Ruiz Blázquez, Ramona, Muñoz Organero, Mario, Sánchez Fernández, Luis. (2018). Evaluation of outlier detection algorithms for traffic congestion assessment in smart city traffic data from vehicle sensors. *International Journal of Heavy Vehicle Systems,* 25(3/4), pp. 308-321.

DOI: 10.1504/IJHVS.2018.10016106

Evaluation of outliers detection algorithms for traffic congestion assessment in smart city traffic data from vehicle sensors

Ramona Ruiz Blázquez*,
Mario Muñoz Organero and
Luis Sánchez Fernández

Department of Telematic Engineering,
Carlos III University of Madrid,
Av. Universidad 30, 28911, Leganés, Madrid, Spain
Email: raruizb@it.uc3m.es
Email: munozm@it.uc3m.es
Email: luiss@it.uc3m.es
*Corresponding author

Abstract: On-board sensors in vehicles are able to capture real-time data representations of variables conditioning the traffic flow. Extracting knowledge by combining data from different vehicles, together with machine learning algorithms, will help both to optimize transportation systems and maximize the drivers' and passengers' comfort. This paper provides a summary of the most common multivariate outlier detection methods and applies them to data captured from sensor vehicles with the aim to find and identify different abnormal driving conditions like traffic jams. Outliers detection represents an important task in discovering useful and valuable information, as has been proven in numerous researches. This study is based on the combination of outlier detection mechanisms together with data classification methods. The output of the outlier detection phase will then be fed into several classifiers, which have been implemented to assess if the multivariate outliers correspond with traffic congestion situations or not.

Keywords: multivariate outliers; traffic jams; outliers detection methods; vehicles telemetry; machine learning.

Reference to this paper should be made as follows: Blázquez, R.R., Organero, M.M. and Fernández, L.S. (xxxx) 'Evaluation of outliers detection algorithms for traffic congestion assessment in smart city traffic data from vehicle sensors', *Int. J. Heavy Vehicle Systems*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: AUTHOR PLEASE SUPPLY CAREER HISTORY FOR EACH AUTHOR OF NO MORE THAN 100 WORDS.

1 Introduction

This work is related to the Smart Cities context where the intelligent use of information and communication technologies has a key role in the performance of better transport solutions with a smart, safe and sustainable mobility. This smart mobility is accomplished thanks to the use of different data provided by mobile sensors that capture in-vehicle telemetry data. In the present study, the target is focused on outlier detection in order to detect and identify traffic anomalies. Our research is especially focused on congested traffic conditions caused by heavy traffic or accidents.

Despite the fact that outlier detection techniques aim to remove anomalous observation from the data, like system faults, human or instrument errors, they can also discover useful abnormal data that reflect significant information depending on the nature of the dataset. A lot of applications that use outlier detection can be found. For instance credit card fraud detection, loan application processing, network intrusion detection, activity monitoring, network performance, fault diagnosis, structural defect detection, satellite image analysis, motion segmentation, time-series monitoring, medical condition monitoring, pharmaceutical research or detecting novelty in the text (Hodge and Austin, 2004).

Nowadays, collecting real traffic data from vehicles is widely available in many scenarios. For the context of this research, a new data-set has been generated using an Android app called 'Smart Driver', developed by our research group, which sends data each second during the driving time to a central processing server.

One of the major interests in detecting and predicting road traffic congestion cases is to use such information so that each vehicle can move from one place to another as quickly and efficiently as possible. This paper focuses on the automatic detection of traffic congestion cases by feeding the output of an outlier detection stage to a final classifier to assess when abnormal traffic conditions are most likely to be due to traffic congestion. The output of the proposed algorithm could then be used to minimise the time wasted in traffic jams or some health problems caused by air pollution due to heavy traffic conditions.

In the next section, the main multivariate outlier detection methods are briefly discussed, followed by the approach taken in this study. Then, the experimental results are presented to end with the conclusions and future works.

2 State-of-the-art

Outlier detection in datasets is a task extensively used in numerous and different domains as stated before. Many researchers are working to improve the accuracy, precision and efficiency of algorithms to find outliers in massive amounts of data. Although many methods and techniques have already been successfully implemented, there is no single universal solution for outlier detection. Each method will be more or less suitable in a combination of different parameters depending on factors such as the structure of the dataset, its size and dimension. Moreover, the final accuracy will depend on the type of data and the type and proportion of outliers. Many reviews and surveys comparing different techniques have been undertaken with similar conclusions (Chen et al., 2010; Zhang, 2013; Penny and Jolliffe, 2001; Gogoi et al., 2011; Ben-Gal, 2005).

Outliers can be detected from univariate or multivariate datasets. As opposed to finding outliers in univariate datasets, finding outliers in a multivariate analysis, with several

mutually dependent variables simultaneously related, will include the correlation among them as a factor to consider in the detection process.

Multivariate outlier detection techniques can be classified based on different criteria. In the next subsections, some of the most common techniques are briefly described.

2.1 Statistical methods

These techniques are based on the assumption that the data are normally distributed. One important aspect of a multivariate normal distribution is that it is completely specified by a mean vector and a covariance matrix.

One observation will be considered as an outlier if it is located relatively far from the centre of the data distribution. The distance is calculated in stochastic units. The 'Mahalanobis Distance' is widely used and depends largely on the estimated parameters of the multivariate normal distribution and is computed as follows:

$$MD(x_i) = \sqrt{(x_i - \overline{x}_n)^T S_n^{-1} (x_i - \overline{x}_n)}, \tag{1}$$

where $\overline{x}_n$ stands for the sample mean vector and S_n^{-1} is the inverse of the sample covariance matrix, and n indicates the total number of observations in the sample.

The Mahalanobis distance can be affected by masking and swamping effects. As a way to accomplish a better detection, robust estimates of the multivariate distribution parameters can be computed. Some robust estimators are the 'minimum covariance determinant' (MCD) and the 'minimum volume ellipsoid' (MVE) (Rousseeuw, 1984, 1985; Rousseeuw and Van Driessen, 1999).

The MCD and MVE estimators are the centre and the covariance of a subsample of size h that minimises the determinant and the volume, respectively, of the covariance matrix associated to the subsample. The value of the h parameter must represent the minimum number of observations which must not be outliers and it is normally taken as captured by equation (2).

$$h = [(n + p + 1)/2] \leq n. \tag{2}$$

Taking into account their statistical and computational efficiency, the MCD is preferred over the MVE (Acuña and Rodriguez, 2004).

2.2 Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction technique, a linear transformation which reduces the dimensionality of the dataset and, therefore, it is suitable for high data dimensionality. In this case, no particular probability distribution is assumed for the data. The object of the analysis is to take p variables and find combinations of these to produce indices that are mutually uncorrelated. The lack of correlation is a useful property because it means that the indices are measuring different dimensions in the data. The first component is a linear combination of the original variables with higher variance; the second component has the second highest variance and so on. However, if the original variables are uncorrelated then the analysis does absolutely nothing (Manly, 1994).

PCA could be used to detect multivariate outliers since a very extreme value will take a principal component and will appear as the end of this component. Unfortunately, although PCA can identify isolated outliers, there is no guarantee that it will work when there are outliers groups due to the masking problem (Peña, 2002).

2.3 Projection pursuit

Unlike PCA, in this case, the variables in the data are considered to follow a multivariate normal distribution, which is completely characterised by its mean vector and its covariance matrix. It consists of constructing an indicator, which can be interpreted as projecting data in a certain direction that reveals the arrangement of points in space. A projection criterion is defined and the direction is found where that criterion is a maximum, since any multivariate outlier observation must appear as an outlier in at least one direction of projection, defined by the line that joins the data centre with the outlier observation (Peña, 2002).

This technique is very effective for multivariate datasets of considerable length; however, the results depend on the chosen indexes.

2.4 Distance-based methods

Distance-based methods for outlier detection are based on the calculation of local distances between objects in the data. An observation is defined as a distance-based outlier, described as $DB(p, d)$, if at least a fraction of p of the observations in the dataset is further than d from it (Knorr and Ng, 1998).

This definition is suitable when the dataset does not fit any standard distribution model. It can discover outliers effectively, but this approach is sensitive to the parameters p and d and the efficiency is low in datasets with a high dimension.

There are other proposals using distance-based outlier detection methods such as the k-nearest neighbours (kNN) algorithm, which calculates the k-distances for all objects and orders the objects in descending order of these values, with the first n objects being considered as outliers.

2.5 Density-based methods

Density-based outlier detection methods estimate the density of the neighbourhood of each observation. An observation is considered as an outlier if it lies in a neighbourhood with low density (Gogoi et al., 2011). A new notion of the local outlier is introduced that measures the likelihood of an object to be an outlier by using the density of the local neighbourhood. The degree of the density of neighbours is called local outlier factor, LOF, and is assigned to each object (Chen et al., 2010). The user decides whether an observation will be considered an outlier based on this degree.

In general, these algorithms are more effective than those distance-based. However, they are more complex and computationally more expensive.

2.6 Clustering techniques

Clustering is an unsupervised learning method in which data are grouped according to similar characteristics. In the clustering outlier detection methods, a cluster of small size, including the size of only one observation, is considered likely to contain outlier observations.

There are several clustering algorithms, like the k-means and k-medoid algorithms.

For outlier detection, the distance to the appropriate centroid (or medoid) of the normal cluster is calculated. If the distance between an object and the centroid is larger than a predefined threshold, the object is treated as an outlier. But these methods are not always

optimised for outlier detection, since their main objective is the grouping of similar samples. They are suitable only if the number of outliers is small (Zhang, 2013).

2.7 Neural networks

Outlier detection could be based on the use of neural networks in general and deep learning structures in particular. Although outlier detection is an unsupervised learning task, it is possible to have some examples of outliers to train the deep learning network. Besides, stacked auto-encoders can be used to capture a representation of the data, in order to detect outliers as those observations that are not explained with that representation.

In Hawkins et al. (2002), a replicator neural networks (RNN) is performed as an outlier detector method, where the RNN is trained from a sample dataset to build a model that predicts the given data. The RNN was able to identify outliers without using class labels and with high accuracy in several datasets, so the effectiveness of the RNN for outlier detection was demonstrated.

2.8 One-class support vector machines

The support vector machines (SVM) are a type of binary classifiers that can be used as regression machines, and for novelty or outlier detection. SVM is a type of kernel-based methods which are applicable to both supervised and unsupervised tasks.

One-class classification algorithms try to find the support of a distribution that is capable of automatically classifying data points as outliers in large amounts of data. A one-class SVM uses an implicit transformation function defined by the kernel to project the data into a higher dimensional space. The algorithm then learns the decision boundary, a hyperplane, that separates the majority of the data from the origin, and the outliers would be those data points allowed to lie on the other side of the decision boundary (Amer et al., 2013). Therefore, the One-Class SVM is an unsupervised outlier detection method that does not assume any parametric form of the data distribution, but is able to capture the real data structure, and it works better when the data is strongly not normally distributed. Strictly speaking, the one-class SVM is not an outlier detection method, but a novelty detection method, where its training set should not be contaminated by outliers as it may fit them (**?**).

2.9 Peña and Prieto algorithm

In multivariate outlier detection, there is no a universal solution and the choice of the best method depends on the different parameters like the number of dimensions of the data or the data type. The combination of different types of algorithms allows optimising the results for this task.

One example is the algorithm described by Peña and Prieto (2001), where the techniques of Projections Pursuit and Statistical Robust Measurements are combined. It is an iterative algorithm in which the observations suspected of being outliers are eliminated from the original data. With a dataset of p variables, the data are projected in $2p$ directions, p directions of maximum kurtosis and p directions of minimum kurtosis. Since any multivariate outlier observation must appear as outlier in at least one direction of projection, the idea of using directions maximising and minimising the kurtosis coefficient of the projected observations, ensures to find those outliers because on the one hand in univariate variables the kurtosis coefficient is increased by the presence of some outliers data, and on the other hand a big group of outliers can cause bimodality and low kurtosis (Peña, 2002). Once none

observation is eliminated, the robust estimates, mean vector and covariance matrix, are computed from the remaining data sample. One observation will be considered as an outlier if its Mahalanobis distance is greater than a given threshold.

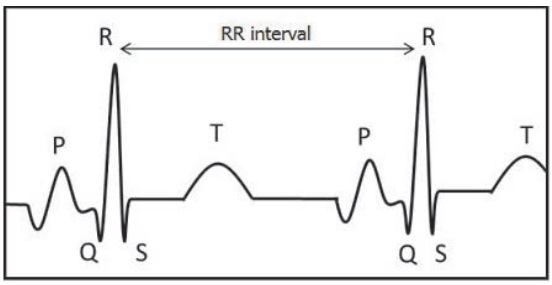
3 Approach

Our intended objective in this work is to detect anomalous driving conditions from the data collected by the drivers themselves. The Android app 'SmartDriver', developed by our research group, gets measurements from two sensors: the GPS in the mobile device itself and a wearable heart rate sensor in a chest band worn by the driver. A set of twenty-five variables, numerical and categorical, are obtained from the sensed raw data. A sample of some of these data is shown in Table 1, where PKE stands for Positive acceleration kinetic energy per distance (Watson et al., 1985), described in equation (3), and the RR interval is the distance between two consecutive R wave peaks which is equivalent to the time between two heartbeats, as shown in Figure 1, and it is measured in milliseconds.

$$\mathrm{PKE} = \frac{\sum(V_f^2 - V_i^2)}{d}, \frac{dv}{dt} > 0. \tag{3}$$

The PKE variable is an indicator that represents the ability to keep the kinetic energy of the vehicle as low as possible; therefore a nervous driving is associated with a high PKE value, while, on the contrary, a smooth driving is associated with a PKE value close to zero (Andrieu and Saint Pierre, 2012). In equation (3), V_f and V_i stand for final velocity and initial velocity respectively, measured in metres per second (m/s), during time intervals in which acceleration is positive, and d is the total distance travelled in metres.

Figure 1 A sample of two ECG waves



The original data consist of 25 variables from different drivers most of them while commuting to work. There are three types of variables:

- periodical observations captured each second
- average values for observations associated with a stretch of road of 500 m
- observations associated particular to events such as acceleration values above a threshold limit.

Table 1 Raw data collected

Variable	*Datum*	*Description*
1	Timestamp	Date and time of the day
2	Event type	High Acceleration, High deceleration, High heart rate, heart rate, high speed, Vehicle speed, Vehicle location, PKE, RR
3	Driver identifier	64-character alphanumeric string
4	Latitude	Latitude for the vehicle
5	Longitude	Longitude for the vehicle
6	Velocity	Kilometres per hour
7	Observation value	High acceleration, high heart rate, High Deceleration, High speed, PKE
8	Average value in a stretch of road	For heart rate & Vehicle speed
9	Median value, in a stretch of road	For vehicle speed
10	Standard deviation in a stretch of road	For heart rate and vehicle speed
11	RR value	For vehicle location
12	Road type	highway, highway_link, trunk, trunk_link, primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, Residential, road, unclassified, service, living_street, pedestrian, track, path, circleway, footway, steps

Since traffic jams are anomalous traffic situations, where traffic flow changes abruptly, these changes must be reflected by outliers, so one of the performed tests to validate this idea is to be able to detect multivariate outliers from the collected data.

Raw data are processed in a first stage using just the periodical observations (vehicle location, velocity and RR), and data corresponding to a single driver in a stretch of highway in both directions, whose length is about five kilometres, during nine days. To have a first statistical characterisation of the data and outlier detection, the sample to test consisted of 2183 observations, taken each second, with six variables. These variables are the average velocity, PKE, instant velocity, instant acceleration, RR and pNN50, where the average velocity, PKE and pNN50 are calculated at each previous 30 s interval.

The pNN50 heart rate variability statistic, in equation (5), is the percentage of the NN50 count, defined as the mean number of times in which the change in successive values (NN, normal to normal or RR) exceeds 50 ms (Miteus et al., 2002). In equation (4) the NN50 formula is presented, where NN_i stands for the actual NN interval, NN_{i-1} is the previous NN interval and k is the total number of intervals.

$$NN50 = \sum_{i=1}^{k} [(NN_i - NN_{i-1}) > 50 \text{ ms}] \tag{4}$$

$$pNN50 = \frac{NN50}{k} \times 100. \tag{5}$$

The statistics for the selected variables, as well as their univariate outliers are shown in Table 2. Several techniques have been used in order to detect the multivariate outliers getting similar results.

Table 2 Variables statistics

Variable	*Mean*	*Std. dev.*	*Median*	*Interq. range*	*Outliers*
Average velocity	22.982	6.0325	23.794	3.7490	237
PKE	0.2756	0.1514	0.2556	0.1815	58
Velocity	22.96	6.2186	23.88	3.9036	244
Acceleration	–0.00328	0.5537	0.0099	0.358	127
RR	0.8696	0.097	0.861	0.126	37
pNN50	20.37	13.326	16.67	20	10

After the previous statistical characterisation, a second experiment is carried out to prove how these multivariate outliers are related to traffic congestion situations. Data from two drivers are collected in the same stretch of highway for 32 days, where each day containing a traffic congestion event has been labelled. Two kinds of classifiers are been used in several tests with different variables, obtaining very promising results.

4 Experimental results

The theoretical results obtained from various tests validate the ideas previously presented. In these tests, the data are captured in a stretch of the M40 highway in Madrid, between the kilometres 21 and 27.

For the first test, the data are collected for a single driver, during nine days, in both directions. With the purpose of detecting multivariate outliers, data are transformed into useful information where each specific variable is processed and must be considered in relation to the other variables. These variables, shown in Table 2, have been taken each second and at 30 s intervals, getting a sample of 2183 observations. One variable is the average velocity in the interval used, and the instantaneous velocity at the final point of the interval, both measured in m/s. Another variable is the instantaneous acceleration, measured in ms^2, and the other variables are the PKE, RR and pNN50, explained in the previous section.

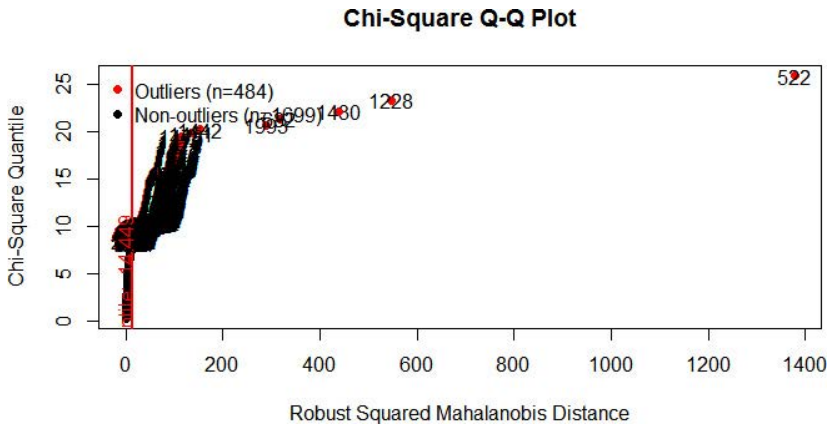
Several methods have been tested to detect outliers. First of all, the Peña & Prieto algorithm and the MCD robust method have been implemented. With these techniques, under the multivariate normality hypothesis, the square of the Mahalanobis distance is distributed as a χ^2 distribution with p (number of variables) degrees of freedom. The threshold used in both methods to determinate if an observation is an outlier has been the 0.975 quantile, that corresponds to a value of 14.449. Although, Hair et al. (1999) suggest a more conservative level of threshold to find the cutoff point for deciding if an observation is an outlier, like for example the 0.999 quantile.

Besides, the LOF algorithm and a k-means clustering algorithm have been performed getting very similar results, as shown in Table 3, where the first five outlier observations are clearly away from the rest of the data, shown in Figure 2, and there are about ten outliers with a Mahalanobis distance greater than 110. But it must be taken into account that the researcher driving the car when getting the data is who really decides whether an observation will be considered as an outlier or not.

The total number of multivariate outliers found by the Peña & Prieto algorithm and a MCD robust method are 461 and 484, respectively, over 2183 observations. Even if these methods assume a normal distribution of the multivariate variables, which is not the case

for our data, the results obtained with an outlier mining technique like a LOF algorithm or a k-means clustering algorithm, using seven neighbours, are very similar, at least with the further observations. Also, a one-class SVM has been performed, where a total of 438 outliers were detected.

Figure 2 Robust outliers detection method (see online version for colours)



It is interesting to note that some multivariate outliers were not found in the univariate analysis, as it is shown in Table 2. That means that each multivariate outlier is not unique in every single variable, but it is unique in the combination of variables (Hair et al., 1999). Like the Mahalanobis distance takes into account the correlation between variables, different results are expected if the variables are permuted, even if the variables have a very low correlation, with exceptions here of the velocity and average velocity, and RR and pNN50 variables. If these correlated variables are eliminated, the results maintain the same five outliers further away, although the number of outlier observations is smaller. In Table 3, the ten furthermost observations for each algorithm are shown.

Table 3 Top 10 outliers for different algorithms

Peña & Prieto		*MCD*		*LOF*	*Clustering* $k-Means$ $k=7$
Observation index	*Mahalanobis distance*	*Observ. index*	*Mahalanobis distance*	*Observ. index*	*Observ. index*
522	1951.1196	522	1378.032	1993	522
1228	785.54793	1228	546.831	522	1228
1430	617.97907	1430	439.118	1451	1430
682	455.68590	682	317.144	1228	682
1993	386.45287	1993	288.455	523	1993
1442	194.14296	1442	152.540	1430	1442
1451	138.60096	1451	133.711	682	1789
1789	121.03030	1478	115.621	1503	354
1613	112.67372	1479	113.164	1442	708
1503	110.83728	1470	111.383	1613	626

Not surprisingly, outliers may be caused by other situations different from traffic jams. One way to verify when a traffic jam occurs is to implement several classifiers to identify when a traffic congestion has actually happened. In this second test, the data used corresponded to the same stretch of highway, collected by two different drivers in a total of 32 days of which five days have been labelled as experiencing a traffic jam.

To calculate the number of outliers per day, only four of the six variables employed before have been used by both drivers. These variables are the average velocity, PKE, instantaneous velocity and acceleration. And the method implemented has been the Peña & Prieto algorithm with a threshold of cut off for the Mahalanobis distance of 0.999.

The five variables related to each of the 32 days used to train the classifiers are the total number of outliers, the maximum Mahalanobis distance, average Mahalanobis distance, the minimum velocity and average velocity. Next, the correlation matrix of the variables shows that there is a high correlation between the total number of outliers in a day with traffic congestion and the minimum and mean velocity. Moreover, there is a high correlation between the minimum and mean velocity with traffic congestion, as expected.

$$
\begin{array}{c}
 \\
Outliers \\
Max.M. \\
MeanM. \\
Min.Vel. \\
Avg.Vel. \\
Jam
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
Outliers & Max.Mah. & MeanMah. & Min.Vel. & Avg.Vel. & Jam
\end{array} \\
\left(\begin{array}{cccccc}
1.00000 & 0.73235 & 0.03528 & -0.79409 & -0.79884 & 0.79811 \\
0.73235 & 1.00000 & 0.47968 & -0.38347 & -0.37648 & 0.38193 \\
0.03528 & 0.47968 & 1.00000 & 0.09223 & 0.08873 & -0.03650 \\
-0.79409 & -0.38347 & 0.09223 & 1.00000 & 0.92356 & -0.90350 \\
-0.79884 & -0.37648 & 0.08873 & 0.92356 & 1.00000 & -0.91128 \\
0.79811 & 0.381937 & -0.03650 & -0.90350 & -0.91128 & 1.00000
\end{array}\right)
\end{array}
$$

A logistic regression linear model (logit) and an SVM with linear kernel have been the classifiers used to determinate if there is a traffic jam on a particular day. Both classifiers have obtained similar results. The confusion matrices are shown in Tables 4 and 5, with hit rates of 93.75% and 100%, respectively.

Table 4 Logit confusion matrix for test 2

Actual vs. predicted	*No traffic jam*	*Traffic jam*	*Actual*	*Recall*
No traffic jam	26	1	27	96.3%
Traffic jam	1	4	5	80%
Predicted	27	5	32	88.15% Avg.recall
Precision	96.3%	80%	88.15% Avg.precision	93.75% Avg.accuracy

A validation experiment only using variables dependent on outliers has also been conducted. In this case, the chosen variables have been the total number of outliers per day, computed again with the Peña & Prieto algorithm, the maximum outliers density in a stretch of about 200 m, and the number of outlier bursts superior to 10 outliers in 200 m. The correlation matrix shows that the presence of a traffic jam is highly correlated with the maximum density of outliers and the outliers bursts in a stretch of highway. The results obtained from both classifiers are identical and shown in Table 6, with hit rates of 96.88%.

$$\begin{array}{c} \\ Outliers \\ Max.Density \\ Bursts \\ Jam \end{array} \begin{array}{c} \begin{array}{cccc} Outliers & Max.Density & Bursts & Jam \end{array} \\ \begin{pmatrix} 1.00000 & 0.97119 & 0.96129 & 0.79811 \\ 0.97119 & 1.00000 & 0.95184 & 0.82374 \\ 0.96129 & 0.95184 & 1.00000 & 0.87130 \\ 0.79811 & 0.82374 & 0.87130 & 1.00000 \end{pmatrix} \end{array}$$

The same test has been performed calculating the total number of outliers per day by means of a one-class SVM. In this case, the correlation matrix shows a lower correlation between the maximum density of outliers and the outliers bursts with the traffic jams.

$$\begin{array}{c} \\ Outliers \\ Max.Density \\ Bursts \\ Jam \end{array} \begin{array}{c} \begin{array}{cccc} Outliers & Max.Density & Bursts & Jam \end{array} \\ \begin{pmatrix} 1.00000 & 0.95791 & 0.95983 & 0.76121 \\ 0.95791 & 1.00000 & 0.97834 & 0.71718 \\ 0.95983 & 0.97834 & 1.00000 & 0.67147 \\ 0.76121 & 0.71718 & 0.67147 & 1.00000 \end{pmatrix} \end{array}$$

The results of this third test, shown in Tables 7 and 8, remain the same for the one-class SVM classifier with an accuracy rate of 96.88%, while for the logit classifier it is a bit lower with an accuracy rate of 93.75%.

Table 5 SVM confusion matrix for test 2

Actual vs. predicted	*No traffic jam*	*Traffic jam*	*Actual*	*Recall*
No traffic jam	27	0	27	100%
Traffic jam	0	5	5	100%
Predicted	27	5	32	100%
				Avg.recall
Precision	100%	100%	100%	100%
			Avg.precision	Avg.accuracy

Table 6 Logit and SVM confusion matrix

Actual vs. predicted	*No traffic jam*	*Traffic jam*	*Actual*	*Recall*
No traffic jam	27	1	28	96.43%
Traffic jam	0	4	4	100%
Predicted	27	5	32	98.2%
				Avg.recall
Precision	100%	80%	90%	96.88%
			Avg.precision	Avg.accuracy

Even though the data collected correspond only a few days, these experiments with real world traffic data show that this approach may be efficient and effective.

Table 7 Logit confusion matrix for test 3

Actual vs. predicted	*No traffic jam*	*Traffic jam*	*Actual*	*Recall*
No traffic jam	26	1	27	96.3%
Traffic jam	1	4	5	80%
Predicted	27	5	32	88.15%
				Avg.recall
Precision	96.3%	80%	88.15%	93.75%
			Avg.precision	Avg.accuracy

Table 8 SVM confusion matrix for test 3

Actual vs. predicted	*No traffic jam*	*Traffic jam*	*Actual*	*Recall*
No traffic jam	27	1	28	96.43%
Traffic jam	0	4	4	100%
Predicted	27	5	32	98.2%
				Avg.recall
Precision	100%	80%	90%	96.88%
			Avg.precision	Avg.accuracy

5 Conclusions and future work

The challenge of searching for multivariate outliers in a dataset by means of different techniques has been considered in this paper, combined with the use of logit and SVM classifiers, with supervised learning, in order to identify traffic jams.

The results from the outlier detection methods, that are summarised in Table 3, show that even if the data are not normally distributed all the techniques present similar outcomes. Although, in general, clustering methods work worse than density-based methods, in this scenario, where data are not normally distributed, a clustering algorithm seems to have a better response.

Regarding the performance of the classifiers, the results obtained were fairly similar, although the SVM showed a slightly better behaviour.

In summary, it is possible to detect traffic congestion situations by means of outlier detection from sensor data collected by drivers followed by a sample classification algorithm.

Future works should be focused on finding a set of variables with best discriminating power, to improve the accuracy of the outliers detection methods and applying the algorithms to different types of road in a smart city. Moreover, it will be necessary to collect more data from different drivers and to label the anomalous traffic situations in order to develop the detection of traffic incident through recognising outliers.

Another approach that will be considered in future work will be the development of algorithms to predict upcoming situations based on current sensor data. Predicting upcoming heavy traffic conditions will help in changing the route before encountering the traffic incident.

Acknowledgement

The research leading to these results has received funding from the '1HERMES-SMART DRIVER' project TIN2013-46801-C4-2-R (MINECO), funded by the Spanish Agencia Estatal de Investigación (AEI), and the 'ANALYTICS USING SENSOR DATA FOR FLATCITY' project TIN2016-77158-C4-1-R (MINECO/ERDF, EU) funded by the Spanish Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF). And the first author was supported by the MINECO Grant BES-2014-070462.

References

Acuña, E. and Rodriguez, C. (2004) *A Meta Analysis Study of Outlier Detection Methods in Classification*, http://academic.uprm.edu/eacuna/paperout

Amer, M., Goldstein, M. and Abdennadher, S. (2013) 'Enhancing one-class support vector machines for unsupervised anomaly detection', *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, Chicago, IL, USA, pp.8–15.

Andrieu, C. and Saint Pierre, G. (2012) 'Using statistical models to characterize eco-driving style with an aggregated indicator', *IEEE Intelligent Vehicles Symposium*, Alcala de Henares, Madrid, Spain, pp.63–68.

Ben-Gal, I. (2005) 'Outlier detection', *Data Mining and Knowledge Discovery Handbook*, pp.131–146.

Chen, S., Wang, W. and Zuylen, H.V. (2010) 'A comparison of outlier detection algorithms for ITS data', *Expert Systems with Application*, Vol. 37, pp.1169–1178.

Gogoi, P., Bhattacharyya, D.K., Borah, B. and Kalita J.K. (2011) 'A survey of outlier detection methods in network anomaly identification', *The Computer Journal*, Vol. 54, No. 4, pp.570–588.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1999) *Análisis Multivariante*, 5th ed., Madrid Prentice-Hall.

Hawkins, S., He, H., Williams, G. and Baxter, R. (2002) 'Outlier detection using replicator neural networks', *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, Aix-en-Provence, France, pp.170–180.

Hodge, V.J. and Austin, J. (2004) 'A survey of outlier detection methodologies', *Artificial Intelligence Review*, Vol. 22, pp.85–126.

Knorr, E.M. and Ng, R.T. (1998) 'Algorithms for mining distance-based outliers in large datasets', *Proceedings of the 24th VLDB Conference*, New York, USA, pp.392–403.

Manly, B.F.J. (1994) *Multivariate Statistical Methods: A Primer*, 2nd ed., Chapman and Hall, London.

Miteus, J.E., Peng, C.K., Henry, I., Goldsmith, R.L. and Goldberger, A.L. (2002) 'The pNNx files: re-examining a widely used heart rate variability measured', *Heart*, Vol. 88, pp.378–380.

Peña, D. and Prieto, F.J. (2001) 'Multivariate outlier detection and robust covariance matrix estimation', *Technometrics*, Vol. 43, No. 3, pp.286–310.

Peña, D. (2002) *Análisis de Datos Multivariantes*, Mc Graw Hill Interamericana de España, SA.

Penny, K.I. and Jolliffe, I.T. (2001) 'A comparison of multivariate outlier detection methods for clinical laboratory safety data', *The Statistician*, Vol. 50, No. 3, pp.295–308.

Rousseeuw, P.J. and Van Driessen, K. (1999) 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics*, Vol. 41, No. 3, pp.212–223.

Rousseeuw, P.J. (1984) 'Least median of squares regression', *Journal of the American Statistical Association*, Vol. 79, pp.871–880.

Rousseeuw, P.J. (1985) 'Multivariate estimation with high breakdown point', in Grossmann, W., Pflug, G., Vincze, I. and Wertz, W. (Eds.): *Mathematical Statistics and Applications*, Vol. B, pp.283–297.

Schölkopf, B., Williamson, R., Smola, A. and Platt, J. (1999) 'Support vector method for novelty detection', *Advances in Neural Information Processing Systems*, Vol. 12, pp.582–588.

Watson, H.C., Milkins, E.E., Holyoake, P.A., Khatib, E.T. and Kumar, S. (1985) 'Modelling emissions from cars', *Proceedings of the 10th Australian Transport Research Forum*, Melbourne, Vols. 1 and 2, pp.87–109.

Zhang, J.(2013) 'Advancements of outlier detection: a survey', *ICST Transactions on Scalable Information Systems*, Vol. 13, No. 1, pp.1–26.

Website

The R Project for Statistical Computing, https://www.r-project.org/

Query

AQ1: PLEASE SUPPLY CAREER HISTORY FOR EACH AUTHOR OF NO MORE THAN 100 WORDS.