Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 11(121)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2024.111100198

Advancing urban transportation planning: a comparative analysis of machine learning and conventional regression-based trip generation models

Anish Kumar[1], Amit Kumar[2, 3*] and Sanjeev Sinha[4]
Assistant Professor, Department of Civil Engineering, Rajkiya Engineering College, Azamgarh- 276201, Uttar Pradesh, India[1]
Research Scholar, Department of Civil Engineering, Dr. A.P.J. Abdul Kalam Technical University, Lucknow-226031, Uttar Pradesh, India[2]
Assistant Professor, Department of Civil Engineering, Rajkiya Engineering College, Azamgarh- 276201, Uttar Pradesh, India[3]
Professor, Department of Civil Engineering, National Institute of Technology, Patna- 800005, Bihar, India[4]

Received: 22-February-2024; Revised: 21-December-2024; Accepted: 22-December-2024

Abstract
Trip generation models are traditionally developed using linear regression. This study focuses on enhancing trip generation models by exploring the potential of support vector machine (SVM) based models. The primary objective is to improve the accuracy and efficiency of travel demand estimation for transportation planning. To achieve this, SVM models were constructed using the radial basis function (RBF) and linear kernel functions and then compared to conventional linear regression models. The performance of these models was evaluated to determine their effectiveness in predicting trip production and attraction. The results revealed that the SVM model with the RBF kernel outperformed both the SVM model with the linear kernel and the traditional linear regression model. Specifically, the R2 for the trip production and trip attraction predictions using the RBF kernel were 1 and 0.95, respectively. Additionally, the Kruskal Wallis (KW) test was employed to assess the statistical significance of the differences between the models. The findings suggest that SVM models, particularly those utilizing the RBF kernel, offer a superior alternative to conventional regression models for trip generation, providing transportation planners with a more efficient tool for travel demand estimation.

Keywords
Trip generation, Support vector machine, Radial basis function kernel, Linear regression, Travel demand estimation, Transportation planning.

1.Introduction
Urbanization is a global phenomenon that continuously reshapes the dynamics of the world [1]. By 2050, the United Nations predicts that almost two-thirds of the world's population will reside in cities [2]. The effective and sustainable management of transportation systems is a crucial factor as cities continue to expand and change [3]. Accurate trip generation prediction is crucial for urban transportation planning to address rising travel demands, reduce congestion, minimize environmental impacts, and enhance overall quality of life [4].

Trip generation models are essential in transportation planning, forming the foundation for estimating travel demand based on various socio-economic and land-use factors.

Traditionally, these models have been developed using linear regression techniques, which rely on a linear relationship between dependent and independent variables. While linear regression models are straightforward to implement and interpret, they often fail to capture the complexities and nonlinearities inherent in real-world travel behavior. As transportation systems become more intricate and data availability increases, there is a growing need for more sophisticated modeling techniques that offer greater predictive accuracy and model robustness. Trip generation models have

*Author for correspondence

historically used conventional regression-based techniques [5–10] to estimate trip generating trends using historical data and predefined socio-economic and land-use associations [11]. There are some challenges associated with the use of linear regression in trip generation modelling. Linear regression assumes a linear relationship between the predictors and the outcome variable. This assumption is often unrealistic in the context of travel behavior, where relationships can be nonlinear and influenced by numerous interacting factors. The predictive performance of linear regression models may be inadequate, especially in complex urban environments where travel patterns are influenced by a wide range of dynamic factors. Modern transportation datasets are large, heterogeneous, and often contain nonlinear relationships that traditional linear models struggle to capture effectively. Linear models can be prone to overfitting or underfitting due to the complexity of the data. Urban planning has found these models to be helpful in the design of transportation infrastructure, traffic control, and urban development. However, conventional models frequently find it difficult to produce accurate estimates in the face of increased urbanization, changing mobility patterns, and socioeconomic conditions [12–15]. Urban transportation planning is one of the many fields where machine learning has recently become a disruptive force [16, 17]. By utilizing the power of artificial intelligence, big data, and data-driven insights, machine learning-based models have the potential to transform the way trip generation is estimated. Machine learning algorithms provide accurate, adaptable, and flexible predictions by autonomously uncovering complex relationships within large datasets and adjusting to evolving urban dynamics [18]. The transition from traditional regression-based trip generating models to the cutting-edge world of machine learning-based models is thoroughly explored in this research article. One of the primary objectives of this research is to assess how machine learning-based trip generation models compare to their conventional regression-based counterparts in terms of predictive accuracy. Accurate trip generation estimations are fundamental to the planning and design of urban transportation systems, as they inform infrastructure investments, traffic management strategies, and overall urban development. Machine learning models have the potential to improve accuracy by extracting complex patterns and relationships from large, multidimensional datasets. Machine learning models can uncover subtle correlations that may not be evident to human analysts, unlike conventional regression models that rely on predetermined variables and linear relationships [19]. This adaptability to diverse and evolving urban contexts is particularly crucial for mid-sized cities like Patna in India, where the interplay of various factors significantly impacts trip generation. In contrast, traditional regression-based models typically rely on static relationships [20–24], which may fail to adequately reflect the dynamic nature of a city's evolving transportation patterns. Inaccurate trip generation estimates can result in either underinvestment or overinvestment in transportation infrastructure, leading to substantial economic, social, and environmental repercussions [25]. By evaluating the predictive accuracy of machine learning models compared to conventional regression models, this research aims to highlight the potential advantages of adopting more advanced, data-driven modeling approaches in urban transportation planning. A critical consideration in evaluating trip generation models is their adaptability to evolving urban dynamics and emerging transportation trends. Cities are dynamic entities, subject to changes in land use, demographic shifts, technological advancements, and transportation preferences. Conventional regression models are often rigid, relying on predefined relationships and variables that may not account for unforeseen changes. On the other hand, machine learning models are renowned for their adaptability [26]. They can continuously learn and update their predictions based on new data, allowing them to adjust to changing urban contexts. This adaptability is particularly relevant for mid-sized cities like Patna, where urban development may be less predictable than in larger, more established urban centers. Machine learning models can handle variables that conventional models may not consider or that may not have been prevalent at the time of the model's development [27]. For example, the rapid rise of ride-sharing services, e-commerce, and electric scooters in urban transportation underscores the need for models that can adapt to these emerging trends. Machine learning models have the potential to integrate new variables and discover nuanced relationships that conventional models may overlook. Assessing the adaptability of machine learning-based trip generation models can provide valuable insights into their practicality in a rapidly evolving urban environment and inform decision-making for more flexible urban transportation planning. The understanding of advantages and limitations of machine learning-based trip generation models facilitates the policymakers to make informed decisions about adopting new modeling approaches.

This may involve adjustments in data collection, computational infrastructure, and regulatory frameworks to accommodate the changing needs of urban transportation planning.

This research seeks to conduct a comprehensive analysis of the shift from traditional regression-based trip generation models to machine learning-based support vector machine (SVM) models within the context of urban transportation planning. By comparing the accuracy, adaptability, scalability, and policy implications of these two modeling approaches, this study aims to contribute to the ongoing discourse in urban planning and provide insights into the potential benefits of adopting more sophisticated, data-driven models. As mid-sized cities like Patna navigate the complexities of urbanization, understanding the capabilities and limitations of these models is essential to making informed decisions that enhance the efficiency, sustainability, and adaptability of urban transportation systems. The primary objective of this study is to enhance trip generation models by exploring the potential of SVM based models. SVMs are a class of machine learning algorithms known for their robustness and ability to handle nonlinear relationships through the use of kernel functions. This study aims to construct SVM-based trip generation models using both radial basis function (RBF) and linear kernel functions. The performance of the SVM models were compared with conventional linear regression models to determine their effectiveness in predicting trip production and attraction. The accuracy and efficiency of the models were assessed using statistical measures such as R-squared (R^2) values and perform the Kruskal Wallis (KW) test to evaluate the statistical significance of the differences between the models.

This paper is structured as follows: Section 2 presents the literature review, Section 3 outlines the methods, Section 4 discusses the results and findings, and Section 5 provides the conclusion.

2.Literature review

Trip generation is the first step of the conventional four stage travel demand model [28, 29]. The first step involves estimating the total number of trips produced and attracted to each traffic analysis zones (TAZ) [30]. The number of persons expected to use the development would be estimated in this stage by using regression method to derive the relation between the trips produced and trip attracted. *Table 1* compiles a few studies that have been undertaken in the area of transport demand modelling. The dependent and independent variables in each of the cases are shown in *Table 2*.

Table 1 Studies related to transport demand modelling

Focus area	Methodology/Model	Key findings	Recommendations	Citation
Transport demand projection for New Delhi	Four-stage planning process, least-square regression models	Found road systems insufficient for future traffic.	Recommended mass rapid transit systems (MRTS) and high-capacity busways.	[31]
Trip generation for small urban communities	Single internal trip purpose for TAZ	No significant difference in using multi-purpose modeling approaches.	Simplified models can be used for smaller communities.	[32]
Home-based trip generation for Thiruvananthapuram	Mathematical modeling of trip generation	Trip generation rate highly dependent on employment status of individuals.	Consider employment status when planning for trip generation.	[33]
Public transport demand in Bangalore	Travel demand modeling	Public bus services remain crucial even with the introduction of MRTS. Bus system projected to handle 52% of demand.	Expand and improve public bus services, integrating land-use planning.	[34]
Mode choice models	Hybrid choice model	High-quality service data is scarce in developing countries, requiring alternate model frameworks.	Use hybrid models for policy planning in data-scarce environments.	[35]
Bus rapid transit system (BRTS) in Indian cities	Review of existing BRTS systems	Public transport lack's reliability, comfort,	Improve public transport systems to reverse the	[36]

Focus area	Methodology/Model	Key findings	Recommendations	Citation
		safety, and security, pushing people towards private car usage.	trend.	
Trip distribution in Alexandria	Gravity model	Shift observed toward non-compulsory trip purposes.	Use gravity models to analyze diverse trip patterns.	[37]
Origin-destination (OD) estimation using location based social networking (LBSN)	LBSN data for OD matrix	LBSN data reduces OD estimation errors caused by traditional gravity models.	Incorporate LBSN data into transport demand analysis.	[38]

Table 2 Variables used in trip generation model

Trip generation model	Independent variable	Independent variable
Trip production	Population	Production
Trip attraction	Employment	Attraction

Trip production: Trip production refers to the number of trips originating from a particular area, such as a household or a TAZ. This includes all types of trips generated by residents for various purposes such as work, education, shopping, and leisure.

Trip Attraction: Trip attraction refers to the number of trips ending in a particular area, such as workplaces, schools, shopping centers, or recreational sites. This variable captures the destinations that draw people to a specific location within a TAZ.

Theoretical background of SVM regression: SVM regression is a powerful machine learning technique used for both classification and regression tasks. In this brief introduction, SVM regression with two commonly used kernels: the RBF kernel and the linear kernel has been focused.

RBF kernel SVM regression: The RBF kernel, also known as the gaussian kernel, is a popular choice for SVM regression. It is particularly effective when dealing with non-linear and complex data. The RBF kernel maps the input data into a higher-dimensional space where it becomes more separable, making it suitable for capturing non-linear relationships in regression problems.For a given set of training data (x_i, y_i) for i = 1 to N, where x_i represents the input features, and y_i represents the corresponding target values, the SVM regression problem with an RBF kernel can be formulated as shown in Equation 1:

$$\text{Minimize: } 1/2 \times \|w\|^2 + C * \sum \xi_i \qquad (1)$$

$$\text{Subject to: } y_i - w\varphi(x_i) - b \leq \varepsilon + \xi_i;\ \xi_i \geq 0$$

Here, w represents the weight vector, $\varphi(x_i)$ is the transformed feature vector using the RBF kernel, b is the bias term, C is the regularization parameter, ε is the error tolerance, and ξ_i are slack variables.

Linear kernel SVM regression: The Linear kernel is a simpler option and is effective when the relationship between input features and target values is believed to be linear. It's computationally efficient and can be a good choice for linear regression tasks.

In the case of SVM regression with a linear kernel, the objective function to be minimized is shown in Equation 2:

$$\text{Minimize: } 1/2 \times \|w\|^2 + C \times \sum \xi_i \qquad (2)$$

$$\text{Subject to: } y_i - wx_i - b \leq \varepsilon + \xi_i;\ \xi_i \geq 0$$

Here, the terms have similar meanings as in the RBF kernel case. In the current study several performance metrices were employed to evaluate the model performance. The details of these metrices are given in *Table 3*.

Table 3 Performance metrics in the study

Performance metric	Mathematical expression
Mean absolute error (MAE)	$(1/n) \sum \lvert y_i - \hat{y}_i \rvert$
Mean squared error (MSE)	$(1/n) \sum (y_i - \hat{y}_i)^2$
Root mean squared error (RMSE)	$\sqrt{(MSE)}$
Mean squared logarithmic error (MSLE)	$(1/n) \sum(\log (1 + y_i) – \log (1 + \hat{y}_i))^2$
Root mean squared logarithmic error (RMSLE)	$\sqrt{(MSLE)}$
R^2	1 - (SSR/ SST)
Mean absolute percentage error (MAPE)	$(1/n) \sum \lvert (y_i - \hat{y}_i) / y_i \rvert \times 100$
Willmott's index of agreement (d)	$1 - [(\sum \lvert y_i - \hat{y}_i \rvert) / (\sum (\lvert y_i - \bar{\hat{y}} \rvert + \lvert \hat{y}_i - \bar{\hat{y}} \rvert))]$
Mielke and Berry Index (B)	$1 - (\lvert \sum (y_i - \hat{y}_i) \rvert / \sum (\lvert y_i - \hat{y}_i \rvert + \lvert y_i - \bar{\hat{y}} \rvert + \lvert \hat{y}_i - \bar{\hat{y}} \rvert))$

Note: y is the predicted, $\hat{y}$ is the observed and $\bar{\hat{y}}$ is the mean of the observed values.

3.Methods

Study area: The current study revolves around Patna urban agglomeration area (PUA). In order to understand the travel pattern of the city, a total of 100 TAZ (*Figure 1*) for Patna were identified based on administrative wards by considering the ease of getting required zonal information. Among these 100 TAZ's, 89 internal zones were identified within the PUA and 11 external zones (outside planning area). The internal zones within municipal boundary are 72 and the areas that fall under Patna Urban area but outside the municipal boundary have been divided into 17 zones based on homogenous land use and traffic generation points. These areas are Phulwari Sharif, Khagaul, Dinapur municipalities and Patliputra Housing Colony. The external zones are outside the planning area boundary by considering the important settlements and towns like Hajipur, Sonepur, Fatwah, Hilsa, Naubatpur, Bihta and Arrah which are given separate zones.

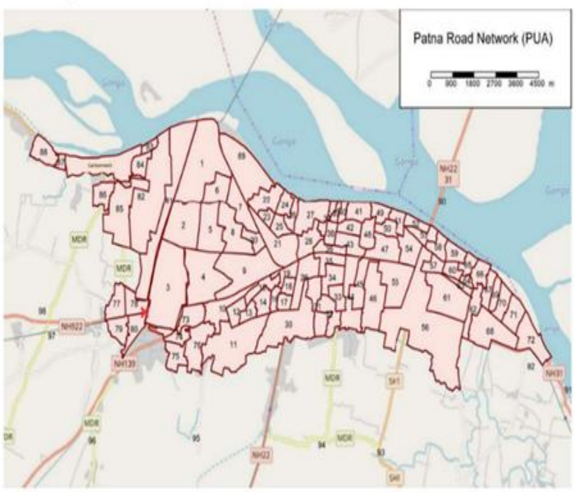

Figure 1 Traffic analysis zones (TAZ)

Data considerations: The base year for the current study was assumed to be 2018 and the predictions were made for the horizon years 2031 and 2041. The detailed zone wise data for population, production, employment, and attraction for the base year 2018 was obtained from the comprehensive mobility plan (CMP), 2018 of Patna. The data in CMP, 2018 related to population, production, employment, and attraction for each TAZ were sourced from two primary methods: census data and home-based interviews. The census data provided a comprehensive and reliable source for population and employment figures. Census data is typically collected by national statistical agencies through a systematic enumeration of the population, which includes detailed demographic, social, and economic information. This census data collection method ensures high accuracy and completeness, making it an ideal source for population and employment data. Specifically, the census data included information on the number of residents and the employment figures within each TAZ. These data points are crucial for understanding the distribution of people and jobs, which are fundamental for modeling trip generation. In addition to census data, home-based interviews were conducted to gather detailed information on trip production and attraction. Home-based interviews involve visiting households within each TAZ and asking residents about their travel behavior, including the number and purpose of trips made. This method provides granular insights into daily travel patterns that are not captured through the census. To ensure that the home-based interviews were representative of the population within each TAZ, a stratified random sampling technique was employed.

This technique involves dividing the population into subgroups (strata) based on certain characteristics, such as geographic location, household size, or socio-economic status, and then randomly selecting samples from each stratum. The rationale behind this approach was to capture a diverse and representative sample of households, ensuring that the data collected reflected the varying travel behaviors across different segments of the population. The combination of census data and home-based interviews was chosen to leverage the strengths of both methods. Census data provided a comprehensive overview of population and employment, ensuring that the foundational demographic and economic characteristics of each TAZ were accurately captured. Meanwhile, home-based interviews allowed for the collection of specific travel behavior data directly from residents, offering a detailed and context-specific understanding of trip production and attraction.

Stratified random sampling in home-based interviews was employed to ensure a representative sample, thereby improving the reliability and validity of the collected travel behavior data. The integration of census data and home-based interviews, along with stratified random sampling technique, provided a comprehensive and accurate dataset for modeling trip generation across different TAZs. This methodological approach ensured that the study could effectively capture the complex dynamics of travel behavior needed to develop robust trip generation models.

The projections for the horizon years 2031 and 2041 based on applicable growth rates, curve fitting, and regression analysis was also taken from the CMP of Patna, 2018. The data from the CMP, 2018 was further divided into training and testing database as shown in *Table 4* and *Table 5* respectively.

Table 4 Summary statistics of the training data

	Variable	Observations	Minimum	Maximum	Mean	Standard deviation
Trip production	Production	60	14000.000	121528.000	35405.717	17332.984
	Population	60	11633.000	106521.000	30521.033	15295.525
Trip attraction	Attraction	60	12133.000	113845.000	36343.217	17658.913
	Employment	60	3967.000	36581.000	10486.633	5380.716

Table 5 Summary statistics of the testing data

	Variable	Observations	Minimum	Maximum	Mean	Standard deviation
Trip production	Population	14	17578.000	46851.000	26949.857	8364.244
Trip attraction	Attraction	14	5525.000	17970.000	9013.714	3319.315

While developing a machine learning model, it is crucial that the training and testing datasets are representative of the same population. If they are not, the model might perform poorly in practice. The Kolmogorov Smirnov (KS) test helps in determining whether the data distributions in the training and testing sets are significantly different. If the training and testing data come from different populations, the model could potentially be overfit to the training data, leading to poor generalization on the testing data. The KS test was used to assess the normality of the data distributions.

This non-parametric test compares the sample distribution with a reference distribution (usually normal) to determine if they differ significantly. A significance level of 0.05 was chosen, meaning there is a 5% risk of concluding that a difference exists when there is none. This threshold is commonly used in statistical analyses to balance the risks of Type I and Type II errors. Choosing this significance level impacts the results by providing a standard measure to judge the normality of the data, influencing subsequent modeling choices and interpretations. In the current study, KS test was performed to address this issue.

The cumulative distribution of the training and testing dataset is shown in *Figure 2*. During the KS-test the null and alternate hypothesis were selected to be: H0: The two samples follow the same distribution, Ha: The distributions of the two samples are different. The computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0. This finding shows that the distribution of zonal data into training and testing dataset holds statistical significance. The results of the KS test are tabulated in *Table 6.*

Table 6 Two-sample KS test / Two-tailed test

Model	Trip production	Trip attraction
D	0.171	0.195
p-value (Two-tailed)	0.828	0.707
alpha	0.05	0.05
Null hypothesis (H_0)	Accepted	Accepted

Methodology adopted in the study: The methodology adopted in the study is shown in *Figure 3*. After acquiring population, production, employment, and attraction data from CMP, 2018, trip production and trip attraction models were prepared for the base year 2018. These models were developed using linear regression analysis. After the development of trip generation model based on linear regression analysis for the base year 2018, trip generation model for the horizon years 2031 and 2041 were developed using linear regression analysis.

In the next part of the study, the trip generation models developed for the horizon years 2031 and 2041 were developed using SVM. A comparative analysis was then made to understand the statistical significance and applicability of SVM based trip generation models.

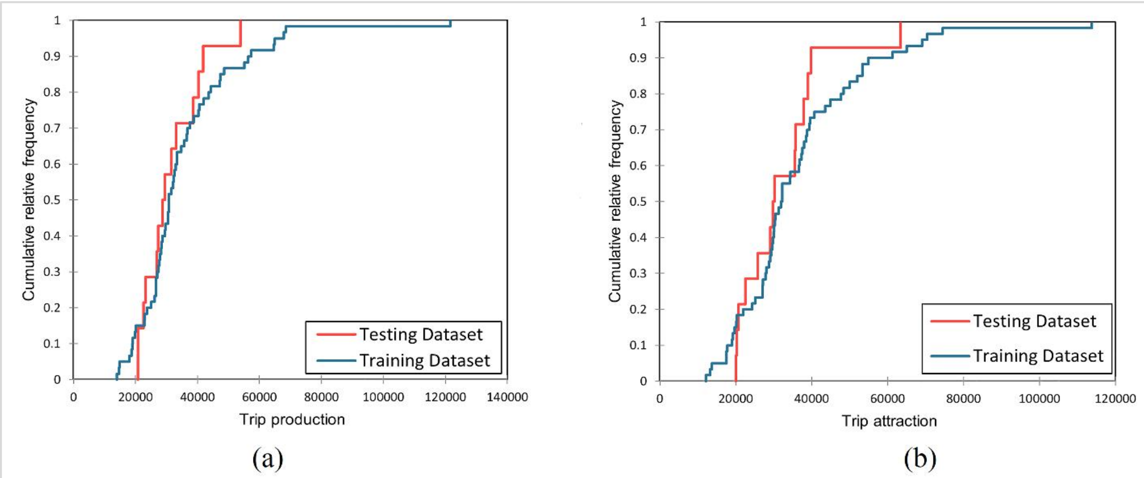

Figure 2 Cumulative distributions of the training and testing database of: (a) Trip production; (b) Trip attraction

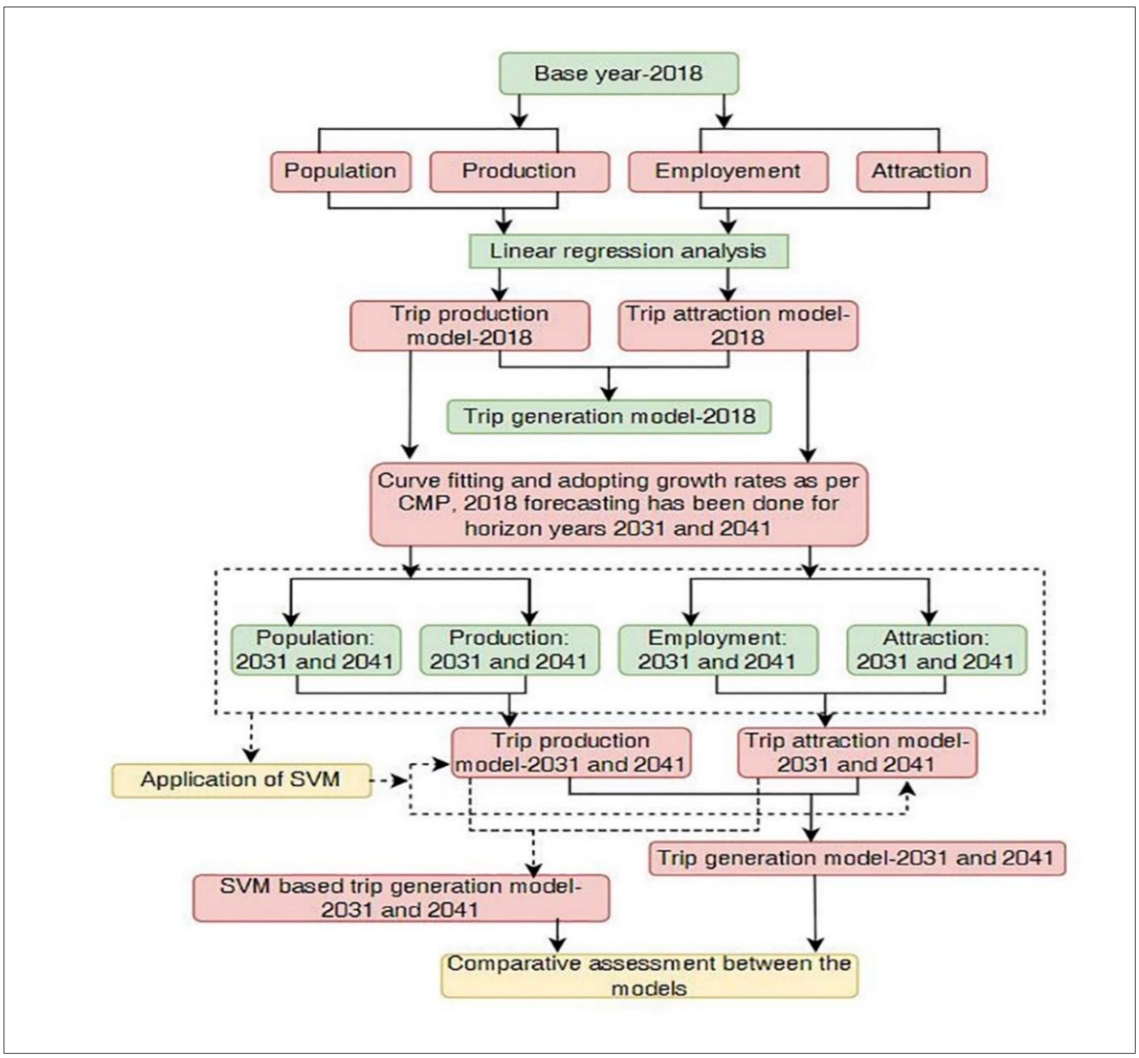

Figure 3 Methodology adopted in the current study

4.Results and discussion

Trip generation model for the base year 2018*:* The trip generation equations for PUA are generated based on linear regression. The trip production equation (Equation 3) for PUA is found to be statistically significant with R^2 value 1. The graphical representation of the trip production model is shown in *Figure 4 (a).*

$$\text{Trip Production (PUA)} = 1.1332 \times \text{Population} + 818.76 \quad (3)$$

The trip attraction equation (Equation 4) for PUA is found to be statistically significant with R^2 value 0.97. The graphical representation of the trip production model is shown in *Figure 4 (b).*

$$\text{Trip Attraction (PUA)} = 3.206 \times \text{employment} + 2820.65 \quad (4)$$

Trip generation model for the horizon years 2031 and 2041*:* Linear regression analysis was employed to develop the trip generation model for the horizon years 2031 and 2041. The data used in the model development was obtained from CMP, Patna. The trip production equation (Equation 5) for PUA is found to be statistically significant with R^2 value 0.70. The graphical representation of the trip production model is shown in *Figure 5 (a).*

$$\text{Trip Production (PUA)} = 1.0755 \times \text{Population} + 4797.326 \quad (5)$$

The trip attraction equation (Equation 6) for PUA is found to be statistically significant with R^2 value 0.99. The graphical representation of the trip production model is shown in *Figure 5 (b).*

$$\text{Trip Attraction (PUA)} = 3.165 \times \text{employment} + 2329.3063 \quad (6)$$

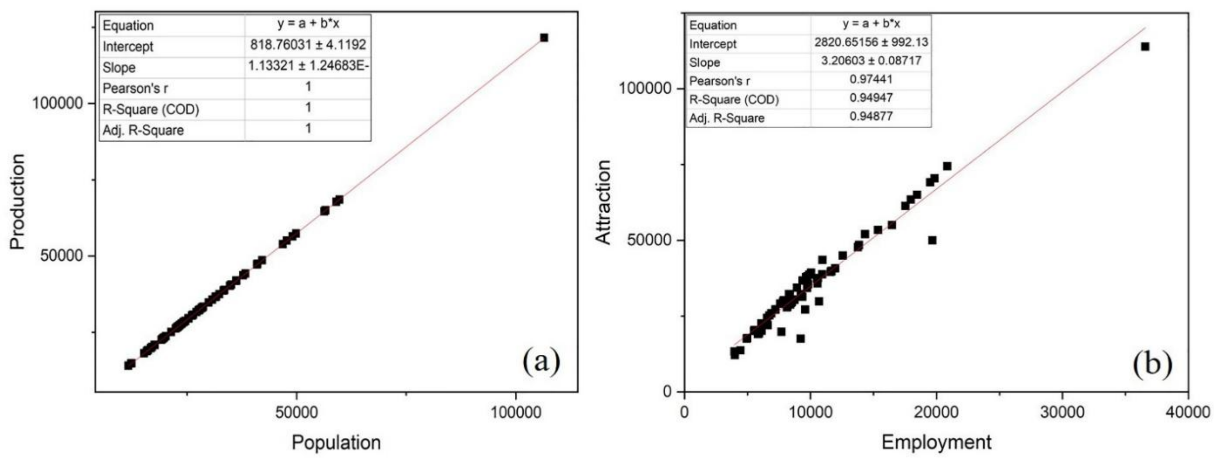

Figure 4 Correlation plot of: (a) Trip production model; (b) Trip attraction model for base year

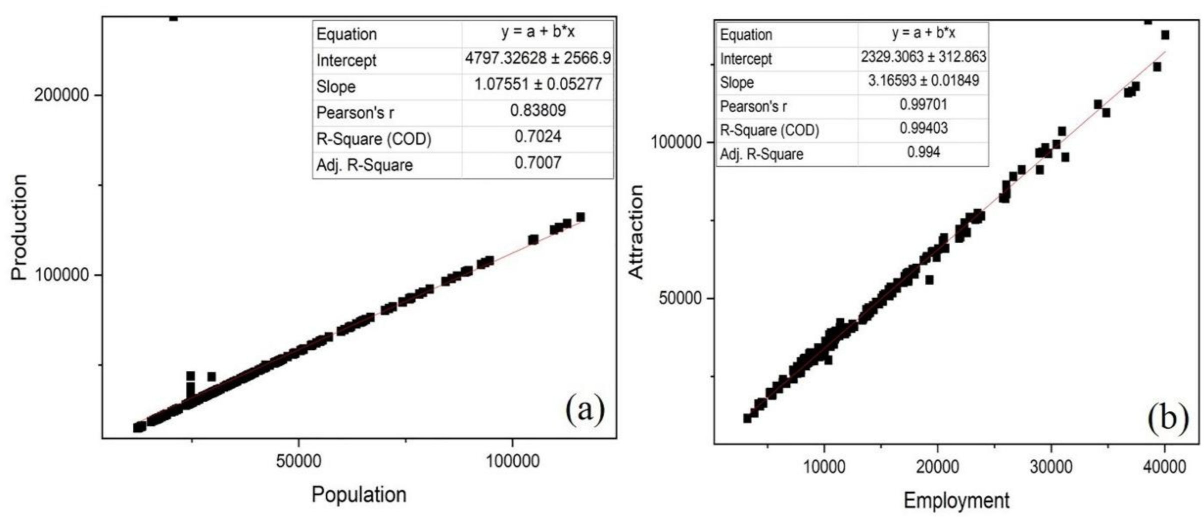

Figure 5 Correlation plot of: (a) Trip production model; (b) Trip attraction model for horizon years

SVM based trip generation model for the horizon years 2031 and 2041*:* RBF kernel and linear kernel based SVM models were developed in the current study. The data from 74 zones were divided into training and testing datasets.

K-fold cross-validation is a crucial technique in machine learning that offers numerous benefits. It provides a robust and comprehensive method for assessing and validating the performance of predictive models. In this technique, the dataset is divided into K subsets, or "folds," and each fold is systematically used as testing data while the remaining folds serve as training data. This process, known as K-fold cross-validation, helps mitigate the risk of biased model evaluation. It reduces the variance in performance estimates by averaging results from multiple iterations, making assessments more stable and reliable.

This technique maximizes data utilization, ensuring all data points contribute to model evaluation, and helps in model selection and hyperparameter tuning. It reveals insights into a model's generalization capabilities, detects overfitting, and allows for a deeper analysis of the bias-variance trade-off. K-fold cross-validation is instrumental in assessing model stability, handling imbalanced datasets, and identifying data quality issues. In the current study, a 10-fold cross validation was done, and the results are tabulated in *Table 7*. In both trip production and trip attraction models, the linear regression model provides more consistent and reliable performance across all folds compared to the RBF SVM model. The linear model's simplicity leads to stable and robust predictions, as indicated by consistently low error metrics and high R^2 values.

The RBF model, despite its capacity to capture non-linear relationships, exhibits significant variability and occasional high errors, likely due to overfitting. This emphasizes the importance of model stability and generalizability, particularly in practical applications like urban planning and transportation forecasting. Despite the variability observed in the RBF model's performance, it remains a viable option for modeling trip production and attraction due to its ability to handle complex, non-linear relationships inherent in urban systems. The higher variability in the RBF model's results suggests suboptimal hyperparameter settings, which can be addressed through careful tuning, such as optimizing gamma and C parameters, and incorporating robust cross-validation and regularization techniques to enhance generalization. Certain folds exhibit comparable or superior performance in R^2 values and error metrics for the RBF model, indicating its potential for high predictive accuracy when appropriately configured. Furthermore, the RBF model's flexibility makes it adaptable to future changes in urban data patterns, providing a robust long-term modeling solution. Mitigating high variability can be achieved by conducting thorough hyperparameter optimization, applying regularization, utilizing ensemble methods, and increasing the training dataset's size and diversity. While the linear regression model shows more consistent performance across folds, the RBF model's ability to capture non-linear patterns and its potential for improvement through careful tuning and validation make it a strong candidate for trip production and attraction modeling, capable of providing more accurate and adaptable predictions over time.

The hyperparameters of the SVM-based models for trip production and attraction, optimized through hyperparameter tuning, are presented in *Table 8*. The performance metrics of trip production and attraction models are given in *Table 9* and *Table 10* respectively. As per the results tabulated in *Table 9* and *Table 10*, SVM models with RBF kernel function performed better than the SVM model with linear kernel function in both trip production and attraction by a small margin.

Table 7 10-fold cross validation of different models

Model	**Trip production**						**Trip attraction**					
Kernel	**RBF**			**Linear**			**RBF**			**Linear**		
K	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**
1	0.24	0.99	0.43	0.10	1.00	0.26	3.49	0.94	1.73	2.04	0.98	1.29
2	0.08	1.00	0.25	0.15	0.99	0.34	2.18	0.16	1.05	33.88	0.95	3.19
3	0.16	0.99	0.37	0.14	0.99	0.35	48.54	0.28	4.65	15.75	0.85	3.37
4	813.05	0.21	10.40	0.15	0.99	0.36	727.30	0.19	12.74	7.84	0.95	2.22
5	983.29	0.19	31.36	0.16	0.99	0.40	3.72	0.98	1.72	3.56	0.98	1.49
6	0.11	0.99	0.30	0.08	1.00	0.24	2.91	0.97	1.34	2.34	0.96	1.36

Model	**Trip production**						**Trip attraction**					
Kernel	**RBF**			**Linear**			**RBF**			**Linear**		
K	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**	**MSE ($x10^6$)**	**R^2**	**MAE ($x10^3$)**
7	0.17	0.99	0.37	0.11	0.99	0.33	14.17	0.92	2.66	9.64	0.97	2.12
8	0.18	0.99	0.40	0.16	0.99	0.38	3.06	0.98	1.44	76.94	0.41	5.41
9	0.18	0.99	0.36	0.30	1.00	0.35	2.06	0.97	1.22	0.71	0.99	0.72
10	0.14	0.99	0.33	0.12	0.99	0.34	3.63	0.98	1.80	3.53	0.98	1.44
Mean	179.76	0.83	4.457	0.147	0.99	0.335	81.106	0.74	3.035	15.623	0.90	2.26

Table 8 Parameters of SVM models trained in the study

Model	**Kernel**	**C**	**ε**	**Tolerance**	**Gamma**	**Bias**	**Number of support vectors**
Trip production	RBF	100	0.03	0.001	0.5	1.685	7
	Linear	100	0.03	0.001	-	0.018	2
Trip attraction	RBF	100	0.03	0.001	0.5	1.513	52
	Linear	100	0.03	0.001	-	0.030	47

The comparative performance of linear regression and SVM models with RBF kernels demonstrates that SVM models generally outperform linear regression in predicting both trip production and trip attraction. In *Table 9*, the RBF model shows lower MAE, MSE, and RMSE values and higher R^2 values compared to the linear model, indicating superior accuracy and predictive capability. Similarly, in *Table 9*, the RBF model again surpasses the linear model with significantly lower MAE, MSE, and RMSE values, as well as higher R^2 values, showcasing its effectiveness in handling complex data patterns. The superior performance of the RBF SVM can be attributed to its ability to capture non-linear relationships between input features and the target variable, which is crucial for modeling real-world data with intricate dependencies. Additionally, SVMs incorporate regularization parameters that prevent overfitting, ensuring better generalization to unseen data. The flexibility in hyperparameter tuning further enhances the RBF SVM's predictive performance, making it a more robust model compared to linear regression

Table 9 Performance metrics of trip production models

Performance metrics	**RBF-Training**	**RBF-Testing**	**Linear-Training**	**Linear-Testing**
MAE	348.291	310.191	328.694	351.684
MSE ($x10^6$)	0.145	0.120	0.125	0.131
RMSE	381.281	346.907	353.323	362.627
MSLE	0.000	0.000	0.000	0.000
RMSLE	0.014	0.013	0.015	0.014
R^2	1.000	0.999	1.000	0.998
MAPE	1.19%	1.13%	1.23%	1.28%
Willmott's index of agreement	0.985	0.979	0.986	0.976
Mielke & Berry index	0.979	0.969	0.980	0.965
Legates & McCabe's index	0.970	0.958	0.972	0.952

Table 10 Performance metrics of trip attraction models

Performance metrics	**RBF-Training**	**RBF-Testing**	**Linear-Training**	**Linear-Testing**
MAE	2076.786	1765.870	2386.877	1998.503
MSE ($x10^6$)	15.11	7.202	17.854	7.638
RMSE	3887.522	2683.693	4225.468	2763.800
MSLE	0.015	0.006	0.015	0.007
RMSLE	0.121	0.080	0.121	0.084
R^2	0.951	0.941	0.942	0.937
MAPE	7.36%	5.69%	7.77%	6.42%
Willmott's index of agreement	0.917	0.894	0.905	0.881
Mielke & Berry index	0.883	0.848	0.867	0.828
Legates & McCabe's index	0.835	0.789	0.810	0.761

The correlation plot for the trip production models and trip attraction models are shown in *Figure 6* and *Figure 7* respectively. It is evident from *Figure 6* and *Figure 7* that almost all the datapoints are located very close to the 45-degree line which adds robustness to the models.

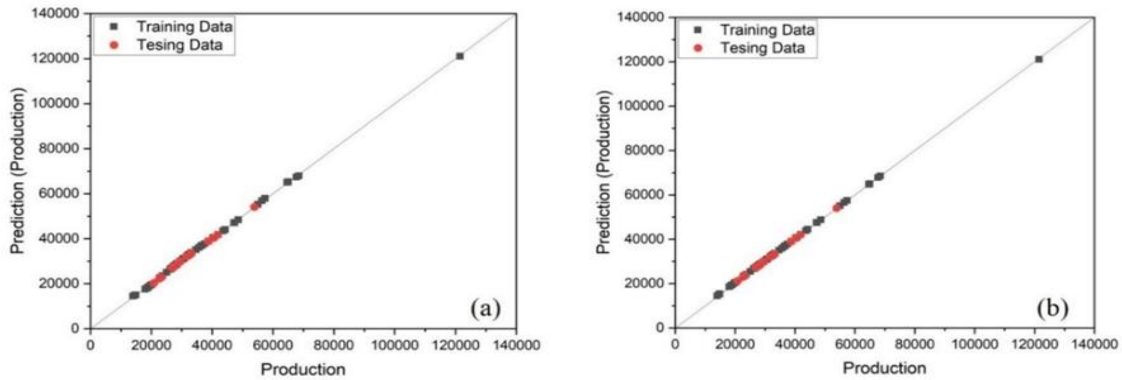

Figure 6 Correlation plot for trip production: (a) SVM-RBF (b) SVM-Linear

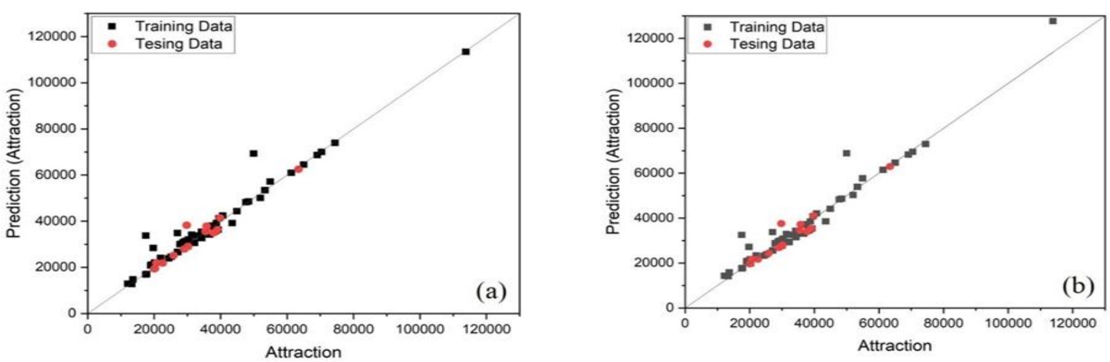

Figure 7 Correlation plot for trip attraction: (a) SVM-RBF (b) SVM-Linear

Comparison between the models developed in the study: In the current study SVM models with RBF and linear kernel functions were trained in the current study along with conventional regression models. After training the models, forecasts for the horizon years 2031 and 2041 were made. The forecasts for the horizon years are shown in *Figure 8*. The graphical plot shown in *Figure 8* highlights almost similar pattern in most of the cases of trip production and trip attraction. The scatter plots showing the correlations between the attributes for trip production and attraction is shown in *Figure 9* and *Figure 10* respectively. The R2 values for the regression models in trip production and trip attraction are 0.70 and 0.99 respectively. The R2 values for trip production and trip attraction models developed using SVM kernel functions are shown in *Table 9* and *Table 10* respectively. KW test: The KW test is a non-parametric statistical test used to determine if there are any statistically significant differences between the medians of three or more independent (unrelated) groups. It is a rank-based test, meaning it works by ranking the data from all groups combined and then comparing the average ranks between groups. The null hypothesis of the KW test is that there are no differences between the medians of the groups. The KW test can help in assessing whether there are statistically significant differences in the predicted values produced by these models. This is crucial for understanding which model performs better on average across different groups or conditions. The null and alternate hypothesis for the test is: H0: The samples come from the same population and Ha: The samples do not come from the same population. The test statistics of KW test are presented in *Table 11*. As the computed p-value in both the cases (trip production and trip attraction) is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H_0. A pairwise comparison is also presented in *Table 12*, which shows whether the models are statistically different from one another or not. The scattergram for trip production and trip attraction models is shown in *Figure 11*.

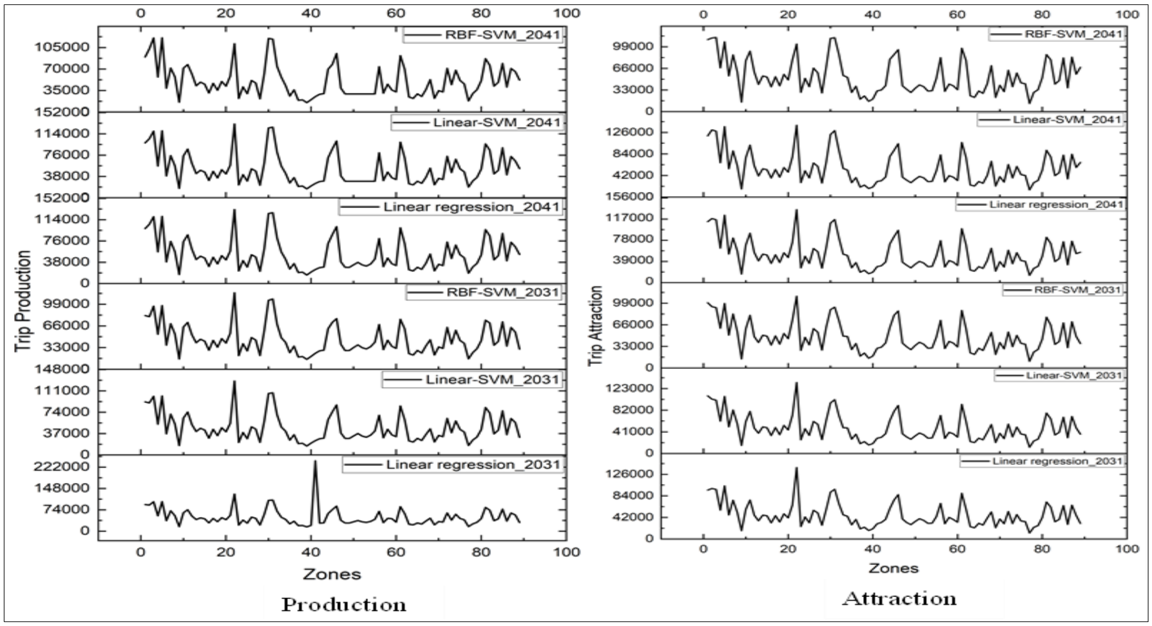

Figure 8 Forecasts for horizon years

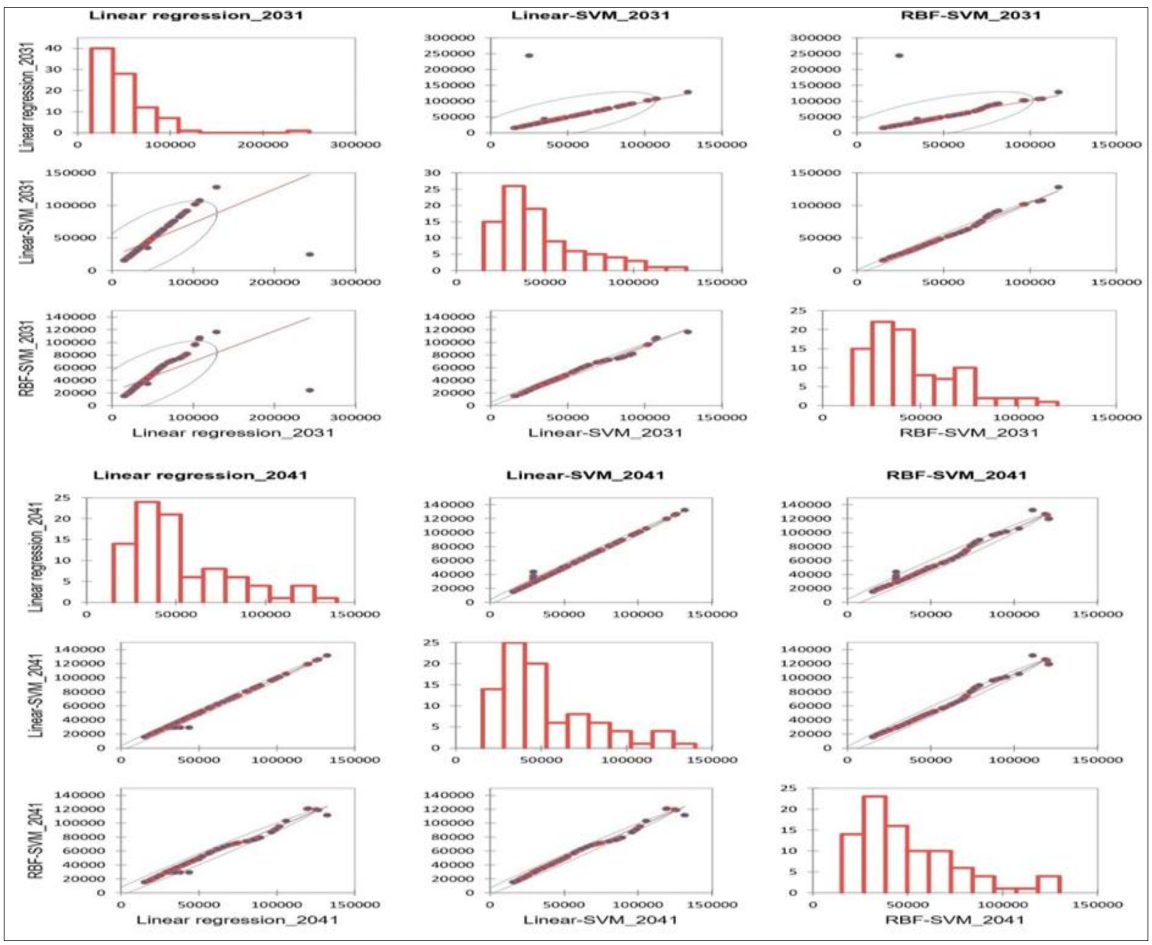

Figure 9 Scatter plot for trip production model

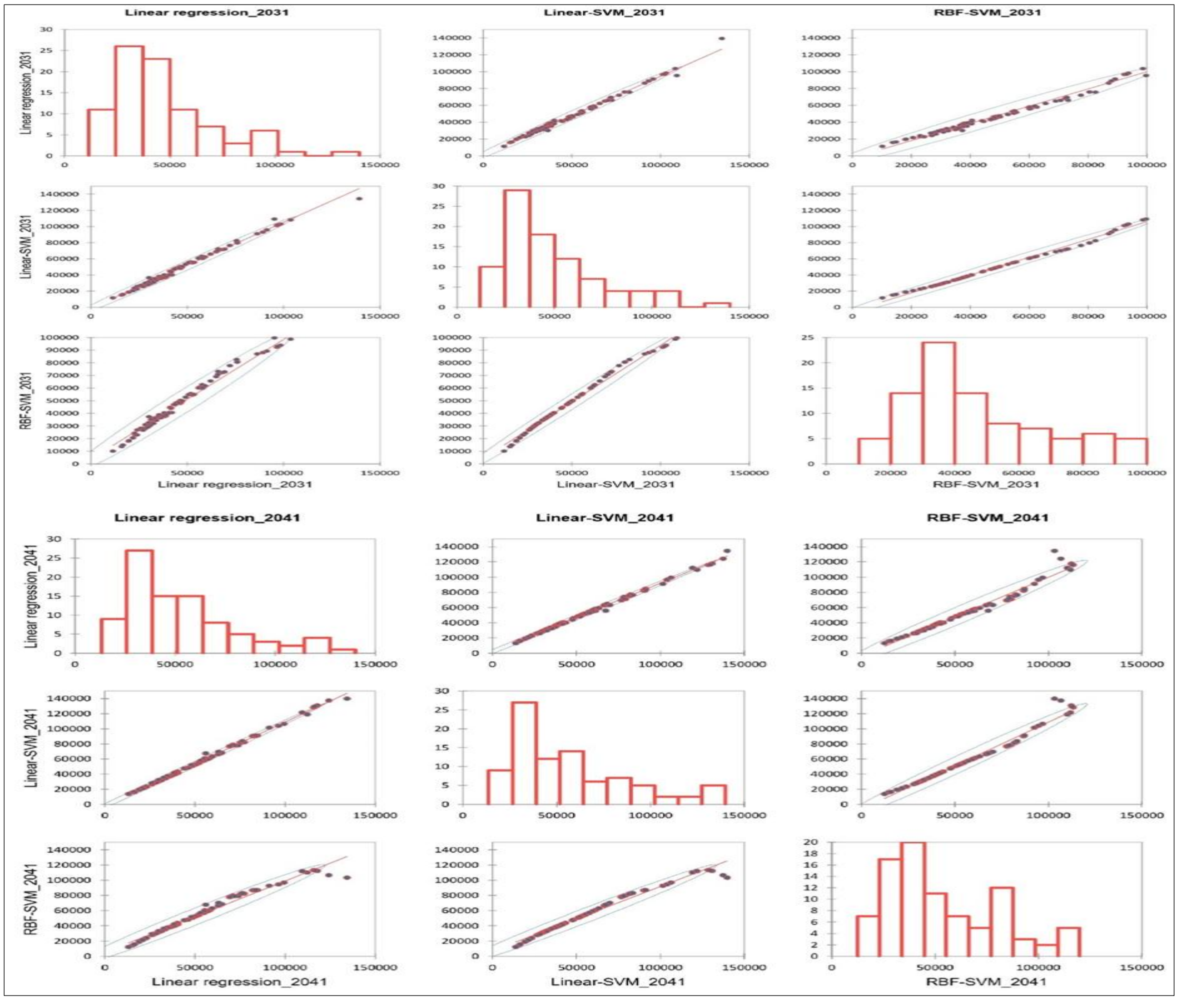

Figure 10 Scatter plot for trip attraction model

Table 11 KW test statistics

	Production	Attraction
K (Observed value)	2.498	6.094
K (Critical value)	11.070	11.070
DF	5	5
p-value (one-tailed)	0.777	0.297
alpha	0.05	0.05

Table 12 Significant difference among trip generation models

Trip generation models	Linear regression 2031	Linear-SVM 2031	RBF-SVM 2031	Linear regression 2041	Linear-SVM 2041	RBF-SVM 2041
Linear regression_2031	-	No	No	No	No	No
Linear-SVM_2031	No	-	No	No	No	No
RBF-SVM_2031	No	No	-	No	No	No
Linear regression_2041	No	No	No	-	No	No
Linear-SVM_2041	No	No	No	No	-	No
RBF-SVM_2041	No	No	No	No	No	-

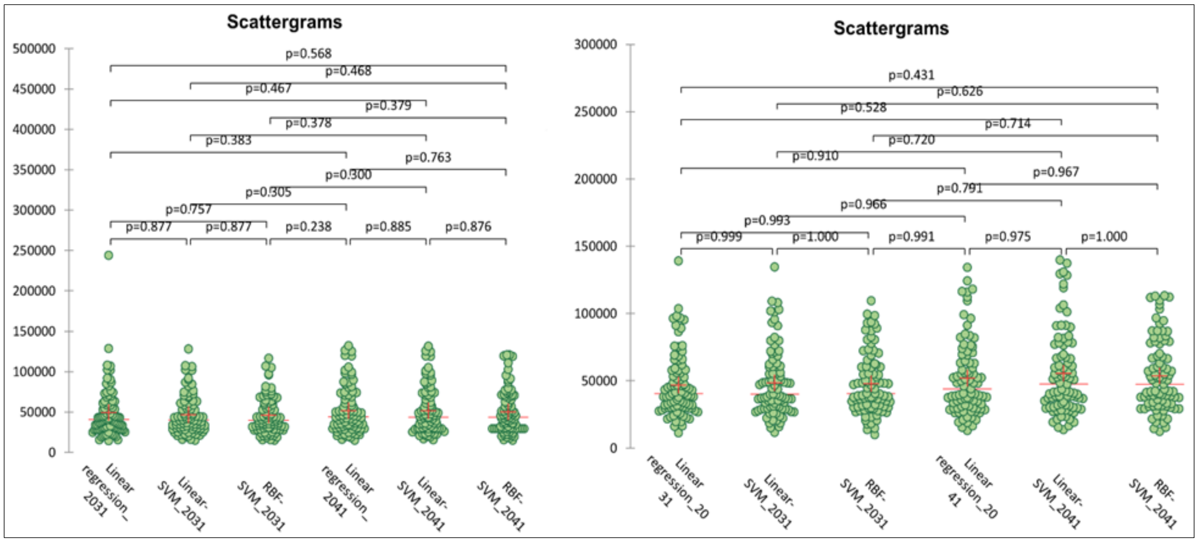

Figure 11 Scattergrams for: Trip production (left) and Trip attraction (right)

Limitations

Using support vector-based machine learning models such as those with RBF kernels for developing trip generation models have several limitations. These models can exhibit high variability in performance and may overfit the training data, reducing their generalizability to new data. They are computationally intensive, requiring significant resources and time for training, especially with large datasets. The performance is highly sensitive to hyperparameter choices, necessitating extensive tuning. Additionally, SVM models are less interpretable than linear models, making it difficult to understand the relationships between input features and trip generation. Scalability issues also arise as dataset size increases, further complicating their practical application. A complete list of abbreviations is listed in *Appendix I*.

5.Conclusion and future work

Machine learning can be efficiently applied to develop trip generation models as an alternative to conventional regression models. SVM-based models developed using the RBF kernel outperformed those using linear kernel functions and conventional regression models. The trip production models demonstrated higher accuracy and predictive performance compared to trip attraction models in both training and testing phases. The SVM model with the RBF kernel performed better than the SVM model with linear kernel function and conventional regression model. The R^2 value for the trip production and trip attraction models developed using the RBF kernel function was found to be 1 and 0.95, respectively. The KW test was employed to assess the statistical significance of the models, revealing no significant differences among them. The findings suggest that SVM models, particularly those utilizing the RBF kernel, offer a superior alternative to conventional regression models for trip generation, providing transportation planners with a more efficient tool for estimating travel demand.

Future work should focus on exploring advanced machine learning algorithms, such as Random Forests, Gradient Boosting, and Deep Learning, to enhance trip generation models. Incorporating additional variables like socio-economic and land-use data, as well as real-time data sources, can improve model accuracy and adaptability. Optimization of SVM kernel functions, testing hybrid or ensemble models, and conducting uncertainty analysis can further refine predictions. Applying these models to larger, diverse datasets and integrating them with decision-support tools for planners will ensure scalability and practical utility. Finally, robust statistical validation across varied scenarios will strengthen confidence in machine learning-based trip generation methodologies.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

Data will be made available on reasonable request to the corresponding author.

Author's contribution statement

Anish Kumar: Writing, editing, review. **Amit Kumar:** Machine learning, editing, review. **Sanjeev Sinha:** Conceptualization, review, editing.

References

[1] Smart A, Smart J. Urbanization and the global perspective. Annual Review of anthropology. 2003; 32(1):263-85.

[2] https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2017_world_population_prospects-2017_revision_databooklet.pdf. Accessed 25 November 2024.

[3] Goldman T, Gorham R. Sustainable urban transport: four innovative directions. Technology in Society. 2006; 28(1-2):261-73.

[4] Oladimeji D, Gupta K, Kose NA, Gundogan K, Ge L, Liang F. Smart transportation: an overview of technologies and applications. Sensors. 2023; 23(8):1-32.

[5] Dodeen AM. Developing trip generation models utilizing linear regression analysis: Jericho city as a case study. Doctoral Dissertation, An-Najah National University, Nablus. 2014.

[6] Sofia GG, Ali AH, Al-zubaidy HA. Trip generation modeling for selected zone in Al-Diwaniyah city. Journal of Engineering and Sustainable Development. 2012; 16(4):167-80.

[7] De OLK, Herédia RT, Bertoncini BV, De ORL. Freight trip generation to buildings under construction: a comparative analysis with linear regression and generalised linear regression. Transportes. 2020; 28(5):28-42.

[8] Monzon J, Goulias K, Kitamura R. Trip generation models for infrequent trips. Transportation Research Record. 1989; 1220:40-6.

[9] Amavi AA, Romero JP, Dominguez A, Dell'olio L, Ibeas A. Advanced trip generation/attraction models. Procedia-Social and Behavioral Sciences. 2014; 160:430-9.

[10] Cordera R, Coppola P, Dell'olio L, Ibeas Á. Is accessibility relevant in trip generation? modelling the interaction between trip generation and accessibility taking into account spatial effects. Transportation. 2017; 44:1577-603.

[11] Chang JS, Jung D, Kim J, Kang T. Comparative analysis of trip generation models: results using home-based work trips in the Seoul metropolitan area. Transportation Letters. 2014; 6(2):78-88.

[12] Triantakonstantis D, Mountrakis G. Urban growth prediction: a review of computational models and human perceptions. Journal of Geographic Information System. 2012; 4(6): 555-87.

[13] Wegener M. Overview of land use transport models. In handbook of transport geography and spatial systems 2004 (pp. 127-46). Emerald Group Publishing Limited.

[14] Linard C, Tatem AJ, Gilbert M. Modelling spatial patterns of urban growth in Africa. Applied Geography. 2013; 44:23-32.

[15] Banister D. Cities, mobility and climate change. Journal of Transport Geography. 2011; 19(6):1538-46.

[16] Liu Z, Liu Y, Meng Q, Cheng Q. A tailored machine learning approach for urban transport network flow estimation. Transportation Research Part C: Emerging Technologies. 2019; 108:130-50.

[17] Müller-hannemann M, Rückert R, Schiewe A, Schöbel A. Estimating the robustness of public transport schedules using machine learning. Transportation Research Part C: Emerging Technologies. 2022; 137:103566.

[18] Chaturvedi V, De VWT. Machine learning algorithms for urban land use planning: a review. Urban Science. 2021; 5(3):1-16.

[19] Surden H. Machine learning and law: an overview. Research Handbook on Big Data Law. 2021:171-84.

[20] Tripathi RP, Pandey K. Gene-expression programming for scour around spur dike. International Journal of Hydrology Science and Technology. 2023; 15(3):295-303.

[21] Talpur MA, Khahro SH, Ali TH, Waseem HB, Napiah M. Computing travel impendences using trip generation regression model: a phenomenon of travel decision-making process of rural households. Environment, Development and Sustainability. 2023; 25(7):5973-96.

[22] Kumar A, Sinha S, Pandey D, Maurya MC, Chauhan VB. Advanced regression models for assessing the strength of multi-walled carbon nanotube-modified high-volume fly ash concrete. Asian Journal of Civil Engineering. 2024; 25(2):2247-68.

[23] Gupta N, Patel J, Gond S, Tripathi RP, Omar PJ, Dikshit PK. Projecting future maximum temperature changes in river Ganges basin using observations and statistical downscaling model (SDSM). In river dynamics and flood hazards: studies on risk and mitigation 2022 (pp. 561-85). Singapore: Springer Nature Singapore.

[24] Parsaie A, Haghiabi AH, Latif SD, Tripathi RP. Predictive modelling of piezometric head and seepage discharge in earth dam using soft computational models. Environmental Science and Pollution Research. 2021; 28(43):60842-56.

[25] Kumar A, Sinha S. Policy options for sustainable urban transportation: a quadrant analysis approach. In risk, reliability and sustainable remediation in the field of civil and environmental engineering 2022 (pp. 487-97). Elsevier.

[26] Ranjan R, Sinha S. Mode choice analysis for work trips of urban residents using multinomial logit model. Innovative Infrastructure Solutions. 2024; 9(10):1-4.

[27] Kumar R, Madhu E, Dahiya A, Sinha S. Analytical hierarchy process for assessing sustainability: indicators of public transportation systems, pedestrians and feeder services in developing countries. World Journal of Science, Technology and Sustainable Development. 2015; 12(4):281-93.

[28] Mcnally MG. The four-step model. In handbook of transport modelling 2007 (pp. 35-53). Emerald Group Publishing Limited.
[29] Shan R, Zhong M, Du D, Lu C. Comparison of trip generation results from activity-based and traditional four-step travel demand modeling: a case study of Tampa, Florida. In conference of the transportation association of Canada 2012 (pp. 1-19).
[30] Mustafa RA. Developing a high-fidelity GIS-based travel demand model framework for improved network-wide traffic estimation. Doctoral Dissertation, The University of New Brunswick. 2019.
[31] Sarna AC, Prakash A, Mukhopadhaya D, Bhatia N, Suthar HH. Planning of mass rapid transit system-a case study of Delhi. In Journal of the Indian Roads Congress 1992; 53(3):553-88.
[32] Anderson MD, Olander JP. Evaluation of two trip generation techniques for small area travel models. Journal of Urban Planning and Development. 2002; 128(2):77-88.
[33] Navya SV, Kumar SS, Kattoor GJ. Trip generation model for the core area of Thiruvananthapuram city. International Journal of Innovative Research in Science, Engineering and Technology. 2013; 2(1):99-106.
[34] Pangotra P, Sharma S. Modeling travel demand in a metropolitan city case study of Bangalore, India. IIM Ahmedabad, 2006.
[35] Walker J, Li J, Srinivasan S, Bolduc D. Travel demand models in the developing world: correcting for measurement errors. Transportation Letters. 2010; 2(4):231-43.
[36] Tiwari G, Jain D. Accessibility and safety indicators for all road users: case study Delhi BRT. Journal of Transport Geography. 2012; 22:87-95.
[37] Abdel-aal MM. Calibrating a trip distribution gravity model stratified by the trip purposes for the city of Alexandria. Alexandria Engineering Journal. 2014; 53(3):677-89.
[38] Jin PJ, Cebelak M, Yang F, Zhang J, Walton CM, Ran B. Location-based social networking data: exploration into use of doubly constrained gravity model for origin–destination estimation. Transportation Research Record. 2014; 2430(1):72-82.

Dr. Anish Kumar holds a doctorate in Civil Engineering and has numerous publications to his credit in prestigious journals and publication houses such as ASCE, Springer, and Elsevier. He has a keen interest in exploring novel applications of sustainable building materials and specializes in applying AI/ML techniques to solve civil engineering problems. He is currently serving as an Assistant Professor in the Department of Civil Engineering at Rajkiya Engineering College, Azamgarh, Uttar Pradesh, India.
Email: anish@gecazamgarh.ac.in

Mr. Amit Kumar holds a postgraduate degree in Transportation Engineering and is currently pursuing Ph.D. at Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India. He is presently serving as an Assistant Professor in the Department of Civil Engineering at Rajkiya Engineering College, Azamgarh, India.
Email: amitkumar@gecazamgarh.ac.in

Prof. Sanjeev Sinha holds a doctorate in Civil Engineering and has successfully supervised numerous master's and doctoral theses. He is the author of several high-impact journal papers and is currently serving as a Professor in the Department of Civil Engineering at the National Institute of Technology, Patna, Bihar, India.
Email: sanjeev@nitp.ac.in

Appendix I

S. No.	Abbreviation	Description
1	BRTS	Bus Rapid Transit System
2	CMP	Comprehensive Mobility Plan
3	KS	Kolmogorov Smirnov
4	KW	Kruskal Wallis
5	LBSN	Location Based Social Networking
6	MAE	Mean Absolute Error
7	MAPE	Mean Absolute Percentage Error
8	MRTS	Mass Rapid Transit System
9	MSE	Mean Squared Error
10	MSLE	Mean Squared Logarithmic Error
11	OD	Origin-Destination
12	PUA	Patna Urban Agglomeration Area
13	R^2	R Squared
14	RBF	Radial Basis Function
15	RMSE	Root Mean Squared Error
16	RMSLE	Root Mean Squared Logarithmic Error
17	SVM	Support Vector Machine
18	TAZ	Traffic Analysis Zones