https://doi.org/10.2352/ISSN.2470-1173.2018.09.IRIACV-276

Outlier Detection in Large-Scale Traffic Data by Regression Analysis

Philip Lam[1], Lili Wang[1], Henry Y.T. Ngan[1], Nelson H.C. Yung[2], Michael K. Ng[1]
[1]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
[2]Department of Electronic & Electrical Engineering, The University of Hong Kong, Hong Kong

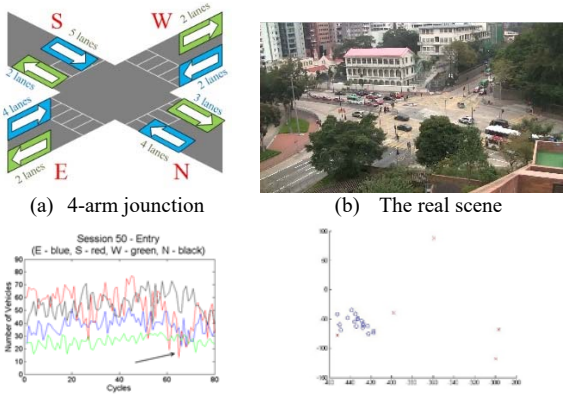
Abstract

A robust outlier detection for large-scale traffic data by an unsupervised regression method is proposed in this paper. Traffic data is collected from loops, sensors and digital cameras all around a city every day. The data size is massive and in a big data format. Outlier is regarded as abnormal traffic situation like traffic jams, low traffic flows, or incidents as well as errors and noise in data storage and transmission. The traffic data to be tackled in this paper is represented by spatial temporal (ST) signals. A principle component analysis (PCA) is used for dimension reduction and to generate a representation of (x, y) −coordinates from the first two component's coefficients in the ST signals. The (x, y) −coordinate points of inliers are measured by Standardized Residual (SR), Hat Matrix (HM) and Cook's Distance (CD) in the regression method so that outliers are assumed to have high changes in these three metrics in the best fit regression model. Experimental result of the proposed method for the Level 1 data achieves detection success rates (DSRs) of 97.37% (SR), 91.19% (HM), 94.28% (CD) for linear regression model, respectively, and 96.80% (SR), 89.71% (HM), 93.14% (CD) for quadratic regression model, respectively. For a finer granularity of Level 2 data, the regression method with the CD metric achieves 94.44% DSR.

1. Introduction

Outlier Detection (OD) [1-3] is a typical topic that related to various fields like military, management, medicine, information technology, etc. This has been becoming more popularly in data science in recent years. Outliers are generally defined [1, 4] as any elements are inconsistent with the majority of data (i.e. inliers). An effective and efficient OD can help to detect any abnormality, to identify the location and to maintain a good data quality for the data science. Many OD methods [1, 5-6] have been developed and adopted in different fields, in which have unique designs in their respective methods in different usages. In transportation, since the traffic data is collected continuously every minute via sensors, digital cameras and detective loops, inspecting any anomaly traffic event or data errors are very important to the transport department or transportation companies to react and deal with the traffic problems such as traffic jams, incidents or abnormal events.

Previously, OD in traffic data is regarded as automatic incident detection (AID) [1, 7-10] and OD in other data (i.e. wireless sensor [2], semi-conductor [3]) is just treated as an abnormal detection. In this paper, we consider the OD as to detect traffic incident as well as data errors upon a real set of traffic data from a 4-arm junction in Hong Kong as shown in Fig. 1 (a),(b). The dataset was collected by a digital camera in 31 days in the junction for two 3-hour sessions per day (i.e. AM: 07:00-10:00 and PM: 17:00-20:00). A total of 764, 027 vehicles were identified by counting in the dataset. The dataset contains traffic jams in different levels and data errors. The traffic



(a) 4-arm jounction (b) The real scene

(c) Session 50: Entries E, W, N are normal, Entry S is abnormal

(d) PM Entry S (With 5 outliers in red crosses)

Fig. 1 (a) A generic diagram of the 4-arm junction; (b) picture of the real scene; (c) abnormal ST signals; (d) Plot of a PCA-projected (x, y) −coordinates.

data has been converted to traffic flow statistics, which is in term of the traffic flow ST signals (Fig. 1(c)). The ST signals, in term of a 3-hour session for each, are further undergone a PCA for dimension reduction to produce *a* (x, y) −coordinates plane for OD (Fig. 1(d)). This paper aims at investigating a statistical approach to tackle the OD by utilizing different metrics such as SR, HM and CD in both linear and quadratic regression models.

Since not many traditional statistical methods have been applied for the OD before, the research significance of this paper is to conduct a preliminary study on the regression model by using different metrics namely SR, HM and CD, performs on OD. Also, we will carry out a comparative study between the linear and quadratic regression models on the OD for Level 1 data. This paper has 3 contributions. The first contribution is to apply an unsupervised regression analysis for the large-scale traffic data by applying the SR criterion. This criterion exploits the leverage points and influence point to determine the CD for the OD. The second contribution is the regression analysis for the Level 1 data achieves accuracies of the OD at 97.37% (SR), 91.19% (HM), 94.28% (CD) for linear regression model, respectively, and 96.80% (SR), 89.71% (HM), 93.14% (CD) for quadratic regression model, respectively. The third contribution is that the regression method with the CD metric is evaluated on OD on a finer granularity of Level 2 data and obtained a 94.44% DSR. The rest of the paper is organized as follows: Section II is the related work of the OD. The details of the proposed OD method is presented in Section III. Performance evaluation results of OD on the original Level 1 data and finer granularity Level 2 data are provided in Sections IV and V, respectively. Finally, conclusions are drawn in Section VI.

2 Related Work

OD [1] in traffic data has become more popular in recent years. In general, OD has several main approaches such as statistical, clustering, proximity and learning approaches. The OD methods are carried in either unsupervised, semi-supervised or supervised manners. In short, the supervised approach exploits the inliers as training data, the semi-supervised approach employs one or some outliers for training the decision boundaries or regions in the training stage. Then, both of them will be undergone a testing stage. In contrast, the unsupervised approach does not require any training process and all data can be directly input for the testing.

In recent time, we have attempted Dirichlet process mixture model (DPMM) (96.67% DSR) [4], kernel density estimation (KDE) (95.20% DSR) [5], Gaussian Mixture Model (GMM) (94.50%) [11] in the statistical approach, modulo-k clustering tree (97.74% DSR) [12] in the clustering approach, distance-based k-nearest neighbor (kNN) (96.19% DSR) [13], density-based local outlier factor (LOF) (93.5% DSR) [14] and quaternion function (97.83% DSR) [15] methods in the proximity approach, one-class support vector machine (59.61% DSR) [5], Naïve Bayes (NB) classifier (93.78% DSR) [11] in the learning approach. Most of them reached over 90% DSR in the same traffic dataset. Most of them were designed in supervised (GMM and NB) and semi-supervised (e.g. one-class SVM, kNN, LOF, modulo-k clustering tree, quaternion) approaches and only DPMM and KDE were in an unsupervised approach. In the review, we discover that the regression model has a potential to be developed as an unsupervised approach.

Previously, quite a lot of AID methods [7-10] have been developed for the OD, however they were evaluated by simulation in computer software (i.e. METANET and IMETANET [10]) or individual collected datasets (i.e. I-880 loop data [7], Seoul loop data [8], Beijing loop data [9], UK motorway M6 [10]). In short, there is no common benchmarking dataset for evaluation. This paper will utilize the recent collected large-scale traffic data from Hong Kong (Fig. 1) as a benchmark testing. Therefore, the evaluated results can be justified in a more objective level.

3. Regression Analysis

In statistics, the regression analysis is a statistical process for estimating the relationships among variables [16]. A typical linear regression model is denoted as

$$y_i = x_i B - \varepsilon_i \quad (1),$$

where $(x_i, y_i), i = 1,2, \dots, n$ are data points from a dataset and $x_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})$ with k independent variables and $B = (\beta_0, \beta_1, \dots, \beta_k)^T$ denotes the $k+1$ parameter set of the regression model. A predicted value $\hat{y}_i$ is defined as

$$\hat{y}_i = b x_i \quad (2),$$

where x_i denotes an independent variable vector, y_i is a dependent variable vector, $b = \hat{B}$, and $\varepsilon_i = |\hat{y}_i - y_i|$ is the predicted error following $\varepsilon_i \sim N(0, \sigma^2)$.

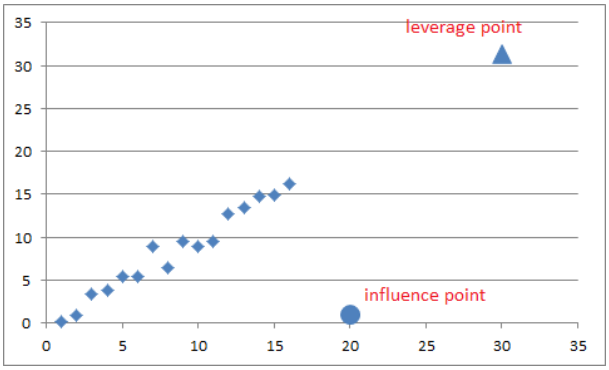


Fig. 2 Example of Leverage point and Influence point.

When researchers are interested to detect an outlier by using a regression method, outliers in the regression model are assumed to be those data points which contain unusual response output values [17]. Usually, data will be classified as outlier if the data's **SR** is too large. The SR formula takes the form of

$$z_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad (3),$$

where $e_i = |y_i - \hat{y}_i|$ is the residual of the data, s denotes the mean square error in the model, and h_{ii} is the i^{th} diagonal component of **HM,** H, which is defined as

$$H = X(X^T X)^{-1} X^T \quad (4),$$

where $X = (x_1, x_2, \dots, x_n)^T$. By empirical results [17], a data point's SR value greater than 3 is sufficient to conclude that that data point is an outlier.

In the regression model, there exist leverage points (shown in Fig. 2) which are those observations made at extreme or outlying values of the independent variables such that the lack of neighboring observations, or a large effect on the predicting regression model. It means that the fitted regression model will be close to that particular observation [17]. For an analysis of such leverage point, the rule of thumb is to determine if h_{ii} is more than a double of the mean leverage. Therefore, it is worth to investigate whether missing some data value will lead to a serious effect on the regression model curve. One criterion is to measure the data's **CD**, D_i, which is denoted as

$$D_i = \frac{e_i^2}{(k+1)s}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \quad (5),$$

where e_i is the residual of the data, s is the mean square error in the model, and h_{ii} is the i^{th} diagonal component of hat matrix H for $k+1$ parameters. A quadratic regression model is just to modify x_i to be a quadratic form such as $y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 - \varepsilon_i$.

Methodology

A flowchart of the proposed regression method is presented in Fig. 3. In a recent study in [5], the correlation between the (x, y) −coordinates is very low on the same traffic signal data. The regression method does not highly rely on the relationship between the (x, y) −coordinates, but depends on the assumption that data

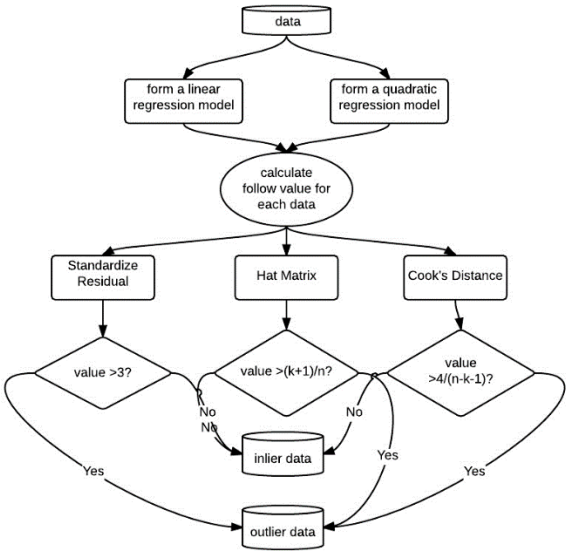


Fig. 3. Flowchart of the proposed Regression method.

residuals on a maximum likelihood estimated (MLE) regression model, for which shows a significant difference between inlier and outlier data points (i.e. $(x, y)-$coordinates) [5]. Therefore, an unsupervised regression method is presented. First, a simple linear or quadratic MLE regression model is constructed based on the $(x, y)-$coordinates. Second, the mentioned outlier criteria in regression such as the leverage and influential points are exploited to detect any outlier.

Although checking SR sometimes is sufficient enough to define an outlier, it is required to check whether the detected influence point and leverage point will be more accurate. In traffic data, some abnormal data are little far away from inlier data. However, in the regression approach, these data's SR could be high but its hat or influential value could be low. Therefore, the hat matrix and CD are employed as another criteria. Procedure of the proposed regression OD is as below:

Step 1. Input the data points

Step 2. Construct a regression model (Eq. (1)) as either
(a) Simple regression model, or
(b) Second order regression model

Step 3. Calculate the criterion value for each metric
(i) z_i Standardized residual (Eq. (3))
(ii) h_{ii} Hat matrix (Eq. (4))
(iii) D_i Cook's distance (Eq. (5))

Step 4. Determine any outlier by empirical thresholds as the rule of thumb (i.e. $z_i > 3$, $h_{ii} > \frac{k+1}{n}$, $D_i > \frac{4}{n-p}$).

4. OD on Original Level 1 Data

Experimental results are given in Table 1 for the original Level 1 data. Typical measurement metrics such as true positive (TP), false positive (FP), true negative (TN), false negative (FN), DSR, True positive rate (TPR), false positive rate (FPR), positive predictive value (PPV) and negative predictive value (NPV) are employed in this paper. The detailed definitions can be referred to [4]. The OD results of linear regression with the CD are illustrated in Fig. 4. The AM Entry West and PM Entry South signals demonstrate FP and FN cases, respectively, which are considered as difficult cases.

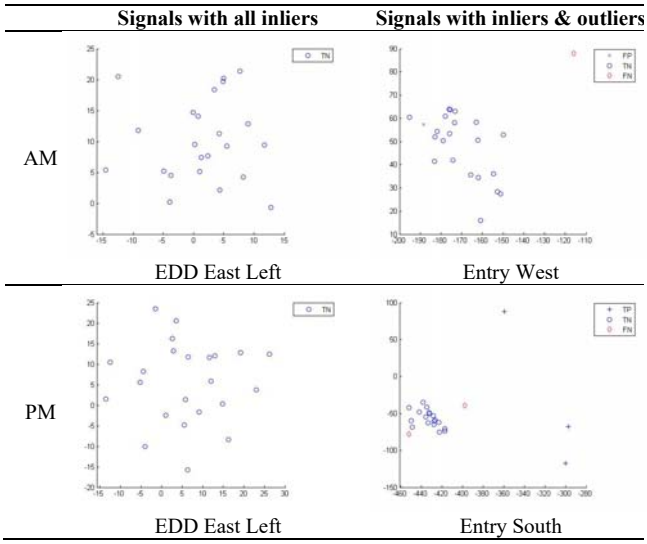


Fig. 4. OD results of linear regression with the CD metric (Regression line is not shown here).

Table 1. Detection success rates (DSR) among various criteria (%).

		Linear regression			Quadratic Regression		
	Criteria	SR	HM	CD	SR	HM	CD
AM	DSR	98.40	91.53	94.05	97.71	90.39	92.91
	PPV	40.00	5.56	12.50	10.00	5.26	5.56
	NPV	99.08	99.00	99.05	98.86	98.98	99.03
	TPR	33.33	33.33	33.33	16.67	33.33	33.33
	FPR	0.69	7.66	5.07	1.15	8.81	6.25
PM	DSR	96.34	90.85	94.51	95.88	89.02	93.36
	PPV	45.45	17.54	39.58	45.00	17.54	34.31
	NPV	97.80	96.59	98.47	97.34	96.56	98.44
	TPR	47.67	19.83	70.17	36.17	19.83	63.50
	FPR	1.63	6.05	4.19	1.65	7.92	5.37
All	DSR	97.37	91.19	94.28	96.80	89.71	93.14
	PPV	42.73	11.55	26.04	27.50	11.40	19.94
	NPV	98.44	97.80	98.76	98.10	97.77	98.74
	TPR	40.50	26.58	51.75	26.42	26.58	48.42
	FPR	1.16	6.86	4.63	1.40%	8.37	5.81

*where red values are best performance, P/NPV (positive/negative predictive value), T/FPR (true/false predictive value)

In the linear regression, the SR performance outperforms to other criteria in the AM sessions, for which the SR could better describe the inlier data points in the center. However, in the PM sessions, although the SR accuracy is still the highest, the sensitivity of CD is 70.17% which is over 20% higher (much better) than the SR's one. Also, the SR and CD criteria are truthful tests for inliers because their NPVs are both over 98% (overall result). Therefore, the SR and CD performance are promising in different situations.

The HM performance is not satisfied where comparing with other two criteria. That shows that the leverage point is not sufficient to say anomaly in the traffic data but influential point is. Furthermore, the overall performance in the quadratic regression model is worse than that in linear regression model. This situation exists due to there is no quadratic relation between (x, y) −coordinates. Therefore, the result would be worst if a wrong model is applied to detect any outlier.

5. OD on Finer Granularity: Level 2 Data

For a deeper study, it is interesting that how would the data be when more PCA processed (x, y) − points in more data. When doing PCA-processed data form the original data to lower data dimension, the OD problem will be simpler. However, for the data, there will be information loss during the processes.

Table 2. Different granularity levels and corresponding ST signals.

Level	Original data Cycles (period)	No. of (x,y)-data points every signal	No. of cycles for abnormal event if outlier(s) exist
Level 1 (original)	80 (3 hours)	23	>4
Level 2	26 (1 hours)	69	≥3
Level 3	13 (0.5 hours)	138	≥1

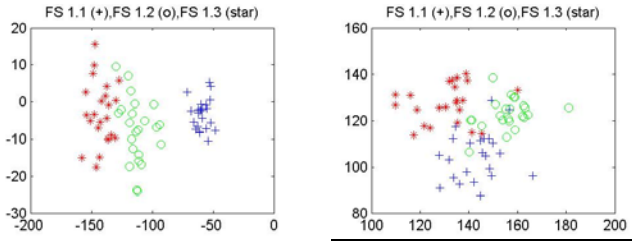
In our original traffic data case, the PCA process compressed the whole data model from around 760,000 vehicles into 874 sessions' data (23 sessions in each signal*2 periods*19 directions). As a PCA penalty, the PCA (x,y)-points show a rough traffic flow information only for a 3-hour session time. In OD, it is hard to detect an abnormal event that is not significantly serious or significantly long duration. In other words, the exact location of OD is hard to trace in a 3-hour session. Furthermore, it is less evidence to suffice the whole inlier data pattern. Therefore, having a finer granularity level of ST signals is a must to study the OD problem for more details of outliers.

One 3-hour Session to Three Finer 1-hour Sessions

This Section has been planned to be conducted the finer granularity level in the list as follows: In the original Level I, PCA-processed data in each signal were processed by a whole 3-hour session of 80 cycles that generates 23 data points (from 23 days). For the ground-truth outlier labeling, a whole session in a signal will be classified as an outlier if any abnormal event is observed across more than 4 cycles. It is considered to affect that specific signal.

In Level II of signal granularity, PCA-processed data in each signal were processed by a 1-hour session (i.e. three 1-hour sessions, each session per 26 cycles) that generates 69 data points. For the ground-truth outlier labeling, a session data in a signal will be classified as an outlier if observed that a session data have any abnormal event across more than or equal to 3 cycle affecting the specific signal. The rest of this Section is to focus on studying the Level II PCA-processed data for OD.

In the level II PCA-processed data, 19 traffic signals * 3 1-hour sessions' PCA-processed data are generated, each traffic direction



(a) (b)

Fig. 5. (a) AM Entry signal (b) PM Entry signal, where separate into three groups by the data period belong to their hour session.

contains total 69 data points. Fig. 5 shows two signal plots. For example, if we call Session 1 in level I, then call Session 1.1, 1.2, 1.3 in Level II. Fig. 27 shows the data plot in some signals where the data are clustered. It is clearly show that there are cluster groups in the AM session for a Entry Signal (Fig. 5(a)) However, there is no the same effect in the PM sessions (Fig. 5(b)). By observation, it is clearly observe the different with level one PCA-processed data is that there is another cluster group in some signal data plot. Different cluster groups may violate the original inliers and outlier assumption or definition that inlier are center a one clustered group. Moreover, if there are two or more outlier groups, there might be the problem that some outlier data may far away some inlier group but mixture with other inlier group.

The principle solution is changing or making more assumptions on inlier and outlier, or assumption in detecting outlier. Another way around is to project the different data group into other domains. The reason of those differences is the pattern of the traffic flow. There are more clusters of data points in each hour group in the AM traffic signal due to the traffic pattern in the AM period is a steady increase. However, less significant clusters in PM period are found because the traffic pattern is become a mess. Therefore, OD seems to be promising in different hour groups in the AM and PM sessions.

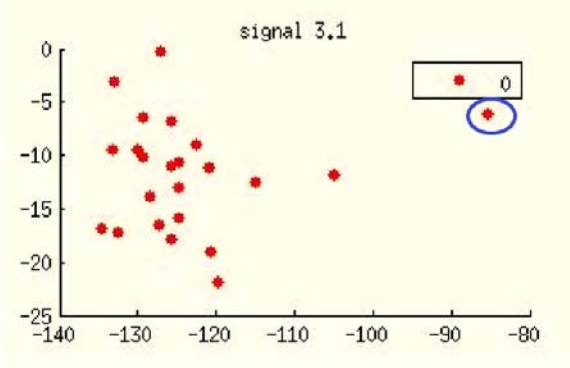
Level II Outlier Labeling

A Level II outlier labeling is newly constructed in this research. In labelling the ground-truth outliers, it is detemined in four kinds of resouces:

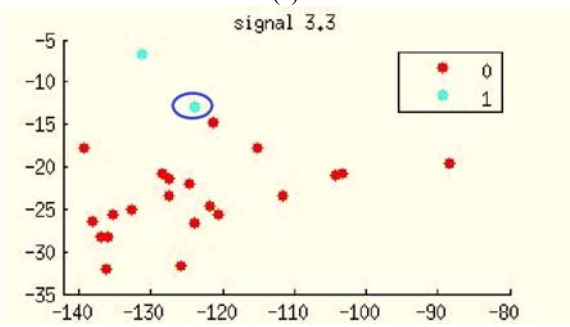
1. Video of the traffic data;
2. The gound-truth remarks on different motion parterns;
3. Raw data plot on different direction ST signals; and
4. Level II PCA processed data scattered plots.

For gound-truth outliers in Level II PCA data, firstly there must cause abnormal events that across over three or more cycles, and the data are reasonly causing a high numerial effect.

By observation, resouces 2, 3 are 4 are hints to the obsever whether there is any datum reasonly causing a high numerial effect. Meanwhile, resouce 1 is for the final confirmation. Five types of anomalies are labeled as the same as the previous work in Level I: Type 1: Hardware failure; Type 2: Frequent congestions in an



(a)



(b)

Fig. 6. PM Entry W signal: (a) the first hour, the blue circled data is far from the data that is not an outlier; (b) the third hour, the blue circled data is near the data that is an outlier. Inlier is labeled as 0 and outlier is labeled as 1.

Entry/Exit; Type 3: Vehicles blocking an Entry/Exit; Type 4: Low volume in an Entry/Exit; Type 5: Congestions in an Exit/Entry leading to low volume in other Entry/Exit.

In Level II PCA domain, although number of data points is increased to 69, there are some problems in labeling ground-truth outlier data. Fig. 6 shows one of the examples. The situation often occurs in the PM session. For ground-truth labelling, a data point observable far from the majority of data is still an inlier (Fig. 6(a)). One of our concerns is that more detailed event can affect the data points seriously and these events are weighted much heavier in the Level II PCA data. Another case is that information of data points fulfill the requirement to treat as Level II ground-truth outlier, but in close to inlier data, for which is higher difficult to be detected as outlier (Fig. 6(b)).

Evaluation on Level II domain

In this Section, we evaluate the OD by the unsupervised linear regression analysis by using the CD metric. We apply the CD metric because it offered a good performance in DSR and TPR for Level 1 data above. In the unsupervised approach, there is no need to select any training data set and all the data are tested as a whole. Table 3 tabulates a summary of the OD results of Level 1 and Level 2. In Level 2 domain, the overall DSR and NPV of the regression analysis by the CD metric are very good with 94.44% and 99.03%, respectively. In short, the performance of the regression analysis in both data granularities are outstanding.

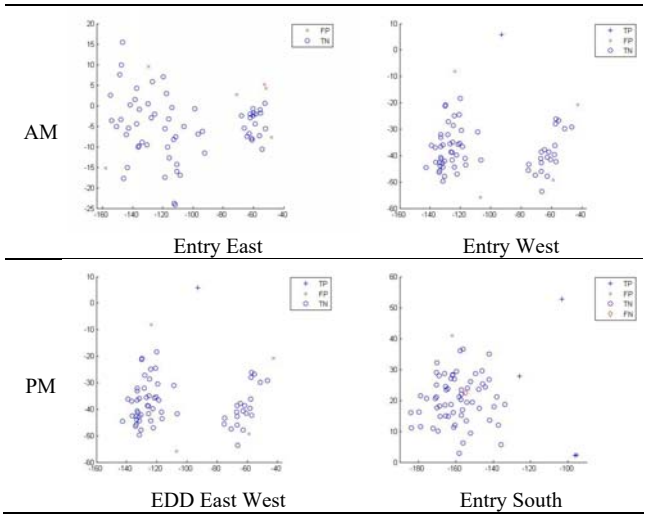


Fig. 7. Level II OD results of linear regression with the CD metric (Regression line is not shown here).

Table 3. Performance of cook's distance criterion in linear regression.

Level I Domain

	DSR	PPV	NPV	TPR	FPR
AM	94.05%	12.50%	99.05%	33.33%	5.07%
PM	94.51%	39.58%	98.47%	70.17%	4.19%
Overall	94.28%	26.04%	98.76%	51.75%	4.63%

Level II Domain

	DSR	PPV	NPV	TPR	FPR
AM	93.90%	8.42%	99.50%	43.75%	5.70%
PM	94.97%	31.98%	98.55%	60.90%	3.77%
Overall	94.44%	20.20%	99.03%	52.33%	4.74%

*where red number represent better performance between two domains.

6. Conclusions

In this paper, we have proposed an OD method based on linear and quadratic regression models, for which three metrics (i.e. SR, HM and CD) are implemented for evaluation. Experimental results show that the proposed algorithm for the Level 1 data can reach accuracies of 97.37% (SR), 91.19% (HM), 94.28% (CD) for linear regression model, respectively, and 96.80% (SR), 89.71% (HM), 93.14% (CD) for quadratic regression model, respectively. These results are comparable to our previous research. We have even attempted a finer granularity of Level 2 data and the regression method with the CD metric offers 94.44% DSR. In the future work, we will try to remove two or more observations in one piece of data the dataset in order to have an in-depth study for the data value's change in the regression model.

Acknowledgment

This research is supported by the grants of Hong Kong RGC GRF: 12201814 and HKBU FRG/14-15/054.

References

[1] S.Y. Chen, W. Wang, and H. Zuylen, "A Comparison of Outlier Detection Algorithms for ITS data," *Expert Systems with Applications*, vol. 37, pp.1169–1178, 2010.

[2] C. O' Reilly, A. Gluhak, M.A. Imran, S. Rajasegarar, "Anomaly Detection in Wireless Sensor Networks in a Non-Stationary Environment," *IEEE Commun. Surveys Tuts.*, 16(3), 1413-1432, 2014.

[3] Z. Li, R.J. Baseman, Y. Zhu, F.A. Tipu, N. Slonim, L. Shpigelman, "A Unified Framework for Outlier Detection in Trace Data Analysis," *IEEE Trans. Semiconductor Manufacturing*, 27(1), pp. 95-103, 2014.

[4] H.Y.T. Ngan, N.H.C. Yung and A.G.O. Yeh, "Detection of Outliers in Traffic Data based on Dirichlet Process Mixture Model," *IET Intelligent Transportation Systems*, vol. 9, no. 7, pp. 773-781, 2015.

[5] H.Y.T. Ngan, N.H.C. Yung, and A.G.O. Yeh, "A Comparative Study of Outlier Detection for Large-scale Traffic Data by One-class SVM and Kernel Density Estimation," *IS&T/SPIE Electronic Imaging,* 94050I-94050I-10, 2015.

[6] S. Sscalera, X. Baro, O. Pujol, J. Vitria, and P. Radeva, *Traffic-Sign Recognition System*, Springer, 2011.

[7] J. Wang and X. Li, "A Hybrid for Automatic Incident Detection," *IEEE Trans. ITS*, vol. 14, no. 3, pp. 1176-1185, 2013.

[8] N.-K. Hong, J-.W. Choi, Y.-K. Yang, "A Study on Incident Detection Model Applying APID Model, Fuzzy Logic and Traffic Pattern," *Proc. IEEE ITSC*, MoD4.3, pp. 196-203, 2007.

[9] L. Chen, Y. Cao, R. Ji, "Automatic Incident Detection Algorithm Based on Support Vector Machine," *Proc. IEEE ICNC*, 864-866, 2010.

[10] A. Gning, L. Mihaylova, R.K. Boel, "Interval Macroscopic Models for Traffic Networks," *IEEE Trans. ITS*, 12(2), pp. 523-536, 2011.

[11] P. Lam, L. Wang, H.Y.T. Ngan, N.H.C. Yung, A.G.O. Yeh, "Outlier Detection in Large-scale Traffic Data by Naïve Bayes Method and Gaussian Mixture Model Method," *IS&T Int'l Sym. Electronic Imaging*, no. 6, pp. 73-78, 2017.

[12] C.H.M. Wong, H.Y.T. Ngan, N.H.C.Yung, "Modulo-k Clustering based Outlier Detection for Large-scale Traffic Data," *Proc. ICITA*, 2016.

[13] T.T. Dang, H.Y.T. Ngan, W. Liu, "Distance-based k-nearest Neighbors Outlier Detection Method in Large-scale Traffic Data," *Proc. IEEE DSP*, pp. 507-510, 2015.

[14] M.X. Ma, H.Y.T. Ngan, W. Liu, "Density-based Outlier Detection Method by Local Outlier Factor on Large-scale Traffic Video Data," *IS&T Int'l Sym. Electronic Imaging*, no. 4, pp. 1-4, 2016.

[15] L-L. Wang, H.Y.T. Ngan, W. Liu, N.H.C. Yung, "Anomaly Detection for Quaternion-valued Traffic Signals," *Proc. IEEE DICTA*, 2016.

[16] J.S. Armstrong, "Illusions in Regression Analysis," *International Journal of Forecasting (forthcoming)* 28 (3): 689, 2012.

[17] W. Mendenhall, and T. Sincich, *A Second Course in Statistics: Regression Analysis*, Seventh Edition, Pearson, 2012.

Author Biography

Philip Lam received the B.Sc. (Hons) degree in Statistics and Operations Research (2015) in Hong Kong Baptist University, China. He is studying the M.Sc. in Actuarial and Investment Science at Hong Kong Polytechnic University, China and is supposed to obtain the degree in 2017. Also, he is currently an actuarial intern in China Taiping Life Insurance (Hong Kong) Company Limited.

Li-Li Wang received the M.Sc. degree from Shandong University, Shandong, China (2007) and the Ph.D. degree from the Hong Kong Polytechnic University, China (2013). She is currently a postdoctoral fellow in the Department of Health Technology and Informatics, the Hong Kong Polytechnic University. Her research interests include digital signal processing, image and video coding, pattern recognition, machine learning, computer vision and big data analysis.

Henry Y.T. Ngan received the B.Sc. degree in Mathematics (2001), the M. Phil. degree (2005) and the Ph.D. degree (2008) in Electrical & Electronic Engineering at The University of Hong Kong, China. He is currently an assistant professor of research in Mathematics, Hong Kong Baptist University. He was a conference chair of IS&T Electronic Imaging 2016, 2017. He is an editor of IS&T Journal of Imaging Science & Technology.

Nelson H.C. Yung received his B.Sc. and Ph.D. degrees from Newcastle University, UK. He has co-authored five books and book chapters and has published over 190 journal and conference papers in the areas of digital image processing. He was a guest editor of the SPIE Journal of Electronic Imaging. He is a Chartered Electrical Engineer, Member of the HKIE, IET and Senior member of the IEEE.

Michael K. Ng received his B.Sc. degree (1990) and M.Phil. degree (1992) in Mathematics at The University of Hong Kong and Ph.D. degree (1995) at Chinese University of Hong Kong, China. He is currently a Chair Professor at the Department of Mathematics, Hong Kong Baptist University. His research interests include Bioinformatics, Data Mining, Operations Research and Scientific Computing. He has published and edited 5 books, and published more than 200 journal papers. He is an SIAM fellow.