Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 104 (2013) 526 – 532

Procedia
Social and Behavioral Sciences

2nd Conference of Transportation Research Group of India (2nd CTRG)

Method of Identifying Low Performance Vehicles in Heterogeneous Traffic on Two-Lane Highways

Pritam Saha[a,]*, Antaripa Bhadra[a], Nagendra S. Reddy[a], Ashoke Kumar Sarkar[b]

[a]*National Institute of Technology Agartala, Tripura (West)-799055, India*
[b]*Birla Institute of Technology and Science Pilani, Rajasthan-333 031, India*

Abstract

This paper demonstrates a method of identifying low performance vehicles in heterogeneous traffic on two-lane highways. Field data collected adopting video photographic survey technique was used to develop the method. Principal Component Analysis was used in the study to identify low performance vehicles based on sample outlier detection. Two parameters namely percent speed reduction and percent slower vehicles were identified and outliers were subsequently detected in multivariate settings. Around 33 percent of the total data points were detected as outliers and out of which around 29 percent data points correspond to non-motorized vehicles. These outliers represent those vehicles that follow slower impeding vehicles by choice; no attempt of passing manoeuvre is usually made even if there is an opportunity to pass or dynamic characteristics of the vehicles do not permit them to perform passing manoeuvre. These vehicles in turn affect the traffic performance and therefore considered as low performance vehicles.

Keywords: Two-lane highways; heterogeneous traffic; low performance vehicles; Principal Component Analysis

1. Introduction

In India, road traffic is mostly described as heterogeneous composing wide spectrum of vehicle types in terms of static and dynamic characteristics. Unlike homogeneous traffic stream, the heterogeneous behaviour is commonly characterized by lack of lane discipline and increased risk ability of driver population. Large speed differential of vehicles results frequent car-following interaction and thus causes impedance to faster vehicles.

* Corresponding author. Tel.: +0-943-676-7170; fax: +0-381-234-6360.
E-mail address: saha.pritam@gmail.com

doi:10.1016/j.sbspro.2013.11.146

However, the speed differential depends on the proportion of low performance vehicles in traffic composition. In most of the cases, these vehicles create impedance to faster vehicles in the traffic stream and thus affect quality of service. This is more critical on two lane highways where interaction occurs in both direction flows. The current research, therefore, is aimed at developing a method of identifying such low performance vehicles.

Principal Component Analysis (PCA) was used in the present study in identifying low performance vehicles based on sample outlier detection. Two parameters: percent speed reduction and percent slower vehicles were used in the analysis and the necessary data was collected from the field. Thus, based on the field data collected on NH-44, a national highway passing through the state of Tripura, India, these two parameters were computed and accordingly applied in PCA. To facilitate visualization of outliers, Score, Influence, Leverage and Hotelling T2 plots were contemplated and the data points that are considerably dissimilar or inconsistent were accordingly identified. The analysis was made with 5 percent significance level. Around 33 percent of the total data points were detected as outliers and out of which around 29 percent data points correspond to non-motorized vehicles. The drivers behaviour and vehicular characteristics (static and dynamic) are the major attributes of low performance vehicles in the traffic stream.

The principle component analysis was deemed to be very promising technique in making a perceptible contribution to the detection of low performance vehicle from heterogeneous traffic composition. Two positively correlated parameters: percent speed-reduction and percent slower vehicles were used in the analysis. The hypothesis of the study was 'percent speed-reduction should increase with a simultaneous increase of the proportion of slower vehicles in the traffic stream'. Disagreement, however, indicates a paradoxical nature of the observed data points, thus could be treated as outliers. The statistical plots that were obtained based on the analysis were investigated in order to detect anomalous sample data. The detected outliers eventually identified those vehicles, speed of which is insensitive to the presence of slower vehicles in the traffic mix. Sensibly, these outliers represent those vehicles that follow slower impeding vehicles by choice; no attempt of passing manoeuvre is usually made even if there is an opportunity to pass or dynamic characteristics of the vehicles do not permit them to perform passing manoeuvre. These vehicles in turn affect the traffic performance and considered as low performance vehicles. Thus, principal component analysis could be considered as a plausible and powerful tool to identify these vehicles in heterogeneous traffic stream.

2. Review of literature

Mining information from a dataset containing multiple dimensions is commonly observed (Aggarwal & Yu 2001). Hence, this is not trivial when a dataset contains multiple outliers. Even a small percentage of outliers can distort the results and render the outcome misleading or useless. To overcome this problem, Alqallaf et al. (2002) proposed robust methods to estimate key parameters such as the mean and covariance/correlation matrix without the negative effect of outliers. Few techniques such as the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD), however, have proven their robustness but are limited to small moderate dimensions (Hubert et al. 2005; Egan & Morgan 1998). Yang and Trewn (2004) suggested the use of multivariate model as it provides a way for engineers and manufacturers to test their products in an environment that provides many advantages over univariate models. They were, however, in accord with the fact that multivariate quality control is inherently more complex than univariate statistical process control. Nevertheless it would be a more realistic representation of the data since the real world practice does not usually have only one variable that is measured independent of all other variables in a system. Montgomery (2005) also suggested that the univariate approaches may lead faulty results since it takes little or no account of the covariance that exists between the observations.

In a study on outlier detection for high dimensional data, Aggarwal and Yu (2001) observed that the process of finding meaningful outliers becomes inherently more complex for multivariate data. This could be exemplified by a dataset obtained from a food processing unit that contains hundreds of uncommon dimensions. Hubert et al.

(2005) proposed the use of projection pursuit (PP) and principal component analysis (PCA) to process and analyze large information of these types of applications. Ben-Hur and Guyon (2003) also demonstrated the use of PCA to extract features relevant to cluster analysis. Stefatos and Hamza (2007) introduced the concepts of principal component analysis to detect multiple outliers in high-dimensional datasets. They suggested, as well, that the proposed algorithm is computationally fast and robust to outlier detection. In a fairly recent study, Saha et al. (2009) demonstrated the utilization of PCA as a means for outlier detection.

In context to the present study, the dataset obtained from heterogeneous traffic contains considerable proportion of outliers because of different vehicular and driver's characteristics. The above literature has made it quite evident that PCA could be used as a robust and effective statistical tool to detect those outliers in multivariate settings. PCA was, thus, applied on the unsupervised traffic dataset in order to detect outliers while identifying the low performance vehicles in the traffic stream on two-lane highways.

3. Interpretation of the PCA plots

PCA operates in an unsupervised manner and is used to analyze the inherent structure of the data. It helps in dimension reduction of the data set by finding an alternate set of coordinates, known as the principal components (Esbensen 2005; Martens & Naes 1989). The influence of the presence of outliers in the dataset could be well interpreted using the following complementary sets of attributes.

3.1. Score plot

The derived principal components together define a plane into the k-dimensional space. This plane is a window into the multidimensional space that can be graphically visualized. When observed data are projected onto this, new co-ordinate of the data points is generated which is termed as scores. The plotting of such projected configuration is known as score plot. Scores describe the data structure in terms of sample patterns and more generally show sample differences or similarities. Each sample has a score on each PC. It reflects the sample location along that PC and is the coordinate of the sample on the PC. Data points with close scores along the same PC are similar whereas the data points are considered to be quite different from each other if the scores differ greatly.

3.2. Influence plot

The influence plot displays the samples residual X-variances against leverages. Residual X-variance is the proportion of X-variance not explained by the PC taken into account. Leverage is the distance from the projected sample to the mean of the dataset. Samples with high leverages have a stronger influence on the model than other samples. They may or may not be outliers, but they are influential. Samples with high residual variance are likely outliers. However, an influential sample with high residual is the worst case and considered as dangerous outlier. It can easily be detected from an influence plot.

3.3. Hotelling T^2 plot

This plot displays the Hotelling T^2 statistics for each sample as a line plot. The statistic gives a measure of significant variation of the dataset. It is obtained by the sum of normalized squared scores divided by their variance. The statistical threshold is calculated using the F-distribution with typical 5 percent significant level. This plot could be used to detect outliers identifying the statistics outside the threshold.

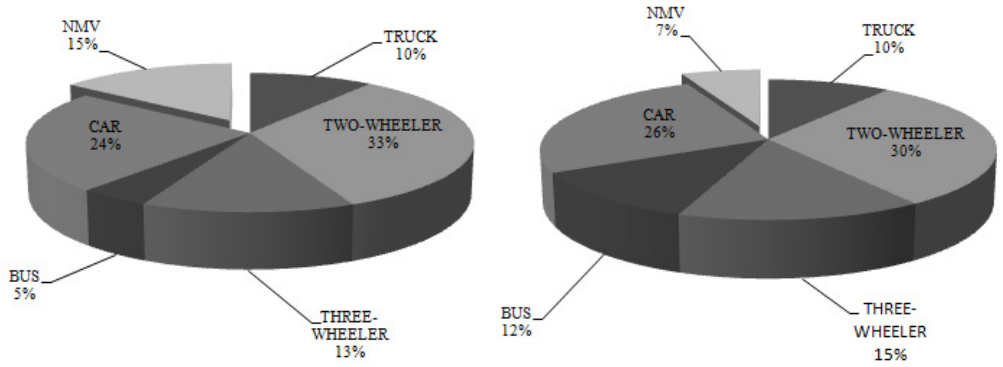
4. Data mining

The rapid economic growth has resulted in the entry of large number of fuel efficient and high engine powered new generation cars/vans into the Indian market during past few years. The dynamic characteristics of these new technology vehicles help in achieving higher speeds. Thus the prevailing road traffic exhibits a real mixed traffic situation as the same road space is shared by a wide range of vehicle category including para transit modes like motorized three wheeler, paddle tricycle etc. This consequences large speed differential in the prevailing heterogeneous traffic stream. However, to simplify the analysis the observed vehicle types were grouped into six categories: car, bus, truck, three-wheeler, two-wheeler and non-motorized vehicle (NMV). Video photographic survey technique was adopted while conducting field study and representative sample data was collected to obtain statistically reliable results (Table 1). *Fig. 1* illustrates the traffic composition of the observed data contemplating both directions traffic separately. The composition exhibits significant proportion of three-wheeler and non-motorized vehicles.

Table 1. Sample traffic data collected on selected highway section

Type of vehicles	Time in	Time out	Lapsed time	Speed (Kmph)	FFS[r]	SR[ϕ]	PSR [ϵ]	PSV[¶]
Truck	00:15:12	00:16:02	00:00:50	36.00	60.21	24.21	40.21	
Truck	00:15:51	00:16:33	00:00:42	42.86	60.21	17.35	28.82	0.00
Two-wheeler	00:15:56	00:16:45	00:00:49	36.73	58.61	21.88	37.32	
Two-wheeler	00:16:11	00:17:07	00:00:56	32.14	58.61	26.47	45.16	0.00
Three-wheeler	00:16:40	00:17:18	00:00:38	47.37	55.52	8.15	14.68	
Bus	00:17:15	00:18:13	00:00:58	31.03	64.21	33.18	51.67	
Truck	00:17:20	00:18:07	00:00:47	38.03	60.21	21.91	36.39	
Truck	00:17:23	00:18:08	00:00:45	25.00	60.21	20.21	33.57	33.00
Car	00:17:32	00:18:26	00:00:54	33.33	67.95	34.62	50.94	
Car	00:17:41	00:18:29	00:00:48	37.50	67.95	30.45	44.81	
Truck	00:17:43	00:18:30	00:00:47	24.30	60.21	21.91	36.39	
Car	00:18:00	00:18:58	00:00:58	31.03	67.95	36.92	54.33	
.....	....	....	....	....	....	....	....	....
.....	....	....	....	....	....	....	....	....

[r]Free Flow Speed; [ϕ] Speed reduction; [ϵ] Percent Speed-reduction ;[¶] Percent Slower-vehicles



(a) (b)

Fig. 1. Observed traffic composition of the data used in PCA: (a) west bound traffic; (b) east bound traffic

Field data were processed to measure the proposed parameters, percent speed-reduction (PSR) and percent slower-vehicles (PSV). The space mean speed of different vehicle types was computed using the traffic data extracted from the video photographic survey at different flow levels. At the same time, free-flow speed of different vehicle types was also calculated in this analysis by developing speed-flow relationship. The free-flow speed of different vehicle types was used to determine the percent speed-reduction (PSR). The limiting speed of slower vehicles causing impedance to faster vehicles in the traffic stream was found to be about 26 kmph. The number of vehicles moving at or below this speed at any instant of time was recorded and expressed in terms of percentage (PSV). These two variables were accordingly used in principal component analysis (PCA) and outliers were subsequently detected in multivariate settings.

5. Analysis and results

PCA was performed over the data set of proposed parameters: percent speed-reduction (PSR) and percent slower-vehicles (PSV), with the aim of observing the inherent structure of the collected field data in terms of similarities and differences. Two principal components were derived for the observed dataset and a window into the two-dimensional space was thus formed. The projected configuration of the data points, termed as score plot is illustrated in *Fig. 2*. The plot identifies the observations that are far from the mean and do not fit the PCA model. These observations were accordingly treated as outliers. *Table 2* reflects the proportion of outliers observed respectively for motorized and non-motorized vehicles. The PC1 (principal component) accounts for 77 percent dispersion of the dataset and PC2 accounts for rest 23 percent. Such dispersion is quantified as variance and more specifically termed as explained variance. At the same time, another complimentary set of plots were developed as well to further inspect the outliers detected by the score plots. *Fig. 3* describes the influence plot of the observed data where samples with high residual variance and leverages were identified as outliers. *Fig. 4* demonstrates the Hotelling T^2 plot, considered as an alternative way of plotting sample leverages. The statistical threshold was calculated for the dataset imported to the software, the *Unscrambler* using the F-distribution with 5 percent significant level and displayed as a red line (Fig. 4). The samples were detected as probable outliers when the statistics cross the red line.

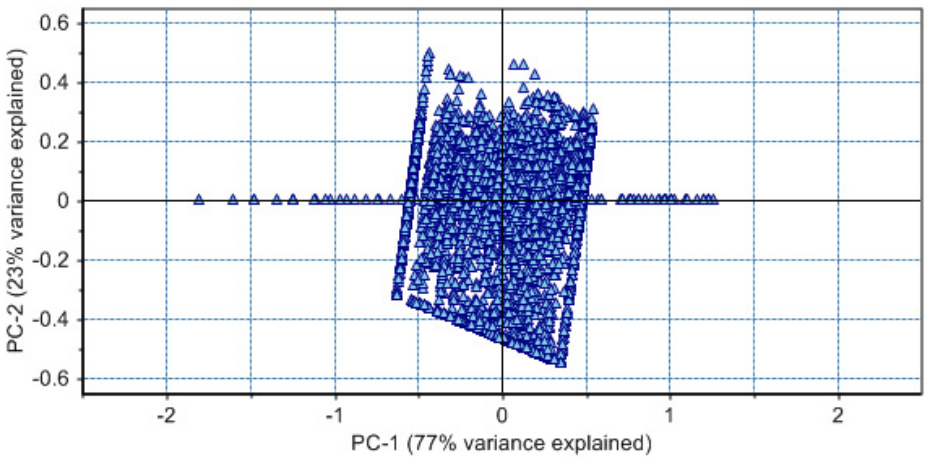


Fig. 2. PCA-Score plots of the dataset: analysis performed on the proposed parameters

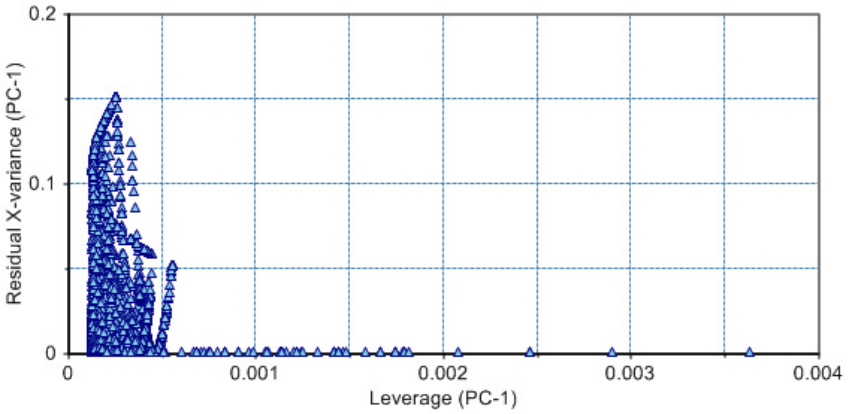


Fig. 3. Influence plot displaying sample residual X-variances (taking PC-1 into account) against leverages

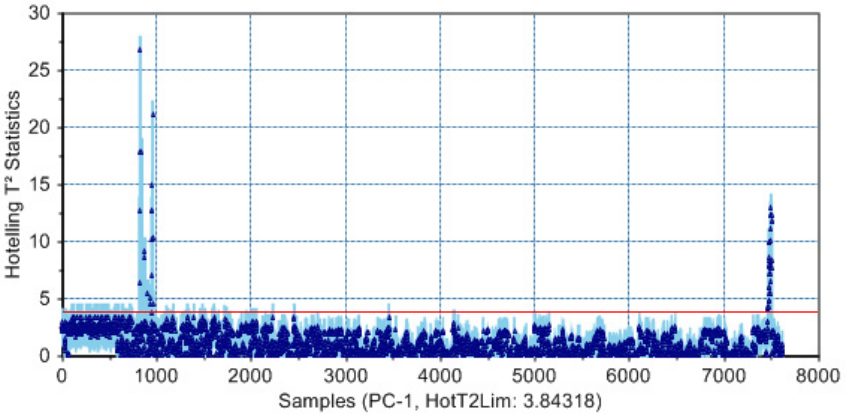


Fig. 4. A line plot displaying the Hotelling T² statistic for each sample in the dataset of the proposed parameters

Table 2 provides the outlier details that indicate majority of the detected outliers correspond to non-motorized vehicles. The reason is 'percent speed-reduction' of these vehicles do not increase with a simultaneous increase of the 'proportion of slower vehicles' in the traffic stream and thus makes a paradox in the dataset. Around 33 percent of the total data points were detected as outliers and out of which around 29 percent data points correspond to non-motorized vehicles. Accordingly, the study results make it clear that performance comparison considering the entire traffic with reference to a common scale is impractical.

Table 2. Vehicle category wise detected outliers from PCA

Vehicle category	Sample size	Outliers detected (%)	Sample size	Outliers detected (%)
Bicycle & Tricycle	-	-	845	28.89
Car	6779	2.02	-	-
Bus		0.26	-	-
Truck		0.67	-	-
Two-wheeler		1.29	-	-
Three-wheeler		0.36	-	-

6. Conclusions

The parameters identified in the present study: percent speed-reduction and percent slower vehicles were used in the analysis on account of positive correlation among them. The premise of the present study is based on the hypothesis that 'percent speed-reduction should increase with a simultaneous increase of the proportion of slower vehicles in the traffic stream'. Accordingly, principal component analysis (PCA) was performed in order to identify the disagreement that indicates paradox of the observed data points while identifying the outliers. The analysis provided a number of plots including Scores, Influence and Hotelling T^2 to visualize and detect the outliers. The detected outliers eventually identified those vehicles, speed of which is insensitive to the presence of slower vehicles in the traffic mix. Logically, these identified outliers represent those vehicles that move in following slower impeding vehicles by choice; no attempt of passing manoeuvre is usually made even if there is an opportunity to pass or dynamic characteristics of the vehicles do not permit them to perform passing manoeuvre. They consequently affect the traffic performance and thus considered as low performance vehicles. The principal component analysis, therefore, could be considered as a plausible and powerful tool in identifying low performance vehicle from heterogeneous traffic composition.

Another significant outcome of the analysis was detection of substantial proportion of non-motorized traffic as outliers. This is attributed to the fact that speed reduction of these vehicles is infinitesimal even when the proportion of slower vehicles is large. This in turn disagrees with the data pattern wherein speed reduction is usually observed high. Clearly, the study results indicate that performance comparison considering the entire traffic with reference to a common scale is impractical.

References

Aggarwal, C. C. & Yu, P. S. (2001). Outlier detection for high dimensional data. *Proc. ACM SIGMOD.*

Alqallaf, F. A., Konis, K. P. & Martin, R. D. (2002). Scalable robust covariance and correlation estimates for Data Mining. *Proc. ACM SIGKDD*.

Ben-Hur, A. & Guyon, I. (2003). Detecting Stable Clusters Using Principal Component Analysis. *In Functional Genomics: Methods and Protocols. Brownstein, M. J. & Kohodursky, A. (eds.) Humana press,* 159-182.

Egan, W.J. & Morgan, S.L. (1998). Outlier detection in multivariate analytical chemical data. *Analyical Chemistry, 70*, 2372-3279.

Esbensen, K. H. (2005). Multivariate Data Analysis - In Practice. *5th Edition, CAMO Process AS, Esbjerg*, Denmark.

Hubert, M., Rousseeuw, P. J., & K. V. Branden. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics, 47*, 64-79.

Martens, H. & Naes, T. (1989). Multivariate Calibration. *Wiley, Chichester*, England.

Montgomery, D.C. (2005). Introduction to Statistical Quality Control. *John Wiley & Sons*.

Saha, B. N., Ray, N. & Zhang, H. (2009). Snake Validation: A PCA-Based Outlier Detection Method. *IEEE Signal Processing Letters, 16,* 549-552.

Stefatos, G & Ben Hamza, A. (2007). Cluster PCA for Outliers Detection in High-Dimensional Data. *IEEE*, 3961-3966.

The Unscrambler X 10.2. *http://www.camo.com*

Yang K. & Trewn, J. (2004). Multivariate Statistical Methods in Quality Management. *Mc Graw Hill Professional*.