TECHNICAL PAPER

JOURNAL OF THE SOUTH AFRICAN INSTITUTION OF CIVIL ENGINEERING
ISSN 1021-2019 (print) | ISSN 2309-8775 (online)
Vol 68 No 1, March 2026, Pages 21–30, Paper 1795

Prediction of operating speed on horizontal curve and tangent: A comparative study of multiple linear and principal component regression

KK Tottadi, A Mehar

Operating speed is considered a crucial factor in evaluating geometric design consistency on highways. The operating speed models developed in earlier studies predominantly employed the approach of multiple regression analysis. Using this approach, it is possible to find the presence of multicollinearity among independent variables in the developed models. To examine the effect of multicollinearity, the current research formulated prediction models for operating speed on both horizontal curves and tangent sections using multiple linear regression (MLR) and principal component regression (PCR). The MLR analysis is applied to model the operating speed, and it showed that multicollinearity (variance inflation factor > 4) exists among various independent variables, which affects the predictability of the model developed using MLR. Therefore, principal component analysis (PCA) was applied to eliminate the multicollinearity between independent variables. The PCA results revealed that only the first two components (PC1 and PC2) with eigenvalues greater than one were selected for PCR analysis. Furthermore, when comparing the results obtained by MLR and PCR, the PCR model yields good predictions for both curved and tangent section cases. Finally, the PCR model was selected as the best-fitted model for predicting operating speed on the curve and tangent sections, as it had the lowest variable inflation factor values and the highest R^2 values compared with the MLR model. According to the findings, this study suggests employing PCR instead of MLR when multicollinearity is present in the dataset.

Keywords: Operating speed, multiple linear regression, principal component analysis, principal component regression

INTRODUCTION

The driver, road, and vehicle are responsible for the occurrence of road accidents in the traffic system. As per the World Health Organization (WHO), around 1.19 million people die annually as a result of road accidents (WHO 2023). In India, 461 312 road accidents occurred in 2021, resulting in 153 972 fatalities and 384 448 injuries (MoRTH 2022). In 2022, there was an average of 11.19% more traffic crashes than in 2021. As per the Ministry of Road Transport and Highways (MoRTH 2022), 54 593 accidents were reported on horizontal curves during 2022 in India. The number of crashes, fatalities, and injuries increased 10% from 2021 to 2022 on curved sections. Driver error is solely responsible for 81.2% of total road accidents during 2022 (MoRTH 2022). However, an accident may not always be caused by the same driver in the same conditions. This indicates how a driver's behaviour in a similar condition can be uncertain or unpredictable. The driver's behaviour mainly depends on the road environment and conditions. A sudden change in road alignment can cause a driver to become confused, leading to an accident. Hence, the horizontal curve locations have been considered critical sections within the roadway network. According to the MoRTH (2022) report, 309 247 accidents were reported on the tangent section in India during 2022. The number of

DR KIRAN KUMAR TOTTADI is an Assistant Professor in the Department of Civil Engineering at Anil Neerukonda Institute of Technology & Sciences (ANITS), Visakhapatnam. He holds an MTech and PhD in Transportation Engineering from the National Institute of Technology Warangal, with research focused on operating speed modelling, highway geometry, and road safety. He is actively involved in teaching, research, and guiding undergraduate and postgraduate projects. His academic interests include highway safety, geometric design consistency, traffic emissions, traffic flow modelling and intelligent transportation systems.

Contact details:
Department of Civil Engineering
Anil Neerukonda Institute of Technology and Sciences (A)
Visakhapatnam 531162
India
kirankumar.ce@anits.edu.in

DR ARPAN MEHAR is an Associate Professor in the Transportation Engineering Division at the National Institute of Technology Warangal, India. He obtained his PhD in Transportation Engineering from IIT Guwahati. His research interests include traffic emissions and level of service, traffic flow modelling, highway capacity analysis, and geometric design of highways. He has made significant contributions to transportation engineering research through publications in reputed journals. He has successfully guided seven PhD scholars and several postgraduate students, and continues to mentor young researchers in advanced transportation systems and sustainable highway design.

Contact details:
Transportation Engineering Division
Department of Civil Engineering
National Institute of Technology Warangal
Warangal 506004
India
arpan@nitw.ac.in

KK TOTTADI, A MEHAR. Prediction of operating speed on horizontal curve and tangent: A comparative study of multiple linear and principal component regression. J. S. Afr. Inst. Civ. Eng. 2026:68(1), Art. #1795, 10 pages. http://dx.doi.org/10.17159/2309-8775/2026/v68n1a2

accidents, fatalities, and injuries increased by 11.2% from 2021 to 2022 on the tangent section. Vehicle speed tends to be high on straight roads in open areas, which has been documented in the high percentage of accidents. Moreover, approximately 80% of accidents occurred on tangent and curved sections. Therefore, curves and tangent sections have been identified as critical highway geometric features associated with significant safety concerns.

One of the primary reasons for highway accidents is speeding. It accounted for 72.3% of road accidents in India during 2022. Speed is resultingly considered an essential parameter for designing highway geometry. Design speed is used to determine the geometric characteristics of a roadway (AASHTO 2011). The primary drawback of the design speed concept lies in its reliance on the design speed of the most constrained geometric feature within the highway segment. Consequently, this approach does not explicitly account for the speeds at which vehicles actually travel on curved and tangent sections. This may result in inconsistencies between successive roadway elements. Operating speed, by contrast, is observed under free-flow traffic conditions (AASHTO 2011). Therefore, operating speed, rather than design speed, should be considered when evaluating the consistency of highway geometric design.

Speed varies among drivers and tends to become more unpredictable when different vehicle classes are present in the traffic stream. Drivers do not maintain a uniform speed when traversing a specific geometric feature. Consequently, the 85th percentile speed (operating speed) is adopted to assess geometric design consistency. Although the operating speed may be the same at two different geometric features, the underlying speed distributions can differ significantly. Earlier studies often assumed that vehicle speeds follow a normal distribution; however, speed profiles may follow non-normal distributions. Variability in speed distribution is therefore an important factor in evaluating the geometric design consistency of horizontal curves on two-lane rural highways (Bassani *et al* 2016; Jesna & Anjeneyulu 2016). However, most studies focus on two-lane highways, with limited research on four-lane divided highways (Tottadi & Mehar 2022; Goyani *et al* 2022; Tottadi & Mehar 2024). Resultingly, there is a need to analyse speed distributions for geometric design consistency evaluation on four-lane divided highways.

OBJECTIVES

The study aims to achieve the following objectives:

- Examine the speed distribution of cars on both the curve and tangent sections of four-lane divided highways.
- Develop operating speed models on curve and tangent sections using multiple linear regression (MLR) and investigate the effect of multicollinearity.
- Predict the operating speed on curve and tangent sections using principal component regression (PCR) by addressing multicollinearity.

LITERATURE REVIEW

Several studies have been conducted to predict operating speeds on two-lane highways, while limited studies have been reported on multi-lane highways. Shankar and Mannering (1998) studied heterogeneity problems involving lane-mean traffic speeds and lane-speed deviations on multi-lane highways. They proposed a fundamental model for predicting mean speed, considering both environmental and traffic factors. The three-stage least squares method was adopted to develop a model for mean speed and speed deviation. According to the research results, there are inherent correlations between lane mean speeds and speed deviations. Additionally, it was reported that the mean speed in the middle lane decreased as the hourly traffic volume in the middle lane increased.

Tarris *et al* (1996) developed operating speed models on horizontal curves using linear regression and panel analysis of multilane urban roads. They observed a negative correlation between the degree of curvature and operating speed, indicating that operating speed decreases as the degree of curvature increases. Poe and Mason (2000) developed models to predict mean speed on horizontal curves using linear regression analysis. Their results indicated that mean speed decreases with increasing curvature and grade, and increases with wider lane widths.

Fitzpatrick *et al* (1995) examined the association between design and operating speed of various vehicle types on highways. The authors reported that the curve radius was an essential parameter on horizontal curves, and the vertical distance was recognised as an important factor on vertical curves. Additionally, it was found that the operating speed decreased with increasing access density.

Medina and Tarko (2005) and Kim and Choi (2013) examined the influence of geometric parameters on the operating speed of four-lane urban and rural highways. They found that more speed was observed on rural highways than on suburban highways. Vehicles travel at speeds exceeding the posted limits when the speed limit is below 90 km/h.

Tseng *et al* (2005) developed exponential models for predicting the mean 85th percentile speed, and established speed distributions for each speed profile. They found that the vehicle type, speed limit, and the distance between signalised intersections are the three most crucial factors that affect free-flow speed. Ali *et al* (2007) developed free-flow speed models on tangent sections as a function of geometric characteristics on multilane urban roads. They revealed that the operating speed was highly influenced by tangent length and the presence of a median. Gong and Stamatiadis (2008) predicted operating speed on both the inside and outside lanes of horizontal curves of four-lane rural highways. They concluded that the significant parameters for estimating operating speed in the inner lane are curve length and the presence of a paved shoulder. Similarly, the operating speed on the outer lane increased with parameters like curve radius and paved shoulders.

He *et al* (2010) established relationships between geometric parameters and the 85th percentile speed on long tangents on multi-lane highways. They concluded that the distance between the curve and the tangent section influences the operating speed. Andrade *et al* (2016) used multivariate analysis for predicting operating speed on expressways. They employed principal component analysis (PCA) to identify the variables and determine the optimal number of principal components. Based on PCA, the authors concluded that three principal components explain the variance. Thereafter, stepwise linear regression was used to develop a free-flow speed model based on the components obtained from the PCA. Finally, the developed models were compared with the Highway Capacity Manual 2010 models. Yan *et al* (2017) used continuous speed profiles of three-dimensional combinations for estimating operating speed on multilane highways. They selected the change rate of curvature, degree of curvature, curve angle, grade, and lane width as variables rather than a single index.

Table 1 Details of operating speed models

Author	Parameters used	Modelling technique	Country	Data collection
Shankar and Mannering (1997)	% of trucks, % of cars, mean speed	Three stage least square	USA	Radar gun
Poe and Mason (2000)	Curve radius, degree of curve, lane width, mean speed	Multiple linear regression analysis	USA	Radar gun
Fitzpatrick *et al* (1995)	Curve radius, vertical sight distance, design speed, operating speed	Linear regression analysis	USA	Radar gun
Gong and Stamatiadis (2008)	Shoulder type, median type, pavement type, length of curve, grade index	Multiple linear regression analysis	USA	Radar gun
Kim and Choi (2013)	Curve radius, curvature change rate, tangent length, deflection angle	Multiple linear regression analysis	China	Trap length method
Yan *et al* (2017)	Curve radius, change rate of curvature, deflection angle of curve, grade, and lane width	Multiple exponential analysis	China	GPS and radar gun
Maji and Tyagi (2018)	Curve radius, curve length, superelevation, gradient	Linear regression	India	Trap length method
Maji *et al* (2020)	Curve length, curve radius, tangent speed, deflection angle, vertical grade	Step wise multiple linear regression	India	Trap length method
Sil *et al* (2019)	Preceding tangent length, degree of curvature, deflection angle	Step wise multiple linear regression	India	Trap length method

Morris *et al* (2014) used the ordinary least squares method to estimate the operating speed of cars and trucks on multilane highways with a combination of horizontal and steep vertical grades. The results revealed that the curve radius has a greater influence on the operating speed of a car than that of a truck. Similarly, vertical grade seems to have a significantly greater influence on truck operating speed compared to a car. Maji and Tyagi (2018) studied the impact of highway geometry on the operating speed of cars and sport utility vehicles along a horizontal curve on a four-lane divided highway. They collected speed at the point of curvature, the midpoint of the curve, the point of tangency, and 50 m before the point of tangency. Their findings indicated that the curve radius had a significant effect on the operating speed at the midpoint of the curve. Maji *et al* (2020) employed a stepwise MLR approach to develop speed prediction models for cars, light commercial vehicles, and heavy commercial vehicles at the 85^{th} and 98^{th} percentiles on four-lane divided rural highways. Variables such as curve length and deflection angle had an impact on the 85^{th} and 98^{th} percentile speeds. Sil *et al* (2019) proposed free-flow speed models for cars, light commercial vehicles, and heavy commercial vehicles at the midpoint of the curve on four-lane divided highways. The authors concluded that the length of the curve, deflection angle, and preceding tangent length are the important independent parameters for model development. The variables, analysis techniques, and vehicle type used in earlier studies are summarised in Table 1.

Most studies employed MLR to predict operating speed. However, this approach often does not adequately identify the presence of multicollinearity among independent variables. To examine multicollinearity, alternative techniques such as PCR should be adopted. Previous models have incorporated geometric variables including curve radius, curve length, degree of curvature, deflection angle, median width, shoulder width, and access density. In earlier studies, free-flow vehicle speeds were typically collected using radar or laser guns. Preceding tangent length has also been identified as a key parameter for estimating operating speed on tangent sections. Therefore, additional studies are required to predict operating speed using modelling techniques beyond MLR and to incorporate a wider range of geometric variables for four-lane divided highways, thereby supporting improved geometric highway design.

SITE SELECTION

The roadways used in this study are located in five regions of India: Assam, Madhya Pradesh, Delhi, Mumbai, and Telangana. Table 1 presents the main criteria employed for selecting the study sites. A total of 44 sections were chosen, comprising 22 tangent sections and 22 curved sections.

Figure 1 (a) Curve section on SH 18 in Madhya Pradesh and (b) tangent section on NH-27 in Assam

The tangent and curved sections were selected based on the following criteria:

- Located far from towns or built-up areas
- Absence of intersections within 1 km of the tangent or curve
- Good and dry pavement conditions
- Free from roadside activities, obstructions, or disturbances near the selected location.

The various study area locations are shown in Figure 1.

DATA COLLECTION

The data collection process included both highway geometric data and free-flow vehicle speeds. Geometric characteristics such as curve radius (R), degree of curvature (DC), and curve length (LC) were estimated by importing Google Maps images of the curves into AutoCAD. The exact locations of curve points were obtained using an Android GPS application called OSM Tracker. Site characteristics, including the number of lanes, lane width (W), shoulder type and width (ST and SW), and median width (MW), were measured using a measuring tape. Details of all study sites are presented in Table 2. Access control type and access point density were also assessed and validated using Google Maps.

The free-flow speeds of small car (CS), big car (CB), heavy commercial vehicle (HCV), two-wheeler (2W), bus (Bus), and light commercial vehicle (LCV) were measured at the mid-point of the curve and tangent section using a radar gun during the day under good weather conditions. The descriptive statistics of the various vehicles are presented in Table 3.

Table 2 Details of selected sites

Parameter	Curve	Tangent
Radius of curve (m)	180 to 745	NA
Length of curve (m)	170 to 1 100	500 to 3 000
Deflection angle (^{0})	32 to 106	NA
Degree of curvature (^{0})	8 to 32	NA
Lane width (m)	7.25 to 7.5	7 to 7.5
Shoulder width (m)	1.3 to 2	1.3 to 3
Median width (m)	4 to 5.6	4.3 to 6.2
Access type (per km)	0 to 12	0 to 13
NA = not applicable		

Table 3 Descriptive statistics of free-flow speed

Statistics	Curve				Tangent			
	Car	2W	LCV	HCV	Car	2W	LCV	HCV
Minimum (km/h)	61	42	45	43	66	49	51	48
Maximum (km/h)	112	87	82	65	126	92	88	75
Mean (km/h)	85	59	60	55	91	62	63	58
Standard deviation (km/h)	14	10	10	9	11	8	8	10

Table 4 Results of the F- and Z-test between the small car and the big car

Vehicle type	CS	CB	Vehicle type	CS	CB
Variance (km/h)	111.72	98.09	Mean speed (km/h)	84.91	85.93
$F_{Calculated}$	1.18	$Z_{Calculated}$	1.73		
$F_{Critical}$	1.36	$Z_{Critical}$	1.96		
Alpha	0.05	Alpha	0.05		
Remarks	Not significant		Remarks	Not significant	

SPEED DATA ANALYSIS

Statistical tests were performed on the speeds of different vehicle classes to determine whether any significant differences exist. Z-tests and F-tests were conducted to assess and compare the means and variances among speeds of various vehicle types. Table 4 provides the results of both tests. The F-tests and Z-tests are based on the following two hypotheses:

- H_0 = There is no significant difference in the means/variances between the big car and the small car.
- H_1 = There is a significant difference in the means/variances between the big car and the small car.

The computed F and Z statistics were found to be below their respective critical values, indicating no significant differences in the means and variances between small and large cars. Therefore, it is appropriate to combine both vehicle types into a single category, referred to as "Car."

Analysis of speed distribution

A vehicle's speed is a stochastic variable that conforms to various continuous statistical distributions. The observed speed data for cars at each highway section is grouped into small class intervals to determine the frequency distribution. The class interval used is based on Strug's rule, and the values of frequency were obtained in that interval. Statistical tests, including the Kolmogorov-Smirnov (K-S) test, were conducted to assess the data's fit to a chosen distribution. The statistical test results showed the null hypothesis was rejected at a 95% confidence level, as the Kolmogorov-Smirnov statistic surpassed the critical value of 0.21. The selection of the best-fit distribution is determined by considering both the maximum p-value and the K-S test within a 95% confidence interval. After analysing the speed distribution, the speed profiles follow the normal, log-normal, beta (with four parameters), Weibull (with three parameters), gamma, and generalised extreme value (GEV) distributions. Furthermore, it was observed that the vehicle type 'Car' followed a beta distribution with shape parameter 4 and a log-normal distribution on the majority of roadway sections, including curves. Figure 2 shows the distribution of the curve and tangent of section 1. Likewise, the 85th percentile speeds were calculated from the developed cumulative distribution curves. The parameters include minimum speed, mean speed, and maximum speed, and the 85th percentile speeds were calculated. These statistics are presented in Table 5.

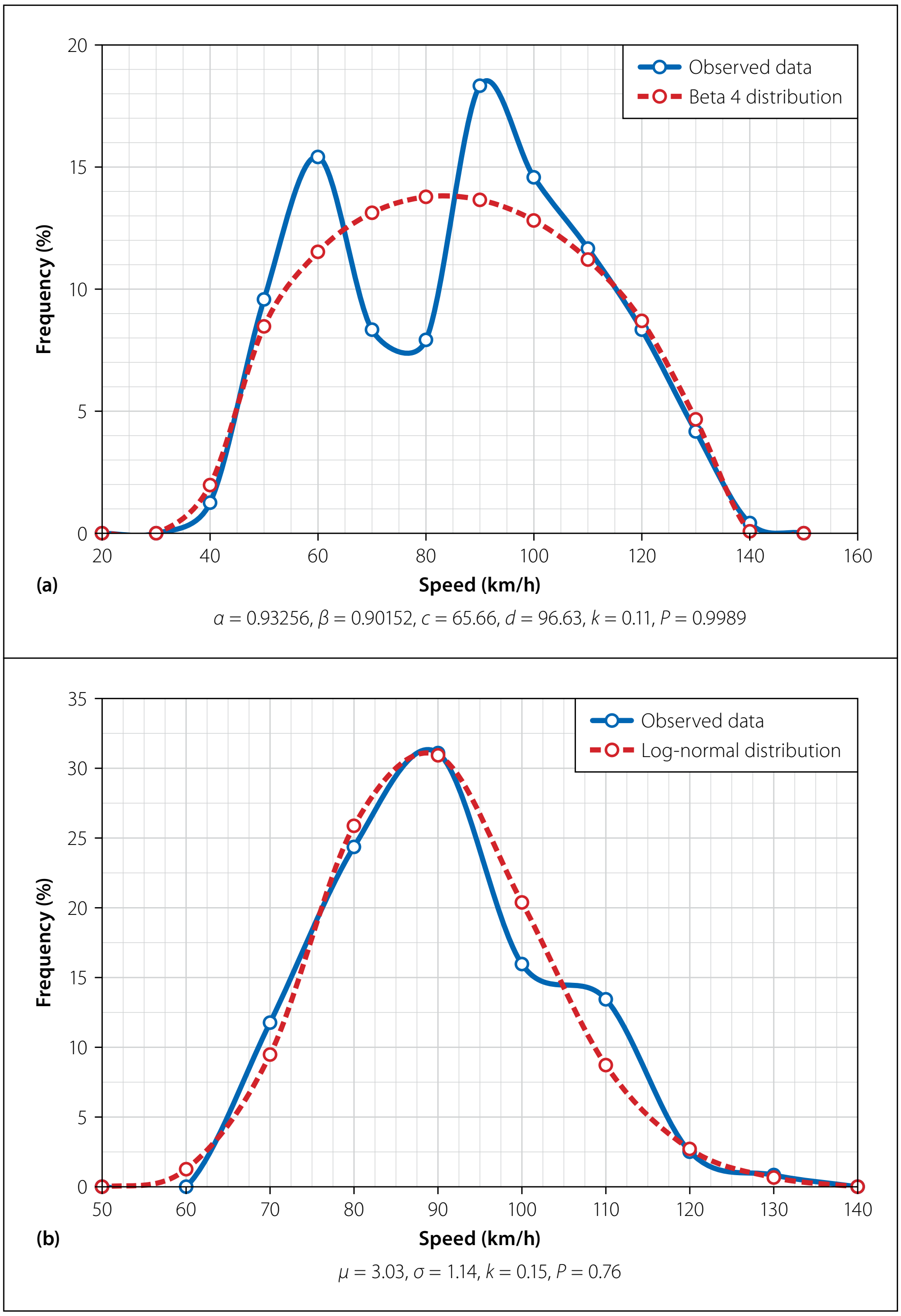

Figure 2 Speed distribution of section 1 (a) curved section and (b) tangent section

Table 5 85th percentile speed of 'Car' for all sites

Statistics	85th percentile speed (km/h)	
	Curve	Tangent
Minimum (km/h)	58	73
Maximum (km/h)	100	110
Mean (km/h)	82	80
Standard deviation (km/h)	11.20	11.85

Table 6 Interpretation rule for the KMO test

KMO value	<0.5	0.5 to 0.6	0.6 to 0.7	0.7 to 0.8	0.8 to 0.9	>0.9
Level of acceptance	Unacceptable	Miserable	Mediocre	Excellent	Meritoire	Marvelous

MODELLING METHODOLOGY

Multiple linear regression

Regression analysis examines the relationship between the dependent and independent variables, explaining the linearity between the two variables. In this study, the 85th percentile speed is the dependent variable, while the geometric parameters serve as the independent variables. The basic form of MLR is as follows:

$$Y_i = a_0 + a_1x_1 + a_2x_2 + a_3x_3 \dots a_nx_n \quad (1)$$

Where
Y_i = dependent variable
a_1, a_2, a_3 and a_n are the coefficients of x_1, x_2, x_3 and x_n.

Principal component analysis

PCA is a mathematical algorithm that aims to minimise the dimensionality of the data (minimising the dataset's feature count while preserving maximal information), while restraining variation in the data as much as possible. This approach aims to extract original information from the dataset and characterise it as a set of novel principal components, which are uncorrelated orthogonal parameters (ideally characterises variability within a singular dataset). It seeks a linear combination of parameters that maximally represents the overall variance among them. This linear combination is referred to as the first principal component, and it accounts for the maximum variance in the data. Subsequently, PCA will look for a second linear combination that accounts for the second largest amount of variance in the data and is uncorrelated with the first principal component. Eigenvalues indicate the proportion of variance explained by each element, while eigenvectors serve as the weights utilised for computing scores of principal components.

Specific steps of principal component analysis

The first step in conducting PCA is to perform the Kaiser–Meyer–Olkin (KMO) test and Bartlett's Test of Sphericity. KMO measures sampling adequacy and the degree of intercorrelation among variables,

Table 7 Results of the developed model

Variables	Coefficient	Standard error	*t*-value	*p*-value	R^2	Adjusted R^2	VIF
Intercept	24.311	0.362	10.23	0.010	0.721	0.697	9.36
R	0.106	0.002	6.32	0.001			5.16
CL	0.005	0.001	2.36	0.002			2.83
DC	-0.033	0.020	-4.56	0.000			6.26
D	-0.001	0.001	-2.21	0.000			1.34
SW	0.091	0.000	3.36	0.002			2.76
MW	7.41	0.000	12.56	0.030			6.24
AD	-0.254	0.001	-3.65	0.020			1.86

Note: R = curve radius; CL = curve length; DC = degree of curvature; D = deflection angle; SW = shoulder width; MW = median width; AD = access density

with values ranging from 0 to 1 (Andrade *et al* 2016). The interpretation of KMO values is presented in Table 6. Bartlett's Test of Sphericity evaluates whether the correlation matrix is an identity matrix, which would indicate that the variables are unrelated and unsuitable for factor analysis. A KMO value greater than 0.5 and a Bartlett's test significance level below 0.05 suggests a strong correlation in the data.

The second step in PCA is to compute the covariance matrix, which quantifies the variance and the degree to which pairs of variables are associated or vary simultaneously. The third step involves calculating the eigenvalues and eigenvectors of the covariance matrix. Eigenvalues represent the magnitude of eigenvectors and, concurrently, quantify the variance contributed by each vector. The fourth step is to select the number of principal components, based on the eigenvalues, which represent the number of principal components in the model using the scree plot.

PRINCIPAL COMPONENT REGRESSION

PCR is the combination of PCA and MLR. It is used to reduce multicollinearity among independent variables and estimate the impact of these variables on the dependent variable. The operating speed is regarded as the dependent variable, while the principal components (PCs) are considered as independent variables. The basic form of PCR is shown in Equation 2:

$$Y_i = \alpha_1 PC_1 + \alpha_2 PC_2 + \ \dots \ \alpha_n PC_n \qquad (2)$$

Where
Y_i = dependent variable
α_1, α_2 and α_n are the coefficients of PC_1, PC_2 and PC_n.

MODELLING OPERATING SPEED USING MULTIPLE LINEAR REGRESSION

The MLR approach was adopted to forecast operating speed on the curve and tangent sections. The results of the developed models are discussed in this section.

Horizontal curve

The dependent variable is the 85th percentile speed, and the independent variables are the geometric variables. Table 7 provides the results of the developed model. It was found that the degree of curvature, deflection angle, and access density have a negative correlation, while curve radius, curve length, shoulder, and median width have a positive correlation. It is clearly indicated that as the degree of curvature and deflection angle increases, the sharpness of the curve also increases, and the corresponding vehicle speed decreases. Furthermore, the curve radius and curve length increase the curve sharpness, while the vehicle speed and stopping sight distance decrease. If the shoulder and median width increase, the speeds of vehicles also increase due to the encouragement of vehicle movement on shoulders and to avoid the glare and influence of opposing vehicles. The standard error, *t*-value, and *p*-value lie within the permissible limits, but the variable inflation factor (VIF) ranges from 1.34 to 9.36. It shows that multicollinearity exists between the relevant independent variables. The general rule of thumb for checking multicollinearity among independent variables is a VIF greater than 4 (Kabacoff 2011). The above results revealed that a few variables exceed the permissible limit value of VIF. Hence, it is necessary to employ an alternative methodology to

Table 8 Results of the developed model

Variables	Coefficient	Standard error	*t*-value	*p*-value	R^2	Adjusted R^2	VIF
Intercept	92.066	0.326	11.26	0.020	0.719	0.685	9.66
SW	5.083	0.020	4.56	0.001			5.21
MW	0.285	0.002	3.96	0.010			4.08
AD	-4.377	0.001	-4.66	0.000			3.66
AT	20.203	0.030	6.66	0.000			8.26
ST	2.723	0.000	4.36	0.001			2.86

Note: SW = shoulder width; MW = median width; AD = access density; AT = access type; ST = shoulder type

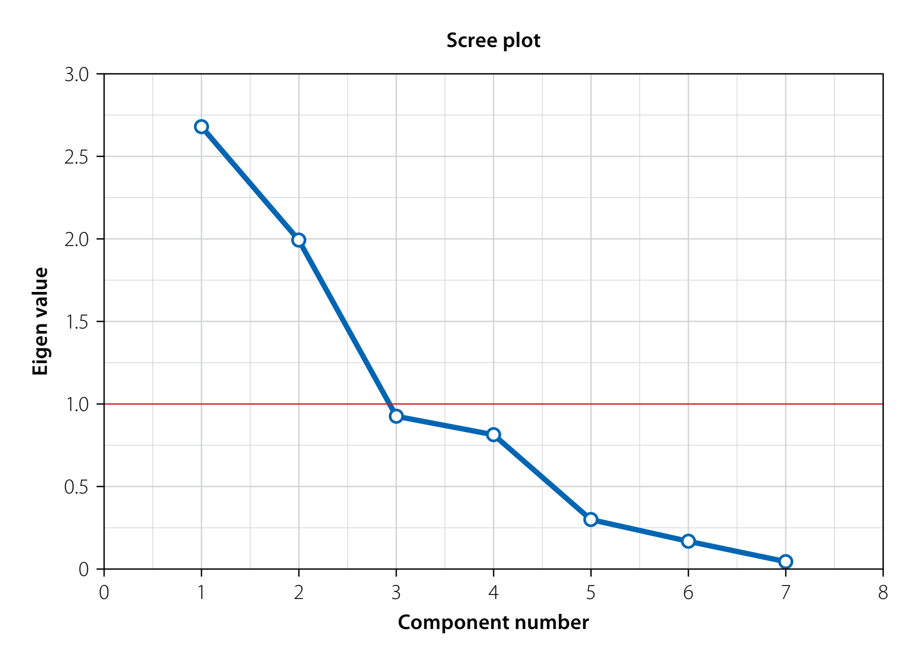

Figure 3 Scree plot

eliminate multicollinearity between the independent variables. In this study, PCA is used to develop models, and the results are presented in the following section.

Tangent

The attainment of maximum speed on a tangent section depends on the geometric and roadside features. The maximum speed observed on a tangent was 110 km/h for a car. The 85th percentile speed was estimated from the cumulative frequency curves, and these speeds were used as the dependent variable. The geometric features, such as shoulder and median width (SW and MW), shoulder type (ST), and other parameters like access density and type (AD and AT), are considered as independent variables. A value of ST = 1 indicates a paved shoulder and ST = 0 indicates an unpaved shoulder. In the same way, AT = 1 denotes no access, while AT = 0 denotes partial or full access. MLR was used to develop the model. The summary of model results is presented in Table 8. It is observed that the shoulder and median width, access, and shoulder type have a positive correlation, while access density has a negative correlation. All statistical measures are within permissible limits except the variable inflation factor. The variable inflation factor ranges from 2.86 to 9.66; it is clearly shown that multicollinearity exists among the independent variables. In the present study, PCA is employed to mitigate multicollinearity, and the results are discussed in the following section.

MODELLING OPERATING SPEED USING PRINCIPAL COMPONENT ANALYSIS

Horizontal curve

Firstly, the KMO and Bartlett's test were conducted to measure sampling adequacy, and to verify the suitability of the data for applying PCA. The null hypothesis for Bartlett's test assumes that the correlation matrix is an identity matrix. In the present study, the KMO value is 0.706, which falls under the 'excellent' level of acceptance, and Bartlett's Test of Sphericity is less than 0.001. Hence, the PCA applies to the collected geometric dataset.

In PCA, the number of principal components is decided based on the eigenvalue criterion. Using this criterion, any element with an eigenvalue of more than 1 is retained. It is observed that only the first two principal components have eigenvalues greater than one (see Figure 3). The examination of the scree plot confirms the presence of two principal components. It also provides information for selecting the number of principal components, as shown in Figure 3.

Therefore, two principal components (PC1 & PC2) are considered for regression analysis, and these two principal components are given sufficient information about the data. The first two principal components explained 66.76% of the variance. Among them, PC1 accounts for 38.28%, while PC2 accounts for 28.47% of the total variance. PC1 is a measure of curve radius, curve length, degree of curvature, and shoulder width, and PC2 is a measure of deflection angle, median width, and access density, as shown in Table 9.

Table 9 PCA results of the curve

Variable	Component	
	PC1	PC2
R	0.957	
DC	-0.957	
CL	0.806	
D		-0.822
SW	0.330	
MW		-0.563
AD		0.908
% of total variance	38.28	28.47

Further, it can be observed that there is a slight difference in the variances of PC1 and PC2 before and after loadings. After loading, the PC1 variance drops to 37.15% from its original value of 38.28%. Similarly, the PC2 variance increased to 29.60% from its original value of 28.47%. The output is suppressed with values less than 0.3. The higher the factor loading, the greater the variable's contribution to the variation of the PCs. The communality values for the variables are as follows: R (91.8%), DC (91.7%), CL (88.7%), D (67.7%), AD (83.2%), SW (54.4%), and MW (42.2%). The communalities values show that the variables accurately represented the variances of the variables in the regression analysis. After finalising the number of principal components, the scores of PC1 and PC2 are estimated. The estimated scores are used to develop a regression model for a horizontal curve. The development of the regression model is explained in the next section.

Principal component regression for curve

The scores of PC1 and PC2 obtained from PCA are used for model development. The scores of PC1 and PC2 are considered as independent variables, and the 85th percentile speeds are regarded as the dependent variable. A multiple regression model was developed for the study. The developed model and its statistics are shown in Table 10.

It can be observed that the coefficient of PC1 is positive, while the coefficient of PC2 is negative. The standard error for the intercept and both components is within

Table 10 Statistical summary of the operating speed model on the horizontal curve

Variables	Coefficient	Standard error	*t*-value	*p*-value	R^2	Adjusted R^2	VIF
Intercept	82.045	0.226	12.45	0.001			1.28
PC1	9.166	0.001	7.48	0.001	0.858	0.787	1.01
PC2	-3.477	0.001	-2.83	0.001			1.08

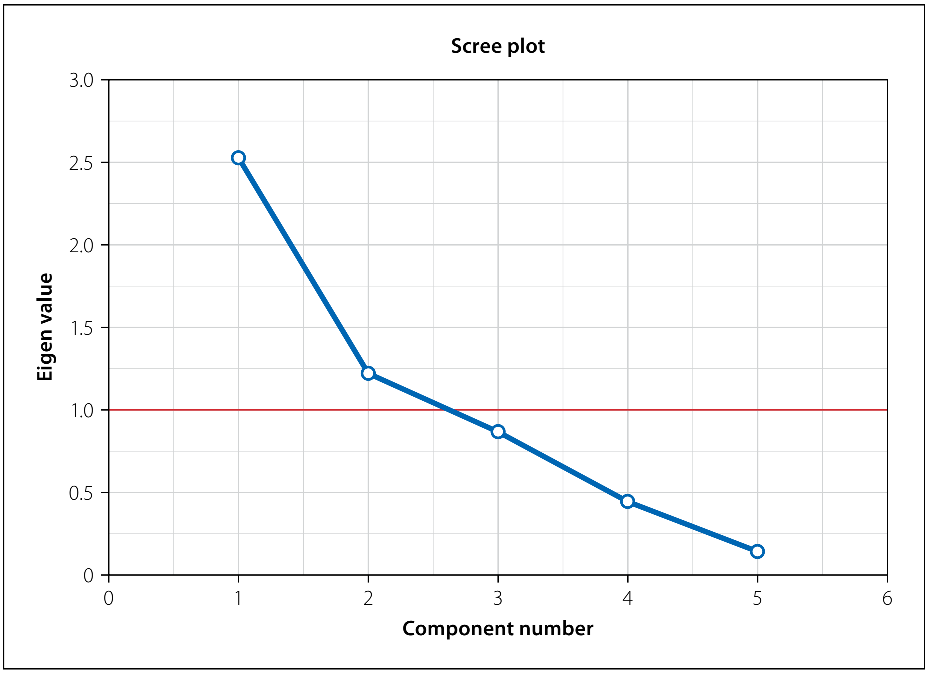

Figure 4 Scree plot

prescribed limits. The *t*-value is high for the intercept compared to the other two components. The value of p is less than 0.05, and hence, all the variables are significant at a 95% confidence interval. The coefficient of determination and adjusted R^2 values are 0.858 and 0.787, respectively.

Tangent section

As noted earlier, the KMO and Bartlett's Test of Sphericity were conducted. The KMO value obtained for the present dataset is 0.787, and Bartlett's test result is less than 0.002, indicating that the geometric data is suitable for PCA. Only the first two principal components have eigenvalues greater than one, and inspection of the scree plot further substantiates the presence of two principal components. Figure 4 illustrates the scree plot used to determine the appropriate number of components.

Accordingly, the regression analysis focuses on two principal components, namely PC1 and PC2, as they capture the essential information within the data. PC1 is associated with shoulder width and median width, while PC2 represents access density, shoulder type, and access type, as summarised in Table 11. Together, the first two principal components accounted for 70.95% of the variance.

Principal component regression for tangent section

The scores derived from the PCA for PC1 and PC2 are utilised in developing the model. The scores attributed to PC1 and PC2 are treated as independent variables, while the 85^{th} percentile speeds are regarded as the dependent variable. The developed model and its statistics are shown in Table 12.

Table 11 PCA results of tangent

Variable	Component	
	PC1	PC2
SW	0.955	
MW	0.890	
AD		-0.866
AT		0.730
ST		0.535
% of total variance	50.53	20.42

It can be observed that the coefficients for PC1 and PC2 are positive. The standard errors for the intercept and both components fall within the specified limits. The value of p is less than 0.05, and hence, all the variables are significant at a 95% confidence interval. The coefficient of determination and adjusted R^2 values are 0.848 and 0.785, respectively.

RESULTS AND DISCUSSION

In multiple regression analysis, one of the assumptions is that the independent variables are not correlated with each other, in other words, there is no multicollinearity among them. The MLR technique was applied to both datasets, including the curved and tangent datasets. The variable inflation factor ranges from 1.34 to 9.36 and 2.86 to 9.66 for the curved and tangent datasets, respectively. It is clearly shown that multicollinearity exists among independent variables. This criterion is not satisfied in both the curved and tangent section data of the present study. Due to multicollinearity, the actual values of the parameters to be estimated cannot be accurately determined. Several techniques are available for mitigating or eliminating

Table 12 Statistical summary of the operating speed model on the horizontal curve

Variables	Coefficient	Standard error	*t*-value	*p*-value	R^2	Adjusted R^2	VIF
Intercept	88.111	0.116	47.23	0.001	0.848	0.785	1.86
PC1	4.053	0.002	2.11	0.001			1.11
PC2	8.271	0.002	4.30	0.002			1.48

the effects of multicollinearity. The most commonly employed techniques using biased estimation results are PCR, ridge regression, and partial least squares regression. In this study, the PCR technique is applied to both curved and tangent section datasets to eliminate multicollinearity.

Using PCR, the variable inflation factor ranges from 1.01 to 1.28 and 1.11 to 1.86 for the curved and tangent datasets, respectively. The VIF values in PCR are less than 4; hence, the criteria are satisfied in this method. The difference in VIF values between MLR and PCR is approximately 7 to 8. Similarly, the standard errors of the MLR coefficients are more than the PCR results. The obtained coefficient of determination (R^2) values are 0.721 and 0.858 in MLR and PCR of the curved section dataset, respectively. Similarly, the obtained coefficient of determination (R^2) values are 0.719 and 0.848 in MLR and PCR of the tangent section dataset, respectively. In both cases, the R^2 values in PCR are higher than those in MLR. This indicates that the PCR model yields better results than the MLR model. Resultantly, PCR is chosen as the most suited prediction model to predict operating speed on horizontal curve and tangent sections of multi-lane highways.

Thus, while conducting multivariate models, it is essential to check for the presence of multicollinearity. To address this issue, a regression method that utilises biased estimation results should be employed. Otherwise, the estimates that must be produced will lead to inaccurate and biased conclusions.

CONCLUSIONS

The primary objective of this study was to compare the effectiveness of PCR and MLR in estimating the operating speed as the dependent variable (Y) against geometric variables as independent variables (Xs) on curve and tangent sections of multi-lane highways. Firstly, statistical tests such as the Z-test and the F-test were conducted to compare the means and variances between big cars and small cars. The results indicated no significant differences; hence, the two vehicle types were combined into a single vehicle category labelled 'Car'. Further, it showed that speed distribution for cars on the curve and tangent sections were different from the normal distribution. It was observed that the car followed a beta 4 distribution on curves and a log-normal distribution on tangent sections. Thereafter, the developed MLR model for the curved section shows that the degree of curvature, deflection angle, and access density are negatively correlated with operating speed, while curve radius, curve length, shoulder, and median width are positively correlated. For tangent sections, the shoulder, median width, and access type exhibit a positive correlation, whereas access density shows a negative correlation with operating speed. In addition, the variable inflation factor is significant for the coefficients of parameters in both models. Hence, it was reported that multicollinearity existed among the independent variables.

To address multicollinearity, PCA was performed. PCA is applied to both curved and tangent section datasets, and the eigenvalues are extracted, from which the principal components are derived. Principal components have the potential to integrate error variance and particular factor variance into the components. It was found that more than 66% and 70% of the total variance was explained by the first two components in the curved and tangent section datasets, respectively. The coefficient of determination values obtained in the PCR model were more than those of the MLR model. Additionally, it was observed that the variable inflation factor in PCR is less than that in MLR. Through the use of orthonormal loadings, the linear independence of the independent variables was verified, demonstrating the absence of multicollinearity.

This study confirms that PCR is preferred over multiple regression analysis for predicting operating speed on curve and tangent sections of multi-lane highways. The present study can be further extended by considering vehicle type, terrain type, environmental factors, and road type to develop operating speed models for both curve and tangent sections. Future studies may also explore other approaches like ridge regression, lasso regression, elastic net regression, and partial least square regression analysis, etc., for better prediction of operating speed on critical sections of highways.

REFERENCES

AASHTO. 2011. *A Policy on Geometric Design of Highways and Streets*, 6th edition. Washington, DC, USA: AASHTO.

Ali, A T, Flannery, A & Venigalla, M M. 2007. Prediction models for free flow speed on urban streets. *Proceedings*, Transportation Research Board 86th Annual Meeting, 21–25 January 2007, Washington DC, USA.

Bassani, M, Catani, L, Cirillo, C & Mutani, G. 2016. Night-time and daytime operating speed distribution in urban arterials. *Transportation research part F: traffic psychology and behaviour*, 42(1):56–69.

De Andrade, G R, Pitombo, C, Cunha, A L N & Setti, J R. 2016. A model for estimating free-flow speed on Brazilian expressways. *Transportation Research Procedia*, 15: 378–388.

Fitzpatrick, K, Blaschke, J D, Shamburger, C B, Krammes, R A & Fambro, D B. 1995. *Compatibility of design speed, operating speed, and posted speed*, report no. FHW A/TX-95/1465-2F. Austin, Texas: Texas Transportation Institute.

Gong, H & Stamatiadis, N. 2008. Operating speed prediction models for horizontal curves on rural four-lane highways. *Transportation Research Record*, 2075(1): 1–7.

Goyani, J, Chaudhari, P, Arkatkar, S, Joshi, G & Easa, S M. 2022. Operating speed prediction models by vehicle type on two-lane rural highways in Indian hilly terrains. *Journal of transportation engineering, Part A: Systems*, 148(3):04022001.

He, Y, Tang, H, Sun, X & Zhao, M. 2010. A study on the operating speed model of passenger cars on long tangent expressway sections. In *ICCTP 2010: Integrated Transportation Systems: Green, Intelligent, Reliable*, 943–949.

Jesna, N M & Anjaneyulu, M V L R. 2016. Reliability analysis of horizontal curves on two lane highways. *Transportation Research Procedia*, 17: 107–115.

Kabacoff, R I. 2011. *Data analysis and graphics with R*. Shelter Island, New York: Manning Publications Co.

Kim, S & Choi, J. 2013. Safety analysis of roundabout designs based on geometric and speed characteristics. *KSCE Journal of Civil Engineering*, 17(6): 1446–1454.

Maji, A, Singh, D, Agrawal, N & Zaman, M. 2020. Operating speed prediction models for tangent sections of two-lane rural highways in Oklahoma State. *Transportation letters*, 12(2): 130–137.

Maji, A & Tyagi, A. 2018. Speed prediction models for car and sports utility vehicle at locations along four-lane median divided horizontal curves. *Journal of Modern Transportation*, 26(4): 278–284.

Medina, AMF & Tarko, A P. 2005. Speed factors on two-lane rural highways in free-flow conditions. *Transportation Research Record*, 1912(1): 39–46.

Morris, C M & Donnell, E T. 2014. Passenger car and truck operating speed models on multilane highways with combinations of horizontal curves and steep grades. *Journal of Transportation Engineering*, 140(11): 1–10.

MRTH. 2022. *Road accidents in India 2022*. New Delhi, India: Ministry of Road Transport and Highways.

Poe, C M & Mason, J M. 2000. Analyzing influence of geometric design on operating speeds along low-speed urban streets: mixed-model approach. *Transportation Research Record*, 1737(1): 18–25.

Shankar, V & Mannering, F. 1998. Modeling the endogeneity of lane-mean speeds and lane-speed deviations: A structural equations approach. *Transportation Research Part A: Policy and Practice*, 32(5): 311–322.

Sil, G, Maji, A, Nama, S & Maurya, A K. 2019. Operating speed prediction model as a tool for consistency based geometric design of four-lane divided highways. *Transport*, 34(4): 425–436.

Tarris, J P, Poe, C M, Mason, J M & Goulias, K G. 1996. Predicting operating speeds on low-speed urban streets: Regression and panel analysis approaches. *Transportation Research Record*, 1523(1): 46–54.

Tottadi, K K & Mehar, A. 2022. Operating speed: review and recommendations for future research. *Innovative infrastructure solutions*, 7(1): 67.

Tottadi, K K & Mehar, A. 2024. Modeling 85th percentile speed differential for evaluating geometric design consistency on two-lane rural highways. *Innovative Infrastructure Solutions*, 9(4): 91.

Tseng, P Y, Lin, F B & Shieh, S L. 2005. Estimation of free-flow speeds for multilane rural and suburban highways. *Journal of the Eastern Asia Society for Transportation Studies*, 6: 1484–1495.

WHO. 2023. *Global status report on road safety 2023*. Geneva, Switzerland: World Health Organization.

Yan, Y, Li, G, Tang, J & Guo, Z. 2017. A novel approach for operating speed continuous predication based on alignment space comprehensive index. *Journal of Advanced Transportation*, 2017: 9862949.