Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 107 – 114

Procedia Computer Science

The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)

Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data

Pritam Saha[a]*, Nabanita Roy[a], Deotima Mukherjee[a], Ashoke Kumar Sarkar[b]

[a]*Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711103, India*
[b]*Birla Institute of Technology and Science Pilani, Rajasthan-333031, India*

Abstract

Level-of-service (LOS) measures of two-lane highways exhibit incompatibility if the prevailing traffic is heterogeneous in character. Thus, such traffic warrants development of LOS criteria on the basis of compatible measures which capture its characteristics. The present paper has suggested the use of percent speed-reduction and percent slower vehicles, as the measures of performance, while defining LOS criteria. Defining such criteria is basically a classification problem and clustering could be applied as an effective technique for its solution. However, heterogeneity in the traffic mix results in the presence of significant proportion of outliers in the data set, which can distort the results and render into misleading or useless outcomes. The study considers principal component analysis to be an efficient technique in detecting outliers from the data set and accordingly applies it on the proposed LOS measures. An iterative process, adopted for removing outliers, indicates that significant proportion of outliers comprises of non-motorized traffic data; this accordingly ensures reliability of the data set. The study concluded the unfeasibility of LOS assessment of the entire traffic, considering both motorized and non-motorized modes, with respect to a common scale.

Keywords: Two-lane-highways; level-of-service; performance measures; heterogeneous traffic data; outlier detection; Principal Component Analysis

* Corresponding author. Tel.:91-9831942049.
E-mail address: saha.pritam@gmail.com

doi:10.1016/j.procs.2016.04.105

1. Introduction

Highway Capacity Manual (HCM)[1] introduced the concept of level-of-service (LOS) for assessing the performance of traffic and provides criteria for it at varied operating conditions. Defining such criteria is basically a classification problem and clustering could be applied as an effective technique for its solution. Clustering is basically an unsupervised approach that classifies the observed data set, under normal and extreme variations, based on certain distance measures. The different vehicular and driver's characteristics in heterogeneous traffic, however, results few observations to considerably differ from the remainders. These are termed as outliers of the dataset, which consequently affect the data clustering. Several researchers have explained the concepts of applying principal component analysis (PCA) as a robust method in detecting these outliers in multivariate settings. The present study, thus, demonstrates the use of PCA as a pre-processing statistical tool to secure a reliable data for clustering in order to define LOS criteria.

The assessment of LOS, on two-lane highways, is a complicated issue because of their unique operational characteristics. HCM provides a method for assessing performance of such highways based on three measures: percent time-spent-following (PTSF), average travel speed (ATS) and percent free flow speed (PFFS). However, a number of researchers have shown that the method is not totally compatible with the highways having heterogeneous traffic with large speed differential; this results in frequent car-following interaction and formation of platoons. Slower vehicles, thus, cause impedance to faster vehicles and compel them to disobey lane discipline and take considerable amount of risk while passing. Keeping this fact in view, the present study identified two major attributes that are responsible for platooning; they are the variation of free-speed characteristics of different types of vehicles: percent speed-reduction (PSR) and, the limiting speed that would differentiate a slower vehicle from other vehicles plying on the highway: percent slower vehicles (PSV).

A real mixed-traffic situation with various kinds of modes, both motorized and non-motorized, is observed on a rural highway, when it approaches a city or a town. The present study, consequently, selected a two-lane highway section, on a National Highway in India, close to a city. The PCA was considered to be a very promising technique in making a perceptible contribution to the detection of outliers from the observed data set and was accordingly applied on the proposed LOS measures. 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 inconsistency in the observed data points, thus compelling them to be treated as outliers. Therefore, the present study aimed at investigating the statistical plots of PCA to detect such anomalous sample data. The outliers thus detected, would represent those vehicles, the speed of which is insensitive to the presence of slower vehicles in the traffic mix.

2. Review of literature

Assessment of traffic performance on highways is imperative, while taking decisions of investments made at different stages of design life. The second edition of the Highway Capacity Manual (HCM)[2] introduced the concept of level of service (LOS) with the aim of expressing highway performances. The HCM LOS measures, however, have undergone significant changes with the advent of new generation vehicles[1-4]. Introduction of such vehicles considerably affects traffic performance, particularly on two-lane roads, because of both directional movement of traffic; this attributes to the formation of frequent platoons and consequent delay. There have been a number of researchers who suggested use of different LOS measures considering speed, platooning and passing [5-7] and also, investigated the relationship of these measures with flow parameters [8-10]. However, studies indicate that the majority of the measures exhibit incompatibility when the traffic is heterogeneous in character and speed differential is quite high. This is prevalent in most of the developing countries including India, where average speed ranges from 100 to 10 km/h on two-lane highways[11]. Couple of fairly recent studies conducted in India proposed the use of alternative measures that take the heterogeneity effect into account[12, 13] and suggested ensuring reliability of the data set, before using them in defining LOS criteria.

Thus, there is a need to apply an outlier detection technique to identify those data points, which are very different from the rest of the data set. Even a small percentage of outliers can distort the results and render the outcome, misleading or useless. This warrants the need of proposing robust methods, which would account the negative effects of outliers, as classical methods of data mining are not reliable if the data set contains outliers[14]. Accordingly, several researches suggested applications of various data-mining algorithms aimed at finding such outliers while clustering data set [15-21]. Techniques such as minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) have proven their robustness; however, their applications are limited to small moderate dimensions[22, 23]. A study 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[24]. The study results also agreed to the fact that multivariate quality control is inherently more complex than univariate statistical process control. At the same time, 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. Moreover, univariate approaches may lead to faulty results, because it takes little or no account of the covariance that exists between the observations[25]. A study on outlier detection for high dimensional data indicates that the process of finding meaningful outliers becomes inherently more complex for multivariate data[21]; this could be exemplified by a dataset, obtained from a food processing unit, which contains hundreds of uncommon dimensions. Principal component analysis (PCA) could be applied in processing and analysing such multivariate data[22, 26, 27]. Couple of recent studies demonstrated the concepts of PCA to detect multiple outliers in high-dimensional datasets [28, 29]; this algorithm is computationally fast and robust in detecting outliers.

This paper exhibits dataset obtained from heterogeneous traffic, containing considerable proportion of outliers; this attributes to the fact that wide range of vehicular and driver's characteristics make the traffic data paradoxical in nature. Most of the above literature made it evident that PCA could be used as a robust and effective statistical tool to detect those outliers in multivariate settings. The present study, therefore, applied PCA in detecting outliers and performing quality control subsequently on the unsupervised traffic dataset. The study also apprises the iterative method adopted in the process of eliminating outliers and securing a data set to produce a reliable cluster.

3. Proposed method: Principal component analysis

Principal component analysis (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[30, 31]. The analysis in the present research was made with the aid of statistical software. As an initial input, a matrix X, with N observations and K variables, is needed to be imported to the software. The dimension of variable space is considered same as the number of variables present in the dataset; each variable represents one co-ordinate axis. Each observation of the X-matrix is placed in the K-dimensional variable space and a swarm of points in this space is formed. First Principal Component (PC1) is the line in the K-dimensional space that passes through the average of the data points and best approximates the data in the least square sense. Generally, the first principal component accounts for maximum of the total variance in the observed variables. However, one principal component is insufficient to model the systematic variation of a data set. A second Principal Component (PC2) is thus calculated and represented by a line, orthogonal to PC1 in k-dimensional space. This line also passes through the average point and improves the approximation of the X-data to the possible extent. *Figure 1* illustrates graphical representation of the principal components that pass through a swarm of data points in a k-dimensional space[32].

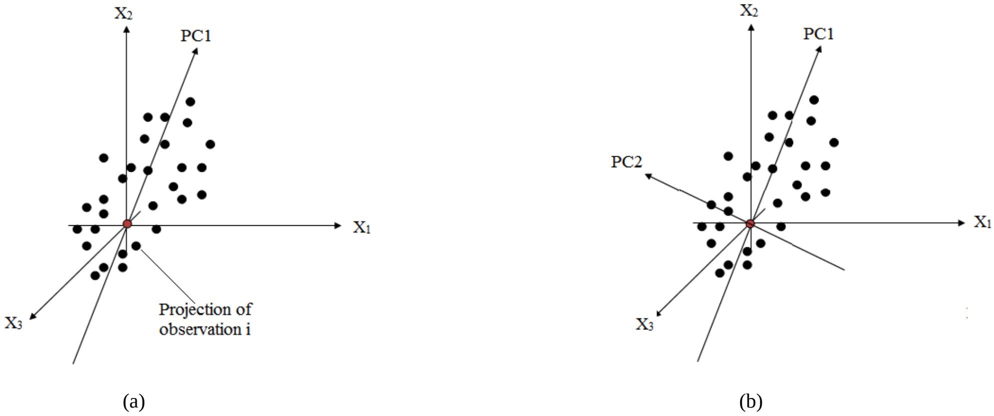


Fig. 1. Graphical representation of the principal components through a swarm of data points in a k-dimensional space. (a) First Principal Component: PC1 (b) Second Principal Component: PC2 orthogonal to PC1

4. Study design

4.1 Approach

The current research is based on the premise that the field data collected from heterogeneous traffic stream contains sizable percentage of outliers that can eventually distort the results and render the outcome misleading or useless. Therefore, the research utilizes PCA as a mean in detecting outliers of the observed dataset. Also, the detected outliers were eliminated from the dataset through an iterative process in order to make it free from outliers. Specifically, PCA was applied on heterogeneous traffic data as a pre-processing step prior to clustering. The analyses in this research were conducted using field data collected from a two-lane highway section in the state of Tripura, India. Eight segments were selected, along a 20 km highway section that approaches Agartala, the capital city of the state, for conducting the study. Thus, a real mixed traffic situation was observed on almost all the study sites.

4.2 Study sites and field data

During the past few decades, the Indian market has experienced an entry of large number of fuel efficient and high engine powered new generation cars/vans; this consequences a wide variety in the prevailing traffic, both in terms of static and dynamic characteristics. They share the same road space and thereby, exhibit heterogeneity in the traffic mix because of large speed differential among them. The present study categorized the observed traffic as car, bus, truck, three-wheeler, two-wheeler and non-motorized vehicle (NMV). While collecting representative field data, a 500 m long road segment was selected and two reference lines were marked on the pavement. Video photographic survey technique was considered appropriate in order to obtain statistically reliable results. Four video cameras were installed on either side of the reference lines; two for each direction, in order to record the entry and exit of vehicles. Video files obtained from the four specified locations were then played on a computer to extract the necessary readings i.e. vehicle type, registration number and time at entry and exit points of the trap. Composition of traffic was observed separately for both directions and proportion of two wheelers and NMV was found to be significant: about 30 percent two wheelers and 10 percent NMVs respectively. Truck and car traffic shared about 10 and 25 percent of the overall composition of traffic.

4.3 Data processing

On the basis of empirical investigations, the present study identified two major attributes that are responsible for platooning of traffic stream, thereby affecting the LOS. They are the variation of free-speed characteristics of different types of vehicles and the limiting speed that would differentiate a slow-moving vehicle from other vehicles plying on the highway. Accordingly, percent speed-reduction (PSR) and percent of slower-vehicles (PSV) in the traffic stream were selected as the performance measures to define the LOS criteria[13].

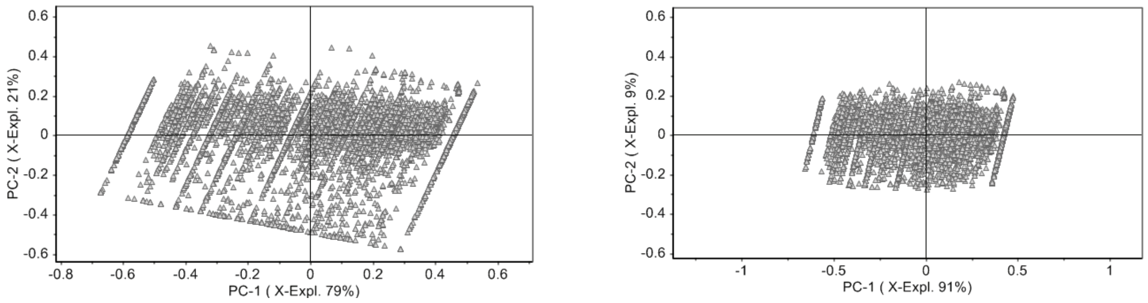
Field data, collected at the study sites, was analyzed to estimate these measures at various flow scopes. Space mean and free flow speeds of different vehicle categories were assessed for determining PSR. Further, analysis of speed data indicates that limiting speed of slower vehicles, causing impedance to faster ones, is about 27 kmph. Those who move at or below this speed, at any instant of time, was recorded and expressed in terms of percentage (PSV). PCA was applied on these positively correlated measures aimed at detecting outliers in multivariate settings. *Table 1* shows the statistical details of the proposed performance measures and an examination of the statistics reveals that coefficient of variation is considerably high, for both the variables, signifying the presence of sizable proportion of outliers in the data set.

Table 1. Detailed statistics of the performance measures used in PCA

Statistical Parameter	Percent-speed-reduction	Percent-slower-vehicles
Maximum	0.924	1.000
Minimum	0.000	0.000
Mean	0.454	0.556
Standard Deviation	0.197	0.320
Coefficient of variation	0.434	0.576

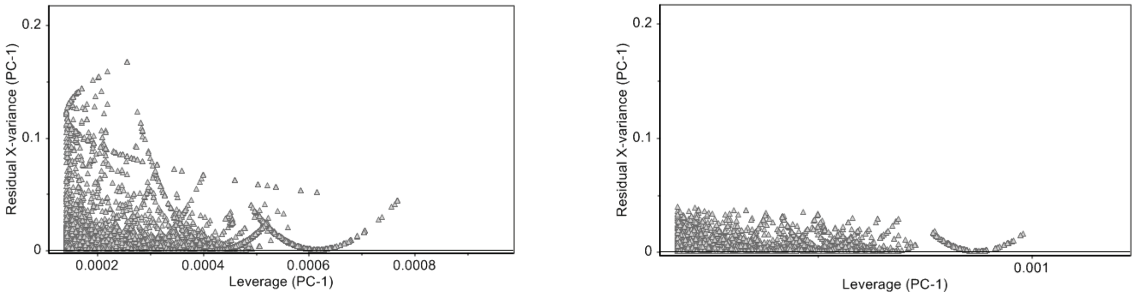
5. Analysis and Results

PCA was applied on the data set of PSR and PSV with the aim of establishing the inherent structure of the field data in terms of their similarities and dissimilarities. Projections of data points onto the principal components (PC1 and PC2), obtained on the basis of least square sense, create new co-ordinates of the data points: they are termed as scores. *Fig. 2* indicates the score plots derived from the projected configuration of the data points. The observations that are far from the mean were identified and accordingly treated as outliers. Since these outliers have significant effect on data clustering, an initiative was taken as the first approach to eliminate such outliers from the dataset. An iterative process was followed for outlier detection and successive elimination. *Table 2* reflects the proportion of outliers observed at each stage of iteration. The proportion of the total variance in the observed variables that the principal component takes into account increases for first PC with a parallel decrease of the second one as the iterative process progresses. This implies a gradual accomplishment of high degree of certainty as the outliers are eliminated from the dataset. However, no variation in the explained variance was observed beyond fourth iteration and proportion of detected outliers was also insignificant. Another complimentary set of plots was also developed to further inspect the outliers, detected by those score plots. *Fig. 3* displays the influence plot of the observed data wherein samples with high residual variance and leverages were identified as outliers. Further, a line plot displaying the Hotelling T^2 statistic was considered as an alternative way of plotting sample leverages. The statistical threshold was calculated for the dataset, using the F-distribution with 5 percent significant level, and displayed as a red line. Outliers were, thus, detected when the statistics cross the red line (Fig. 4). An interpretation of the plots, thereby, elucidates gradual increase of reliability of the dataset during the iterative process. This could be substantiated by the leverage and residual x-variance scales which were observed to decrease with the successive elimination of the detected outliers.



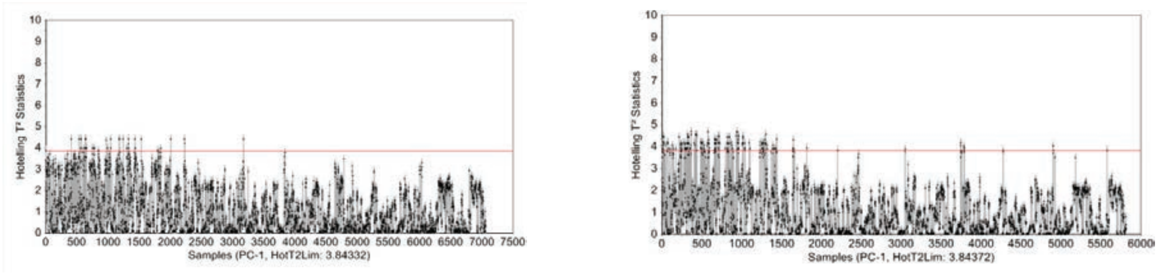
(a) (b)

Fig. 2. PCA-Score plots of the dataset before and after eliminating outliers



(a) (b)

Fig. 3. Influence plots displaying sample residual X-variance (taking PC-1 into account) against leverages of the dataset: before and after eliminating outliers



(a) (b)

Fig. 4. A line plot displaying the Hotelling T² statistic for each sample in the dataset of the dataset: before and after eliminating outliers

Table 2. X-explained variance values obtained from PCA scores analysis for performance measures and the observed outliers

Iteration	Sample Size		Observed outliers (%)		X-explained variance	
No.	MV*	NMV†	MV	NMV	PC1@ (%)	PC2 (%)
1	6465	600	1.87	47.00	79	21
2	6344	318	1.73	54.40	86	14
3	6234	145	1.89	28.29	90	10
4	6116	104	0.28	16.35	91	9
5	6099	87	0.15	5.75	91	9
6	6090	82	5.37	2.44	91	9

† Non-motorized vehicles; * Motorized Vehicles; @ Principal Components

Table 3. Vehicle category wise detected outliers from PCA

Iteration No.	Car (%)	Bus (%)	Truck (%)	2-Wheeler (%)	3-Wheeler (%)	NMV† (%)
1	0.66	0.07	0.30	0.58	0.23	47.00
2	0.59	0.07	0.17	0.69	0.18	54.40
3	0.44	0.06	0.20	0.91	0.35	28.29
4	0.13	0.00	0.00	0.13	0.02	16.35
5	0.06	0.02	0.00	0.04	0.02	5.75
6	1.59	0.41	1.33	1.74	0.70	2.44

† Non-motorized vehicles

Table 3 provides the detected outlier details based on vehicle category and indicates that the proportion of non-motorized vehicles is significant in the detected outliers; this attributes to the fact that speed of these vehicles remains unaltered even if the proportion of slower vehicles increases in the traffic stream. Accordingly, performance comparison considering the entire traffic, with reference to a common scale, is impractical because of such paradox in the dataset. Thus, this is quite evident from the present study that the data obtained from automobile mode should be taken into consideration at the time of data clustering. The recent edition of Highway Capacity Manual[4] also, suggested the use of automobile and bicycle mode separately for performance assessment. The manual also introduced a different performance measure, 'bicycle level of service (BLOS)', based on traveler's perception model for the performance assessment of such mode.

6. Conclusions

Clustering is one of the most efficient classification techniques and could be effectively applied to the data, mined for the proposed performance measures. However, in order to produce a reliable cluster of these data, it is imperative to ensure that the data set is free from outliers. Thus, the cluster analysis involves a pre-processing step to identify the existence of outliers within a data set. The present study applies the concept of principal component analysis (PCA), a statistical tool, at this step. The proposed level-of-service (LOS) measures: percent speed-reduction (PSR) and percent slower vehicles (PSV) were used in the analysis and PCA generated a number of plots including Scores, Influence and Hotelling T^2 to visualize and detect the outliers. The iterative process revealed an increase of the variance explained by the PC1, whereas a simultaneous decrease was observed in case of PC2 as the process progresses. This change was insignificant after a number of iterations, however, outlier detection process and subsequent removal from the data set was continued. This indicates an improvement of reliability of the field data. The Scores plots demonstrated an ascending improvement of reliability during the initial iterations which, however, becomes insignificant in subsequence. This finding was further supplemented by the Influence and Hotelling T^2 plots. Moreover, the study revealed that significant proportion of outliers is for non-motorized traffic data; this attributes to the fact that reduction of speed is insignificant for these vehicles even if they are in substantial proportion. Accordingly, they exhibit paradox in the data pattern wherein speed reduction is usually observed high. The study results, thus, made it evident that LOS assessment considering the entire traffic, with reference to a common scale, is unfeasible and suggests the use of automobile data for the proposed performance measures. The recent edition of *Highway Capacity Manual*[1] also, indicates the use of automobile mode separately for LOS assessment.

References

1. Highway Capacity Manual, TRB, National Academics, Washington, D.C., 2010.
2. Special Report 87: Highway Capacity Manual, 2nd ed. HRB, National Research Council. Washington, D.C.,1965.
3. Special Report 209: Highway Capacity Manual, 3rd ed. (1997 update). TRB, National Research Council. Washington, D.C.,1985.
4. Highway Capacity Manual, TRB, National Research Council. Washington., D.C., 2000.

5. Van As SC, Van Niekerk A. The operational analysis of two-lane rural highways. Proc., 23rd Southern African Transport Conf. (SATC 2004), National Institute for Transport and Road Research (NITRR), CSIR, Pretoria, South Africa; 2004.
6. Al-Kaisy A, Durbin C. Evaluating new methodologies for estimating performance on two-lane highways. Can. J. Civ. Eng. 2008; 35(8), 777–785.
7. Al-Kaisy A, Karjala S. Indicators of performance on two-lane rural highways. Transportation Research Record 2008; 2071, 87 –97.
8. Van As C. The Development of an Analysis Method for the Determination of Level of Service on Two-Lane Undivided Highways in South Africa, Project Summary. South African National Roads Agency, Limited,Pretoria; 2003.
9. Hashim IH, Abdel-Wahed TA. Evaluation of performance measures for two lane roads in Egypt. Alexandria Eng. J. 2011; 50, 245–255.
10. Moreno AT, Lorca C, Sayed T, Garcia A. Field evaluation of traffic performance measures for two-lane highways in Spain. 93rd Transportation Research Board Annual Meeting, Transportation Research Board, Washington, DC. 2014; Paper 14-0847.
11. Dey PP, Chandra S, Gangopadhyay S. Speed studies on two-lane Indian highways. Indian Highways: 7 Journal of the Indian Roads Congress 2008; Vol. 36, No. 6,pp. 9-20.
12. Penmetsa P, Ghosh I, Chandra S. Evaluation of Performance Measures for Two-Lane Intercity Highways under Mixed Traffic Condition. Journal of Transportation Engineering, ASCE 2015; ISSN 0733-947X/04015021
13. Saha P, Sarkar AK, Pal M. Assessment of Level-of-Service of Two-Lane Highways with Heterogeneous Traffic. Transportation Research Board 94th Annual Meeting (No. 15-2723). Transportation Research Board, Washington, DC; 2015.
14. Alqallaf FA, Konis KP, Martin RD. Scalable robust covariance and correlation estimates for Data Mining. Proc. ACM SIGKDD; 2002.
15. Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS. Fast algorithms for projected clustering. In ACM SIGMoD Record 1999; Vol. 28, No. 2, pp. 61-72, ACM.
16. Aggarwal CC, Yu PS. Finding generalized projected clusters in high dimensional spaces. ACM 2000; Vol. 29, No. 2, pp. 70-81.
17. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. ACM SIGMOD Conference Proceedings; 1998.
18. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd 1996;Vol. 96, No. 34, pp. 226-231.
19. Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. In ACM SIGMOD Record 1998; Vol. 27, No. 2, pp. 73-84, ACM.
20. Zhang T, Ramakrishnan R, Livny M. BIRCH: AnEcient DataClusteringMethod forVery Large Databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Canada; 1996.
21. Aggarwal CC, Yu PS. Outlier detection for high dimensional data. Proc. ACM SIGMOD; 2001.
22. Hubert M, Rousseeuw PJ, Branden KV. ROBPCA: a new approach to robust principal component analysis. Technometrics 2005; 47, 64-79.
23. Egan WJ, Morgan SL. Outlier detection in multivariate analytical chemical data. Analyical Chemistry 1998; 70, 2372-3279.
24. Yang K, Trewn J. Multivariate Statistical Methods in Quality Management. Mc Graw Hill Professional; 2004.
25. Montgomery DC. Introduction to Statistical Quality Control. John Wiley & Sons; 2005.
26. Ben-Hur A, Guyon I. Detecting Stable Clusters Using Principal Component Analysis. In Functional Genomics: Methods and Protocols. Brownstein, M. J. & Kohodursky, A. (eds.) Humana press 2003;159-182.
27. Mills RT, Kumar J, Hoffman FM, Hargrove WW, Spruce JP, Norman SP. Identification and visualization of dominant patterns and anomalies in remotely sensed vegetation phenology using a parallel tool for principal components analysis. Procedia Computer Science 2013; 18, 2396-2405.
28. Stefatos G, Ben HA. Cluster PCA for Outliers Detection in High-Dimensional Data. IEEE 2007; 3961-3966.
29. Saha BN, Ray N, Zhang H. Snake Validation: A PCA-Based Outlier Detection Method. IEEE Signal Processing Letters 2009; (16), 549-552.
30. Esbensen KH. Multivariate Data Analysis - In Practice. 5th Edition, CAMO Process AS, Esbjerg, Denmark; 2005.
31. Martens H, Naes T. Multivariate Calibration. Wiley, Chichester, England; 1989.
32. *http://umetrics.com/sites/default/files/books/sample_chapters/multimega_parti-3_0.pdf* last accessed October 07, 2015.