BibTeX:

Analyzing the Correlation Among Traffic Loop Sensors to Detect Anomalies in Traffic Loop Data Streams

Gustavo Souto, Thomas Liebig

Dortmund University, Germany
gustavo.souto@tu-dortmund.de, thomas.liebig@tu-dortmund.de

Abstract. This work aims to analyze whether traffic loop data sensors hold any correlation among them which could support the process to detect anomalies in traffic data stream. In order to find out such a correlation among them we apply a Statistical Baseline Method along with a Sensor Correlation Analysis (SCA) approach. The statistical model analyzes in an unsupervised manner the data distribution in order to detect the events that are three times standard deviation or greater than a threshold ($3 \times \sigma^2 + \mu$) and then passes them to the SCA which in turn analyzes whether an event in a sensor S_k also affected its nearest sensor in time period ΔT after the statistical model detects it. We evaluate our approach by comparing the detected anomalies against traffic alerts which are emitted by Traffic Agents on Twitter.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: anomaly detection, data stream, spatio-temporal correlation, traffic loop sensors

1. INTRODUCTION

Anomaly detection is the process of finding patterns which deviate much from the normal behavior of the data. As result, this process might find one of the following types of anomalies: point anomaly, contextual anomaly, and collective anomaly [Chandola et al. 2009]. The literature also refers anomaly as outliers, abnormalities, discordant or deviants [Aggarwal 2013] and an Event can be described as an occurrence of an anomaly in a certain place during a particular interval of time, Equation 1 [Artikis et al. 2014; Souto and Liebig 2015]. Anomaly detection has applications in Stocks Exchange, Health Care, Network Security as well as in other fields of the industry and science.

$$E =< timestamp, location < lat, long >, cause > \tag{1}$$

In literature, a data stream is defined as a continuous, high-speed and unbounded source of data in which the data arrives as an uncontrollable sequence. This paradigm has recently emerged due to the continuous data problem [Bifet et al. 2011], and therewith this process holds important challenges, specially in the field of anomaly detection. Data stream analysis process imposes some constraints such as processing of the data in a limited amount of memory and in a limited quantity of time, be able to process at any point, and receive a data point at a time and inspect it in at most only once. An approach for anomaly detection in data stream depends also on some particular factors about the data domain. For instance, an approach which desires to detect anomalies in *spatio-temporal* data should take into account the autocorrelation between spatial and temporal features. The vehicle traffic data is an example of spatio-temporal data which has gained more attention in recent years

Fig. 1. Locations of SCATS sensors (marked by red dots) within Dublin, Ireland. Best viewed in color.

due to its importance in city traffic planning. By analyzing traffic data is possible to detect some events such traffic jams and accidents. The Figure 1 depicts the SCATS[1] sensors within Dublin, Ireland. See Section 4.1 for more details. Unfortunately, the SCATS data emitter and Dublinked[2] do not provide training dataset or ground truth which could provide us insights about what is normal and/or what is an anomaly in Dublin traffic data. Therefore, building a classification model to detect traffic anomalies is not possible since we do not have such a training dataset directly. It is known supervised methods are more reliable than unsupervised ones, but the task to label data could be very time-consuming depending on the size of data as well as, in the most of the cases, a domain expert must manually label the data. Therefore, our aim is to analyze whether the traffic loop data sensors hold any correlation among them which could indicate low-level anomalous events in traffic loop data stream. This work applies a basic statistical model ($3 \cdot \sigma^2$) which is baseline method along with *Sensor Correlation Analysis* (SCA) approach to detect low-level anomalous events in traffic loop data stream through the spatio-temporal correlation among traffic loop sensors. This statiscal model is applicable to SCATS data, because it is modeled by a Gaussian distribution. The statistical model analyzes the data distribution in an unsupervised manner in order to detect the events that are three times standard deviation or greater than this threshold and then passes them to the *SCA* which in turn analyzes whether an event in a sensor S_k also affects its nearest sensor in the time period $\Delta T1$ after the statistical model detects it. Some important questions arise from this approach and we aim to answer them in this work: "*Does an event at a sensor S_k affect its nearest sensor S_w within a time-period t?*","*How often is the nearest Sensor affected by an event which takes place at another Sensor?*", and "*Does the correlation among traffic loop sensors help the detection of traffic anomalies?*"

This work is structured as follows: Section 2 discusses the related works about anomaly detection in traffic data streams, Section 3 describes our approach to analyze the correlation among traffic loop data sensors, Section 4 presents our experiments, and finally, the conclusion in the Section 5.

2. RELATED WORK

Stolpe et. al. propose [Stolpe et al. 2013] a Vertically Distributed Core Vector Machines (VDCVM) algorithm for anomaly detection which is based on Core Vector Machine (CVM) algorithm [Bădoiu and Clarkson 2002]. The VDCVM has two components, the Central Node P_0 which coordinates the entire system and the Data Node $P_1 ... P_k$ which detects the anomalies in a distributed manner. The Data Node has two more sub-components, the Worker and Data Repository. The anomaly is

[1]Sydney Coordinated Adaptive Traffic System (SCATS)
[2]Dublinked (http://www.dublinked.com/) is a data sharing network which provides different datasets from Dublin, Ireland.

detected locally by each Worker through a local model and sent to the Central Node along with a small sample of all observations. Then, the Central Node trains a global model on such a sample and used to define whether the sent observation is an anomaly or not. The advantage of this work is the good communication cost between Workers and the Central Node in the training phase, but this approach cannot detect anomalies which are global due to a combination of features, and that is its disadvantage.

In [Yang et al. 2014], Yang et. al present a non-parametric Bayesian method, or Bayesian Robust Principal Component Analysis (RPCA) - BRPCA, to detect traffic events on road. This method takes the traffic observations as one dimension data (1-D) and converts it into a matrix format which in turn decomposes it into a superposition of low-rank, sparse, and noise matrices. The idea of BRPCA is to improve the traffic detection by sharing a sparsity structure among multiple data streams affected by the same events. Such an approach uses multiple homogeneous data streams and a static weather data source in the detection process. The advantage of this work is the generation of a ground truth by 3 expertises in the traffic domain which reviewed different plots. However, the approach is limited to detect only 3 types of traffic events which are Slow down, Unexpected high traffic volume and Traffic jam.

Guo et al. [Guo et al. 2014] propose a traffic flow outlier detection approach which focuses on the pattern changing detection problem to detect anomalies in traffic conditional data streams. The traffic data comes from inductive loop sensors of four regions in United State and United Kingdom, as well as this works makes use of a short-term traffic condition forecasting system to evaluate the proposed approach. This approach performs the analysis of the incoming data point after the data point be processed by Integrated Moving Average filter (IMA) which captures the seasonal effect on the level of traffic conditional series, and then Kalman filter picks up the local effect flow levels after IMA, and GARCH filter models and predict time-varying conditional variance of the traffic flow process. These filters constitute together the integrated forecast system aforementioned. Although the results present good performance about the detection of outliers. This work does not apply another procedure to verify the uncertainty of the detection (e.g. check a different source such as traffic alerts on social networks), that is, whether that event is a real anomaly, or not.

Trilles et al. [Trilles et al. 2015] propose a variation of CUmulative SUM (CUSUM) algorithm in Storm Framework[3] to detect anomalies in data streams near to Real-Time. This approach is only applied when the observations are in-control, that is, the data is normally distributed. In the anomaly detection process the CUSUM is obtained by computing $Y_i = Y_{i-1} z_i$, where z_i is the standard normal variable which is computed as follows $z_i = \frac{x_i - \bar{x}}{s}$, where the s is the Standard Deviation of time series. The events are detected by the Equation 2, if Y_{H_i} exceeds the threshold (CUSUM control charts) $\hat{A} \pm h\sigma_x$ ($h = 5$ and σ_x is the Standard Deviation), then it is an *Up-Event* due its increasing and if Y_{L_i} is greater than threshold (CUSUM control charts) $\hat{A} \pm h\sigma_x$ ($h = 5$ and σ_x is the Standard Deviation), then it is an *Down-Event* due its decreasing. The k variable ("Slack") is the reference value which is usually set to be one half of the mean. The advantages of this work are the application of a simple approach for Real-Time anomaly detection and the dashboard application to visualize the detected events. However, the work does not present experiments with a data source wich has high refresh rate such as SCATS data stream.

$$Y_{H_i} = MAX[0, (z_i - k) + Y_{H_i} - 1] \qquad Y_{L_i} = MIN[0, (z_i - k) + Y_{L_i} - 1] \tag{2}$$

Other works also propose solutions to detect anomaly traffic events such as [Yang and Liu 2011], [Liu et al. 2011], [Pang et al. 2013], [Pan et al. 2013], [Yang et al. 2014], [Liu et al. 2014], [Liu et al. 2014]. However, these solutions make use of moving sensors such as GPS, and we have been focusing

[3]Storm Framework: https://storm.apache.org/

on Static sensors (e.g., SCATS sensors) since our work deals with such a kind of sensors as well as the literature present fewer works using Static sensors than Moving sensors.

Although these works present some substantial advances in the field of anomaly detection in data streams, the field is still in its early stage, and therewith it is possible to see that such works hold some drawbacks which were already discussed as well as open tasks such as incorporate expert knowledge in anomaly detection in traffic of vehicles. Incorporation of expert knowledge data is an interesting research direction which should receive more attention in future, because expert knowledge on the relationship between events may improve detection of anomalous event patterns. None of presented related works approached expert knowledge, but [Schnitzler et al. 2014] and [Liebig et al. 2013] are good references. These works use *Street Network* from OpenStreetMap[4] (OSM) that is a kind of expert knowledge in the process to detect traffic anomalies.

3. TRAFFIC LOOP SENSOR ANALYSIS

In order to find out whether the traffic loop sensors hold some spatio-temporal correlation among them which might support in the anomaly detection process. We apply a statical baseline method along with a SCA approach. The statistical model analyzes the SCATS data stream in order to find (vehicle) flow values which are above some threshold. The detected events are sent to SCA process which analyzes the spatio-temporal correlation of anomalous events over a close sensor, at this process we make use of Street Network data from OpenStreetMap which is a kind of expert knowledge to find close sensors. Our approach to find the spatio-temporal correlation among sensors in the anomaly detection process consists of the following components: Feature Selection, Data Segmentation, Data Summarization, Anomaly Detection and Sensor Correlation Analysis (SCA). These components are implemented on the Storm Framework which was designed to process data streams. The idea to analyze the spatio-temporal correlation among anomalies is possible since the position of all sensors are static and a sensor holds its nearest sensor at close range as seen in Figure 1.

The *Feature Selection (Input) Component* makes the connection to the data source which receives the data stream in a JSON format. It also selects the set of features for the next processes, see more about the SCATS data stream in 4.1.

In order to check a fixed time period of the vehicle traffic the *Data Segmentation Component* performs a segmentation of traffic flow of each traffic sensor according to a specific traffic time period $\Delta T2$ (e.g. 15, 30, 45 or 60 Minutes of traffic). A Fixed Sliding Window approach is applied and the segmentation process adds the most recent data point and discard the oldest one in the segment.

The *Data Summarization Component* summarizes the segment of a time period $\Delta T2$ by computing statistical measures, the mean (μ) and standard deviation (σ^2) (Equation 3), and Upper Bound Limit (Equation 4).

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{3}$$

The *Anomaly Detection Component* analyzes the traffic flow of each sensor and whether this component detects a value above the (upper bound limit) threshold, Equation 4, (i.e., the statistical model in this work considers solely the *upper bound limit* since there is not negative traffic flow), it considers that the sensor holds an anomalous event and send the event for further analysis to *SCA* component, otherwise the component discards the event, because our aim is to analyze the correlation among the sensors and their influence on the detection of traffic anomalies. The event is sent in the form of Equation 1; *cause* of the anomaly is the trigger condition of the anomaly detection component:

[4] openstreetmap.org

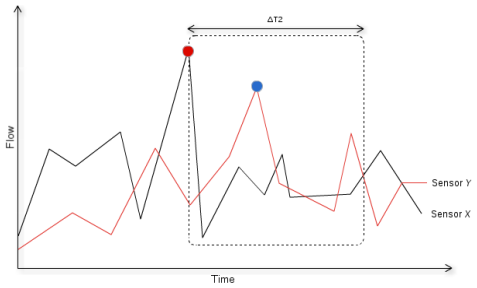

Fig. 2. SCA Approach.

'unexpected high traffic'.

$$Threshold = 3 \times \sigma^2 + \mu \tag{4}$$

The *Sensor Correlation Analysis (SCA) Component* analyzes the correlation among sensors by checking the spatio-temporal correlation among detected events and close sensors. SCA approach works as following: an event E takes a place at sensor S_x and whether during a time period $\Delta T1$ in the future (e.g. 30 Minutes) ($\Delta T1 \neq \Delta T2$) its nearest sensor S_y is affected by the event E, then the event E should be more reliable than the one which does not hold any correlation between two close sensors. The Equation 5 depicts the main principle of SCA component to check the correlation among sensors. The process to find nearest sensors makes use of Street Network data from OSM which is a kind of expert knowledge. The process queries the Street Network data every time an anomaly is sent to this component, the data is stored in PostgreSQL DB by applying the extension for spatial data called PostGIS. Whether the correlation does not hold true, the component discards the event. Figure 2 depicts the SCA approach.

$$Sen_x \Rightarrow Sen_y \Leftrightarrow E(t, Sen_x) \wedge E(t + \Delta t, Sen_y) \tag{5}$$

4. EXPERIMENTS

In order to check whether the SCATS sensors hold some spatio-temporal correlation in the process of anomaly detection we have performed some experiments which apply a statistical baseline model along with the SCA approach as well as compare the detected events against a ground truth. We also apply map matching by plotting both data to compare the results.

Dublin traffic agents such as AARoadWatch[5] and GardaTraffic[6] emit traffic alerts on Twitter. In our experiments these alerts (*Tweets*) are used as ground truth data and compared against the detected events in order to find out how much the SCATS sensors are correlated among them in the process of anomaly detection. On 26 June 2015 the traffic agents has informed 4 events about the traffic in Dublin. For instance, the alert "*DUB: Crash on D'Olier St before College St. This will add to delays in the area.*" was emitter by *AA Roadwatch* at *09:25*.

4.1 Data source

The Sydney Coordinated Adaptive Traffic System (SCATS) is an adaptive urban traffic management system that synchronizes traffic signals to optimise traffic flow across a network [McCann 2014]. SCATS data is time series, because SCATS sensors measure the traffic flow and density over the time,

[5] http://www.theaa.ie/aa/aa-roadwatch.aspx/
[6] http://www.garda.ie/Controller.aspx?Page=111

Table I. Number of anomalous events according to the size of segment by applying SCA approach and not applying SCA (NoSCA) and the number of anomalous events suing SCA which match to any alert from the ground truth data (MGT).

Size	**15**	**30**	**45**	**60**
NoSCA	1929	5234	6210	6759
SCA	32	138	173	223
MGT	0	0	0	0

Table II. Comparing the detected anomalies by applying SCA against traffic alerts (GT) in order to check whether they match (MGT) as well as the percentage of loss candidates per day (LC).

Day	**17/06/2015**	**18/06/2015**	**19/06/2015**	**20/06/2015**	**21/06/2015**	**22/06/2015**
NoSCA	1849	1867	1755	2036	2362	2001
SCA	27	37	24	29	37	35
GT	30	32	9	6	4	6
MGT	0	0	0	0	0	0
LC	98.53%	98.01%	98.63%	98.57%	98.43%	98.25%

that is, it provides information about flow of vehicles and the rate of use (density) of the streets. In Dublin, 506 SCATS sensors are present in their 4 non-overlapped regions (CCITY, NCITY, SCITY and WCITY). The SCATS data stream is emitted in a JSON format and it is high-dimensional with 74 features. However, this work uses a small set of features as follows: *sensor number*, *timestamp*, *latitude*, *longitude* and *flow*, because our approach evaluates the flow of sensor and uses coordinates to find its nearest sensor. The feature selection occurs in the data stream component as can be seen in Section 3. In our experiments we have used SCATS data stream which was measured from 17 to 22 June 2015 as well as 26 July 2015.

4.2 Results

The Table I depicts the number of anomalous events according to the size of segment on 26 June 2015 by applying the SCA approach and without SCA (NoSCA) as well as how many anomalous events (by using SCA approach) match with traffic alerts from the ground truth data at the same day. The result indicates that different segment sizes do not influence the SCA approach in the process of anomaly detection, and thus we evaluate the traffic flow by applying a 15 Minutes segment. Table II presents the result of the detection of anomalies by applying the SCA approach from 17 to 22 June 2015 as well as describes whether any anomaly detected by SCA matches (MGT) with any traffic alert (GT) which was emitter by traffic agents on the same time period. The percentage of loss candidates is also presented and it describes a high rate of loss. None anomaly detected by SCA approach has matched with the traffic alerts as in the experiment performed on 26 June 2015.

Figure 3 shows the map matching between the detected anomalies by applying SCA and the traffic alerts from traffic agents on 26 June 2015. The magenta dots and lines describe the events which are informed by traffic agents in Dublin and the red dots are the anomalous events detected by checking the spatio-temporal correlation among the sensors (SCA). The percentage of loss candidates by applying the SCA approach is 98.34%, that is, only 1.65% of the candidates are considered as anomalous events by the spatio-temporal correlation among SCATS sensors. Considering the low number of events provided by the ground truth such a drastically reduction might be a good, but another reliable source should be considered in order to check the candidates which are discarded in the process. Figure 4 shows the number of anomalies per hour by applying SCA approach in 3 different days which describes the SCATS sensors correlate more among them at night than in the morning or in the afternoon, that is, low traffic flows make the SCATS traffic sensors be more correlated among them. Therefore, considering all results the use of SCA approach is unfortunately poor for detection

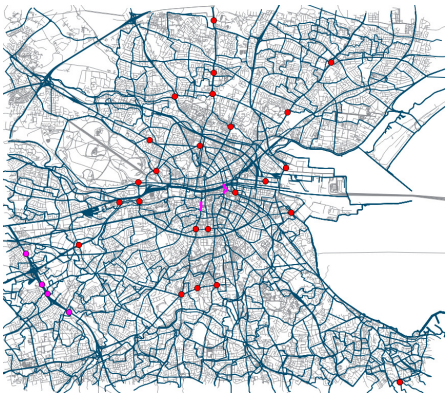

Fig. 3. Comparing ground truth data against the detected events by using SCA approach on 26 June 2015. The magenta dots and lines describe the events which are informed by traffic agents in Dublin and the red dots are the anomalous events detected by using SCA approach.

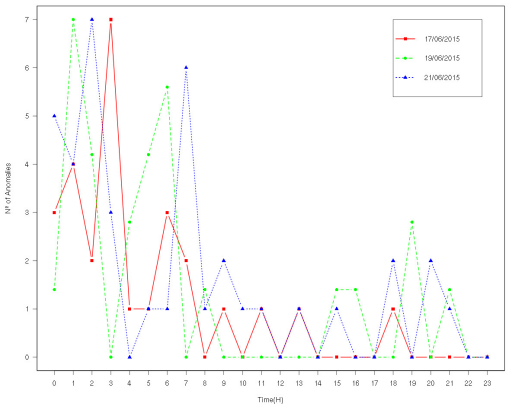

Fig. 4. Number of detected anomalies on 17, 19 and 21 June 2015 by applying SCA approach.

of traffic anomalies.

5. CONCLUSIONS

This work analyzes the spatio-temporal correlation among SCATS sensors in order to find whether such a correlation might support in the process of anomaly detection in an unsupervised manner. However, according to our results the sensors hold a strong correlation at night, but in the morning and in the afternoon such a correlation is weak. We also compare the anomalous events detected (by applying SCA approach) against the traffic alerts which are emitted by traffic agents in Dublin on Twitter. Unfortunately, none of the anomalies have matched with any of the 90 traffic alerts from 17 to 22 June 2015 as well as on 26 June 2015. Therefore, the spatio-temporal correlation among SCATS sensors (SCA approach) is poor for detection of traffic anomalies on static sensors. For future works, we intend to work on an online version of Core Vector Machine (CVM) with uses expert knowledge and traffic alerts to detect anomalies.

Acknowledgements

This research was supported by the National Council for Scientific and Technological Development (CNPq), the European Union's Seventh Framework Programme under grant agreement number FP7-318225, INSIGHT and from the European Union's Horizon 2020 Programme under grant agreement number H2020-ICT-688380, VaVeL. Additionally, this work has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, project A1.

REFERENCES

Aggarwal, C. *Outlier Analysis*. Vol. 1. Springer, New York, 2013.

Artikis, A., Weidlich, M., Schnitzler, F., Boutsis, I., Liebig, T., Piatkowski, N., Bockermann, C., Morik, K., Kalogeraki, V., Marecek, J., Gal, A., Mannor, S., Gunopulos, D., and Kinane, D. Heterogeneous stream processing and crowdsourcing for urban traffic management. In *Proc. 17th International Conference on Extending Database Technology (EDBT), Athens, Greece, March 24-28, 2014*. OpenProceedings.org, pp. 712–723, 2014.

Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B. *Data Stream Mining: A Practical Approach*. The university of Waikato, 2011.

Bădoiu, M. and Clarkson, K. L. Optimal core-sets for balls. *DIMACS Workshop on Computational Geometry*, 2002.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* 41 (3): 15:1–15:58, July, 2009.

Guo, J., Huang, W., and Williams, B. M. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. *Transportation Research Part C: Emerging Technologies*, July, 2014.

Liebig, T., Xu, Z., and May, M. Incorporating mobility patterns in pedestrian quantity estimation and sensor placement. In *Citizen in Sensor Networks*. Springer Berlin Heidelberg, pp. 67–80, 2013.

Liu, S., Chen, L., and Ni, L. M. Anomaly detection from incomplete data. *ACM Trans. Knowl. Discov. Data* 9 (2): 11:1–11:22, Sept., 2014.

Liu, S., Ni, L. M., and Krishnan, R. Fraud detection from taxis' driving behaviors. *IEEE Transactions on Vehicular Technology* 63 (1): 464–472, Jan., 2014.

Liu, W., Zheng, Y., Chawla, S., Yuan, J., and Xing, X. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. ACM, New York, NY, USA, pp. 1010–1018, 2011.

McCann, B. A review of scats operation and deployment in dublin. Tech. rep., ntelligent Transportation Systems, Dublin City Council, Wood Quay, Dublin, 2014.

Pan, B., Zheng, Y., Wilkie, D., and Shahabi, C. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL'13. ACM, New York, NY, USA, pp. 344–353, 2013.

Pang, L. X., Chawla, S., Liu, W., and Zheng, Y. On detection of emerging anomalous traffic patterns using gps data. *Data Knowl. Eng.* vol. 87, pp. 357–373, Sept., 2013.

Schnitzler, F., Liebig, T., Mannor, S., Souto, G., Bothe, S., and Stange, H. Heterogeneous stream processing for disaster detection and alarming. In *IEEE International Conference on Big Data*. IEEE Press, pp. 914–923, 2014.

Souto, G. and Liebig, T. On event detection from spatial time series for urban traffic applications. In *Solving Large Scale Learning Tasks: Challenges and Algorithms*, S. Michaelis, N. Piatkowski, and M. Stolpe (Eds.). Springer International Publishing, pp. (to appear), 2015.

Stolpe, M., Bhaduri, K., Das, K., and Morik, K. Anomaly detection in vertically partitioned data by distributed core vector machines. *ECML PKDD - Lecture Notes in Computer Science* vol. 8190, pp. 321–336, 2013.

Trilles, S., nd Óscar Belmonte, S. S., and Huerta, J. Real-time anomaly detection from environmental data streams. In *AGILE 2015*, F. Bacao, M. Y. Santos, and M. Painho (Eds.). Lecture Notes in Geoinformation and Cartography. Springer International Publishing, pp. 125–144, 2015.

Yang, S., Kalpakis, K., and Biem, A. Detecting road traffic events by coupling multiple timeseries with a non-parametric bayesian method. *IEEE Transactions on Intelligent Transportation Systems* 15 (5): 1936–1946, March, 2014.

Yang, S. and Liu, W. Anomaly detection on collective moving patterns. *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing* vol. 7, pp. 291–296, October, 2011.