestrian exposure estimates which may differ contextually in ways that are not known a priori. Data

from North Carolina is subdivided into training and test dataset. Training data is applied to develop a traditional NB regression model and an MBRP model, and the two models’ predictive performance is assessed based on the predictions for the test dataset.

The use of the MBRP algorithm resulted a model indicating that a pedestrian count’s relationship with explanatory variables varies with the natural logarithm of parcel count, which suggests that pedestrian counts in denser urban environments are more sensitive to presence of alcohol sales locations, and less sensitive to presence of sidewalks, crosswalks, bus stops, and roadways with higher speed limits. These trends are not observable through traditional NB models, which are typically more of a ‘one-size-fits-all approach which does not consider how the relationships between explanatory variables may vary with context.

The model developed using the MBRP algorithm improved predictions of pedestrian counts over the test dataset by approximately 10% as measured by MAE, MAPE, and RMSE. The model’s fitness was also

shown to be an improvement over the NB regression model through lower AIC and the MBRP plotting more points within the confidence interval and nearer to zero on a CURE plot.

The proposed use of the MBRP algorithm to improve pedestrian exposure estimates overcomes the shortcomings of traditional NB regression methods by capturing relationships between explanatory variables in context. Fortunately, the methodology is quite transferable to other regions outside of North Carolina, given that the MBRP algorithm inherently contextualizes coefficient estimates and naturally defines its own optimal splitting points. However, given the difficulty of retaining statistically significant variables across sub-models in the MBRP model, it should be noted that the variables selected for the MBRP model presented in this paper may not be statistically significant in another region. Further optimization may be required for application elsewhere. The MBRP algorithm and other machine learning methods are also known to be ‘data hungry’, and therefore the predictive power of this model may be limited by the size of the available dataset. In practice, a larger dataset would yield better results and could allow for a deeper understanding of the relationships between explanatory variables, which may be more complex than the restricted model presented in this paper suggests.

CRediT contribution statement

Jakob C. Wiegand: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing. **Vikash V. Gayah:** Conceptualization, Data curation, Methodology, Supervision, Validation, Writing—review & editing.

Declaration of competing interests

The authors declare no competing interests.

Funding

No external funding was used in this research.

Acknowledgements

The data used for this project comes from a previous project supported by the North Carolina Department of Transportation which evaluated systemic risk factors

in pedestrian safety outcomes. The authors would like to thank Daniel Carter and Brian Mayhew of NCDOT for providing the data, Ian Hamilton, Lauren Blackburn of VHB, and Ilgin Guler of Penn State for their collaboration.

An earlier version of this work was presented at the 9th Road Safety and Simulation conference, held in Lexington, KY, USA, on 28–31 October 2024.

Ethics statement

The research performed in this paper does not qualify as human subjects research and thus does not require any IRB protocol.

Declaration of generative AI use in writing

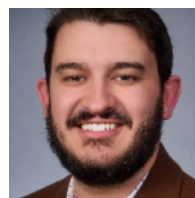
During the preparation of this work the authors used GPT-4 to check writing for grammatical errors and improve clarity of language in the final manuscript. The outputs were reviewed and revised by the authors who take full responsibility for the content of the publication.

References

- Behnam, J., B. G. Patel (1977), ‘Pedestrian volume estimation by land-use variables’, *Transportation Engineering Journal of ASCE*, 103(4), 507–520, <https://doi.org/10.1061/TPEJAN.0000649>.
- Frank, L. D., M. A. Andresen, T. L. Schmid (2004), ‘Obesity relationships with community design, physical activity, and time spent in cars’, *American Journal of Preventative Medicine*, 27(2), 87–96, <https://doi.org/10.1016/j.amepre.2004.04.011>.
- Gayah, V. V., S. I. Guler, H. Liu, L. Blackburn, I. Hamilton (2022), ‘Quantification of Systemic Risk Factors for Pedestrian Safety on North Carolina’, FHWA, NCDOT Project 2022-11, <https://connect.ncdot.gov/projects/research/Pages/ProjDetails.aspx?ProjectID=2022-11>.
- Griswold, J. B., A. Medury, R. J. Schneider, D. Amos, A. Li, O. Grembek (2019), ‘A pedestrian exposure model for the California State highway system’, *Transportation Research Record*, 2673(4), 941–950, <https://doi.org/10.1177/0361198119837235>.
- Hankey, S., G. Lindsey (2016), ‘Facility-demand models of peak period pedestrian and bicycle traffic comparison of fully specified and reduced-form models’, *Transportation Research Record*, 2586, 48–58, <https://doi.org/10.3141/2586-06>.
- Hankey, S., G. Lindsey, X. Wang, J. Borah, K. Hoff, B. Utecht, Z. Xu (2012), ‘Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN’, *Landscape and*

- Urban Planning*, 107(3), 307–316, <https://doi.org/10.1016/j.landurbplan.2012.06.005>.
- Hankey, S., T. Lu, A. Mondschein, R. Buehler (2017), ‘Spatial models of active travel in small communities: Merging the goals of traffic monitoring and direct-demand modeling’, *Journal of Transport & Health*, 7, 149–159, <https://doi.org/10.1016/j.jth.2017.08.009>.
- Hauer, E. (2015), *The Art of Regression Modeling in Road Safety* (New York: Springer), <https://doi.org/10.1007/978-3-319-12529-9>.
- Haynes, M., S. Andrzejewski, C. Saghir, L. yang Feng (2010), ‘GIS based bicycle & pedestrian demand forecasting techniques’, U.S. Department of Transportation, TMIP webinar, <https://www.fhwa.dot.gov/planning/tmip/community/webinars/summaries/20100429/index.cfm>.
- Kashani, A. T., A. S. Mohaymany (2011), ‘Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models’, *Safety Science*, 49(10), 1314–1320, <https://doi.org/10.1016/j.ssci.2011.04.019>.
- Lagerwey, P. A., M. J. Hintze, J. B. Elliott, J. L. Toole, R. J. Schneider (2015), ‘Pedestrian and bicycle transportation along existing roadway—ActiveTrans priority tool guidebook’, National Cooperative Highway Research Program, Report 803, <https://doi.org/10.17226/22163>.
- Lee, J., M. Abdel-Aty, I. Shah (2019), ‘Evaluation of surrogate measures for pedestrian trips at intersections and crash modeling’, *Accident Analysis & Prevention*, 130, 91–98, <https://doi.org/10.1016/j.aap.2018.05.015>.
- Lee, K., I. Sener (2020), ‘Emerging data for pedestrian and bicycle monitoring: sources and applications’, *Transportation Research Interdisciplinary Perspectives*, 4, <https://doi.org/10.1016/j.trip.2020.100095>.
- Lindsey, G., Y. Han, J. Wilson, J. Yang (2006), ‘Neighborhood correlates of urban trail use’, *Journal of Physical Activity and Health*, 3, 139–157, <https://doi.org/10.1123/jpah.3.s1.s139>.
- Lindsey, G., J. Wilson, E. Rubchinskaya, J. Yang, Y. Han (2007), ‘Estimating urban trail traffic: Methods for existing and proposed trails’, *Landscape and Urban Planning*, 81(4), 299–315, <https://doi.org/10.1016/j.landurbplan.2007.01.004>.
- Liu, X., J. Griswold (2009), ‘Pedestrian volume modeling: A case study of San Francisco’, *Association of Pacific Coast Geographers*, 71(1), 164–181, <https://doi.org/10.1353/pcg.0.0030>.
- Lu, T., A. Mondschein, R. Buehler, S. Hankey (2018), ‘Adding temporal information to direct-demand models: Hourly estimation of bicycle and pedestrian traffic in Blacksburg, VA’, *Transportation Research Part D: Transport and Environment*, 63, 244–260, <https://doi.org/10.1016/j.trd.2018.05.011>.
- Miranda-Moreno, L. F., D. Fernandes (2011), ‘Modeling of pedestrian activity at signalized intersections land use, urban form, weather, and spatiotemporal patterns’, *Transportation Research Record*, 2264(1), 74–82, <https://doi.org/10.3141/2264-09>.
- NHTSA (2023a), ‘FARS Encyclopedia 2023’, National Highway Traffic Safety Administration, <https://www-fars.nhtsa.dot.gov/Main/index.aspx>.
- NHTSA (2023b), ‘Traffic Safety Facts, 2021 Data: Pedestrians’, National Highway Traffic Safety Administration, DOT HS 813 458, <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813458>.
- Pulugurtha, S. S., S. R. Repaka (2008), ‘Assessment of models to measure pedestrian activity at signalized intersections’, *Transportation Research Record*, 2073(1), 39–48, <https://doi.org/10.3141/2073-05>.
- Schneider, R. J., L. S. Arnold, D. R. Ragland (2009), ‘Pilot model for estimating pedestrian intersection crossing volumes’, *Transportation Research Record*, 2140(1), 13–26, <https://doi.org/10.3141/2140-02>.
- Schneider, R. J., T. Henry, M. F. Mitman, L. Stonehill, J. Koehler (2012), ‘Development and application of volume model for pedestrian intersections in San Francisco, California’, *Transportation Research Record*, 2299(1), 65–78, <https://doi.org/10.3141/2299-08>.
- Seibold, H., A. Zeileis, T. Hothorn (2016), ‘Model-based recursive partitioning for subgroup analyses’, *The International Journal of Biostatistics*, 12(1), 45–63, <https://doi.org/10.1515/ijb-2015-0032>.
- Tang, H., E. T. Donnell (2019), ‘Application of a model-based recursive partitioning algorithm to predict crash frequency’, *Accident Analysis & Prevention*, 132, 105274, <https://doi.org/10.1016/j.aap.2019.105274>.
- Zeileis, A., K. Hornik (2007), ‘Generalized M-Fluctuation tests for parameter instability’, *Statistica Neerlandica*, 61, 488–508, <https://doi.org/10.1111/j.1467-9574.2007.00371.x>.
- Zeileis, A., T. Hothorn, K. Hornik (2008), ‘Model-based recursive partitioning’, *Journal of Computational and Graphical Statistics*, 17, 492–514, <https://doi.org/10.1198/106186008X319331>.

About the authors



Jakob C. Wiegand received his BS in Civil Engineering from Valparaiso University in 2022 and is currently pursuing a Ph.D. in Transportation Engineering at The Pennsylvania State University. His research interests broadly cover transportation safety, but primarily focus on safety of vulnerable roadway users – especially pedestrians. Jakob’s recent research aims to emphasize the importance of accurate exposure estimates in crash prediction modeling and equity of protections for pedestrians.



Vikash V. Gayah is a professor in the Department of Civil and Environmental Engineering at The Pennsylvania State University, where he also serves as the Interim Director of the Larson

Transportation Institute. He received his B.S. and M.S. degrees from the University of Central Florida and his Ph.D. degree from the University of California, Berkeley. Dr. Gayah's research focuses on urban mobility, traffic operations, traffic flow theory, traffic safety and non-motorized transportation. Dr. Gayah currently serves as an editorial advisory board member of *Transportation Research Part C: Emerging Technologies and Accident Analysis and Prevention*, an editorial board editor of *Transportation Research Part B: Methodological*, an associate editor for *Transportation Letters* and the *IEEE Intelligent Transportation Systems Magazine* (an international peer-reviewed journal), and a handling editor for the *Transportation Research Record*.



All contents are licensed under the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).