Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Lightweight traffic anomaly detection: A case study with SCATS volume data of Melbourne

Iman Taheri Sarteshnizi, Majid Sarvi, Saeed Asadi Bagloee, Neema Nassir,

and Zay Maung Maung Aye[1]

[1]Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Australia

Email for correspondence: itaherisarte@student.unimelb.edu.au

Abstract

In this paper, we evaluate performance of an anomaly detection framework with real traffic count data collected by SCATS (Sydney Coordinated Adaptive Traffic System) loop detectors in Melbourne. The goal is to detect anomalous daily volume profiles within temporally large historical traffic data utilizing a lightweight and parameter-free approach and use it for live applications. To achieve this, daily volume profiles are first compressed into two dimensions benefiting from the Principal Component analysis (PCA). Then, a parameter-free version of DBSCAN is applied to the data with unique days of the week. Results from more than 20 different locations in Melbourne are fully visualized and the advantages and disadvantages of the method are discussed. We found that, with this approach, anomalous volume profiles can be accurately detected in a wide range of spatiotemporal data without any pre-training, parameter setting, or using complex learning methods.

1. Introduction

Intelligent transportation systems (ITS) that incorporate advanced technologies such as sensors, smartphones, cameras, and communication networks can further enhance traffic safety, reduce congestion, and improve mobility (Kazemeini et al., 2022; Sarteshnizi et al., 2022; Zarei Yazd et al., 2022). Among these emerging technologies, loop detectors are also being vastly used and this enables the collection of large amounts of traffic data. Detection of anomalous parts within this data is of great importance as it is necessary for conducting downstream traffic-related tasks by authorities (Taheri Sarteshnizi et al., 2022).

Sydney Coordinated Adaptive Traffic System (SCATS) is widely used in major cities around the world, such as Melbourne, to improve mobility efficiency and safety. Within this system, loop detectors are installed at each intersection, and volume data (number of passing vehicles at each time interval) is one type of data they collect to adjust the signal timing (Yazdani et al., 2023). In addition to this aid, loop detector data can be also used for future planning and modeling to amend other aspects of traffic flow (Emami et al., 2019). However, prior to any other analysis or modeling, we need to assure that the data is healthy and normal since the faulty performance of loop detectors or rare and extreme events may significantly change the underlying distribution of data. These anomalous parts may be misleading for statistical or learning models as they do not correspond to the true and periodic behavior of traffic.

Numerous methods are developed in the literature to address anomaly detection in time series data (Goswami et al., 2022) and specifically urban traffic data (Kalair & Connaughton, 2021). In some cases, authors believe that statistical methods are still surprisingly outperforming recent deep learning and machine learning methods (Nakamura et al., 2023). Furthermore, the evaluation of anomaly detection tasks is also shown to be challenging and controversial as there is no exact and totally reliable ground truth label set for anomalies in time series data (He et

al., 2023). Since anomalies are rare, the problem of anomaly detection is always integrated with an imbalanced dataset, and it is also shown that evaluation metrics also do not perfectly reflect the performance even if we have access to the ground truth (Hwang et al., 2022).

In addition to the hurdles mentioned previously, anomaly detection in spatiotemporal traffic data also proposes further challenges. Depending on the objectives, we may want to focus only on specific types of anomalies. Moreover, we may want to calibrate a model for single-site data, or we may need a single calibrated model to apply to data collected from multiple sites.

In this work, our focus is to mine large historical spatiotemporal traffic data (6 years long) and explore the performance of a light and fast anomaly detection framework (Taheri Sarteshnizi et al. (2023)) to deal with traffic data. The framework is able to successfully detect "complete anomalous daily volume profiles" without any pre-training, where traffic data is not aligned with historical data for a long period during the day. However, the results with this approach are shown to be satisfactory with data from a few numbers of sites. In this paper, we apply the framework to the data of more than 20 different locations within Melbourne, Australia and visually analyze its performance as a case study. It is shown that the proposed method can be successfully used to label spatiotemporal large volume datasets and prepare them for other down-stream tasks.

The contributions of this paper are listed below:

- We focus on "historical and large" spatiotemporal traffic data and our target is to detect totally anomalous profiles.
- A fast and parameter-free method is targeted in this paper to show the benefits of anomaly detection in this way.
- Instead of using metrics, we provide several visualizations of data in a systematic way to avoid ground truth-related problems.
- Evaluation is conducted based on data from more than 20 different locations surrounded by various types of land uses. Experts' opinions were utilized when choosing these locations.

2. Literature review

The detection of anomalous parts in traffic data is studied in different ways. The majority of the previous works are related to incident detection. The main goal of these studies is to promptly recognize and locate accidents that happen within a specific section of the road (specifically on highways). They generally benefit from supervised machine learning algorithms and incident reports recorded by local authorities. Several different algorithms like Logistic Regression (Agarwal et al., 2016), SVM (Xiao, 2019), XGBoost (Parsa et al., 2020), and Random Forest (Jiang & Deng, 2020) are tested and evaluated for this aim. Dealing with unbalanced data in supervised methods is a challenging task and different methodologies like SMOTE (Fang et al., 2020) are used in this regard. Furthermore, the authors used dimension reduction to avoid using unnecessary features in model training to make the procedure less complicated (Shang et al., 2021).

Research on unsupervised detection of anomalies in traffic data is also another direction of this domain. Authors in this area usually utilize trajectory data of connected vehicles traveling within a city and they are generally concerned with specific events happening in specific areas like concerts or sports matches (Gao et al., 2021). They partition the area of their interest into different grids and then use the data of each grid as input to their models. Prediction-based models, clustering approaches, and dimensionality reduction (Wang & Sun, 2021) are applied in these research papers.

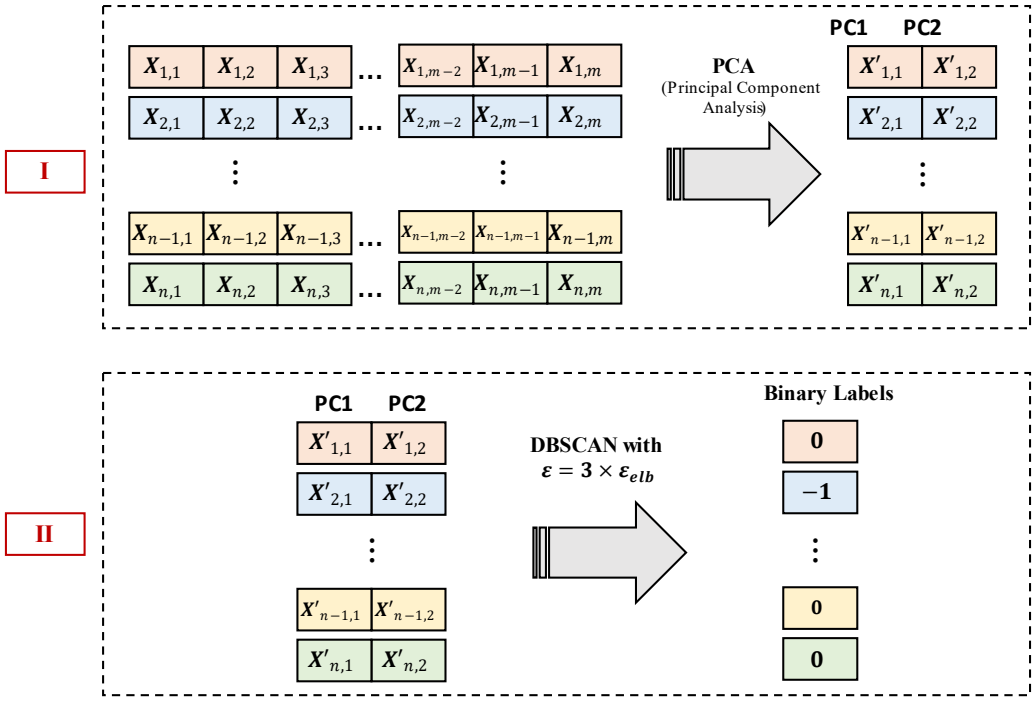
Overall, GPS data is predominantly utilized for capturing urban dynamics and detecting city-wide events. Conversely, loop detector data is employed for monitoring individual road segments to detect short-term anomalies, such as incidents. However, the detection of anomalous daily volume profiles in long historical data is not specifically investigated. On some occasions, a simple but effective method is needed to filter out the data which is not aligned with others. These anomalies may not be related to a major event, but they may have negative effects on future model training. Furthermore, model calibration is needed in almost all of the works in the literature on traffic data anomaly detection, however, we may not have enough time and resources to calibrate a model and find its parameters.

Therefore, to cover this gap, we focus on this problem and evaluate the performance of a lightweight and parameter-free method with data from more than 20 different locations. A comprehensive set of visualizations from the data of all locations is provided to completely reflect the performance of the model. With the plots of data provided in this paper, readers would be able to judge the performance on their own along with our discussion and decide if the method is suitable for their use case or not. To the best of our knowledge, this is the first work that contains a large number of figures representing the performance of a model on traffic volume data.

3. Method

Before describing the method, we should note that our aim is to detect anomalous "daily volume profiles". In other words, we intend to find profiles with anomalous points being the majority not the ones with a few minor anomalies. Therefore, volume profiles containing such minor (contextual) anomalies will not be our interest.

Figure 1: an overview of the method applied to the data.



A demonstration of the method used in this paper is shown in Figure 1. Traffic volume data is illustrated by $\boldsymbol{X}_{n,m}$ where n represents the day index and m is the timestep index on each specific day that the data point is collected. According to this figure, volume profiles are first compressed to two dimensions (PC1 and PC2) applying the PCA method (Taheri Sarteshnizi

et al., 2023). The number of principal components is chosen to be two, as in the main study, since it provides a good base for visualization and also is a good choice for DBSCAN. Furthermore, since we are detecting totally anomalous profiles, two dimensions are enough for this aim and help to keep the problem as lightweight as possible. It is also experimentally shown that with only two principal components, more than 90% of the variation within the data can be captured. Based on the theory of PCA, profiles with anomalous values will be located far from the others in the final 2-D space and we can benefit from this to determine the anomalies with the DBSCAN algorithm.

Applying DBSCAN (Ester et al., 1996) on the two-dimensional data prepared by the PCA needs two different parameters to be specified: $\boldsymbol{minPts}$ (the minimum number of points to create a cluster) and $\boldsymbol{\varepsilon}$ (the reachability distance). Although DBSCAN is one of the most powerful clustering and anomaly detection approaches adopted previously for 2-D data, the selection of its parameters becomes challenging when it comes to different applications. Selection of $\boldsymbol{minPts}$ can be achieved based on the size of data, however, $\boldsymbol{\varepsilon}$ cannot be determined before model implementation and needs extra effort to be specified. One popular approach to automate this selection is presented by Schubert et al. (2017). They suggest that, in some cases, $\boldsymbol{\varepsilon}$ can be determined by the following equation:

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{elb} \quad (1)$$

where $\boldsymbol{\varepsilon}_{elb}$ is the distance of the elbow point if we sort the points based on their distance to their $\boldsymbol{K^{th}}$ nearest neighbors. The elbow point is where the maximum curvature happens in this set of sorted points.

Based on experiments conducted with traffic data, it is found that Equation 1 should be revised as below to increase this value by a coefficient:

$$\boldsymbol{\varepsilon} = \boldsymbol{C} \times \boldsymbol{\varepsilon}_{elb} \quad (2)$$

and $\boldsymbol{C}$ is a constant coefficient proposed to enlarge the value suggested by the elbow method. It is experimentally shown that $\boldsymbol{C} = 3$ is a proper choice to be used in the proposed methodology in Figure 1 and it handles many real-world situations when it is applied. In this way, there is no need to specify the exact value of $\boldsymbol{\varepsilon}$, the most challenging parameter of DBSCAN. Furthermore, it is claimed that the performance is not sensitive to the choice of $\boldsymbol{minPts}$ and $\boldsymbol{K}$, and these parameters can be set to 30 and 5, respectively (for 6 years of data). For different, data sizes, we should certainly change the $\boldsymbol{minPts}$, however, minimum cluster size can be easily predicted. It should be also mentioned that DBSCAN should be applied to the data of each specific day of the week separately as the underlying pattern of data for each one may be unique. Therefore, after the PCA implementation, we separate the data from different days of the week (Monday and Tuesday to Saturday and Sunday) and then use DBSCAN with suggested parameters to determine anomalies. In this paper, we show that with the suggested parameter selection approach, the method can be automatically applied to the data of different locations with no requirement of any location-specific parameter adjustment.

After anomaly detection in historical data, normal data can also be used as a base for live anomaly detection. The distance of the streaming data to its counterparts in normal historical profiles will provide anomaly scores for the new upcoming data.

4. Data and results

To evaluate the method, we used one random leg data of 21 different intersections in Melbourne from 2014 to 2019 (6 years). The exact location of these stations on the map is shown in Figure 2. This type of data is being collected from more than 4800 intersections of Melbourne for all

the upcoming traffic flows to the intersections. Therefore, it is necessary to find and evaluate a parameter-free and lightweight method that is able to clean this big data and prepare it for further analysis. The data used in this study is collected every 15 minutes and therefore each daily volume profile includes 96 data points. Applying the PCA, we summarize the pattern of these 96 points in only two features. To avoid the model being affected by severe single-point anomalies we used B-spline fitting to smoothen the profiles. Other approaches like moving average filtering may be tested and replaced. Furthermore, we rescaled the profiles by dividing the values by the average of all values within the day.

Figure 3 showcases the performance of the PCA and DBSCAN architecture we described. According to the literature (Wu & Keogh, 2023), data visualization is still among the best approaches (if applicable) to validate anomaly detection and needs to be done, at least partially, even with the possibility of using other metrics like F1-score or AUC-ROC. To keep the results within a reasonable size, we only visualized the data collected on Mondays, Wednesdays, and Sundays for each location with anomalies being labeled in red color. The name of each intersection is also mentioned above plots with the first street or road being the main road where the data is collected and the second one showing where exactly the intersection is located. The direction of the flow is also shown afterward.

According to Figure 3, one important point with this approach is that along with an overall good detection of anomalies, it never misses severe anomalous profiles. No point can be found in these subfigures that is considerably far from the others and not selected as an anomaly. This is very important as we can make sure that the model certainly spots severely different profiles. However, the border anomalous-detected points around the normal region may require secondary evaluation depending on the downstream task.

Although we may find some questionable detections around the normal samples (like Figure 3 (r) Sunday), the opinion of observers in such situations may be heterogeneous and differ from one expert to another. In other words, it is also difficult for human eyes to specifically draw a precise line around the normal region and we will come across multiple results if we ask different people to do so. Therefore, the overall performance of the model is “reasonably well” considering that the model does not need any pre-training or parameter adjustment. The abnormal points in all 21 locations are spotted precisely if we generally look at all these plots and with this approach one can get a good label set using the data of any location.

Figure 2: Spatial distribution of 21 stations used for evaluation in this study.

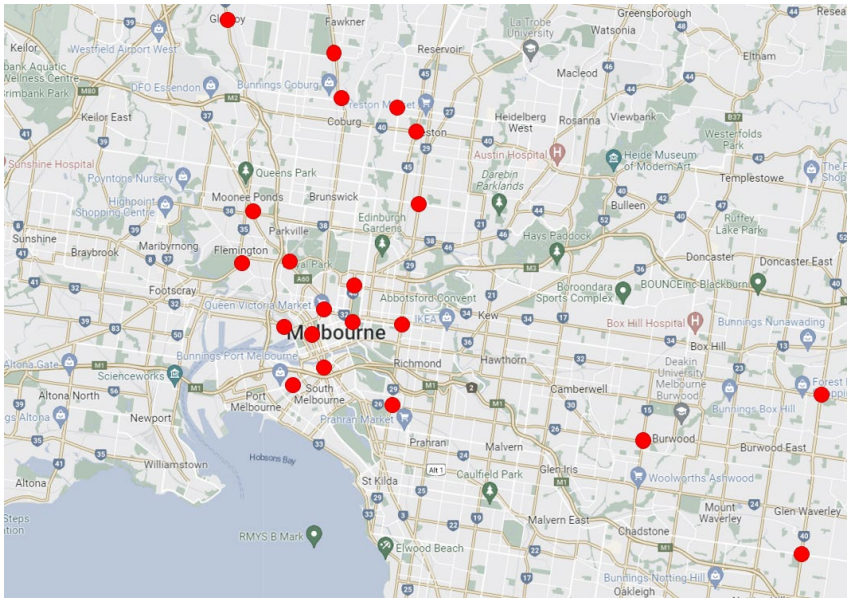


Figure 3: Detected anomalies by DBSCAN after dimension reduction with PCA in real traffic data

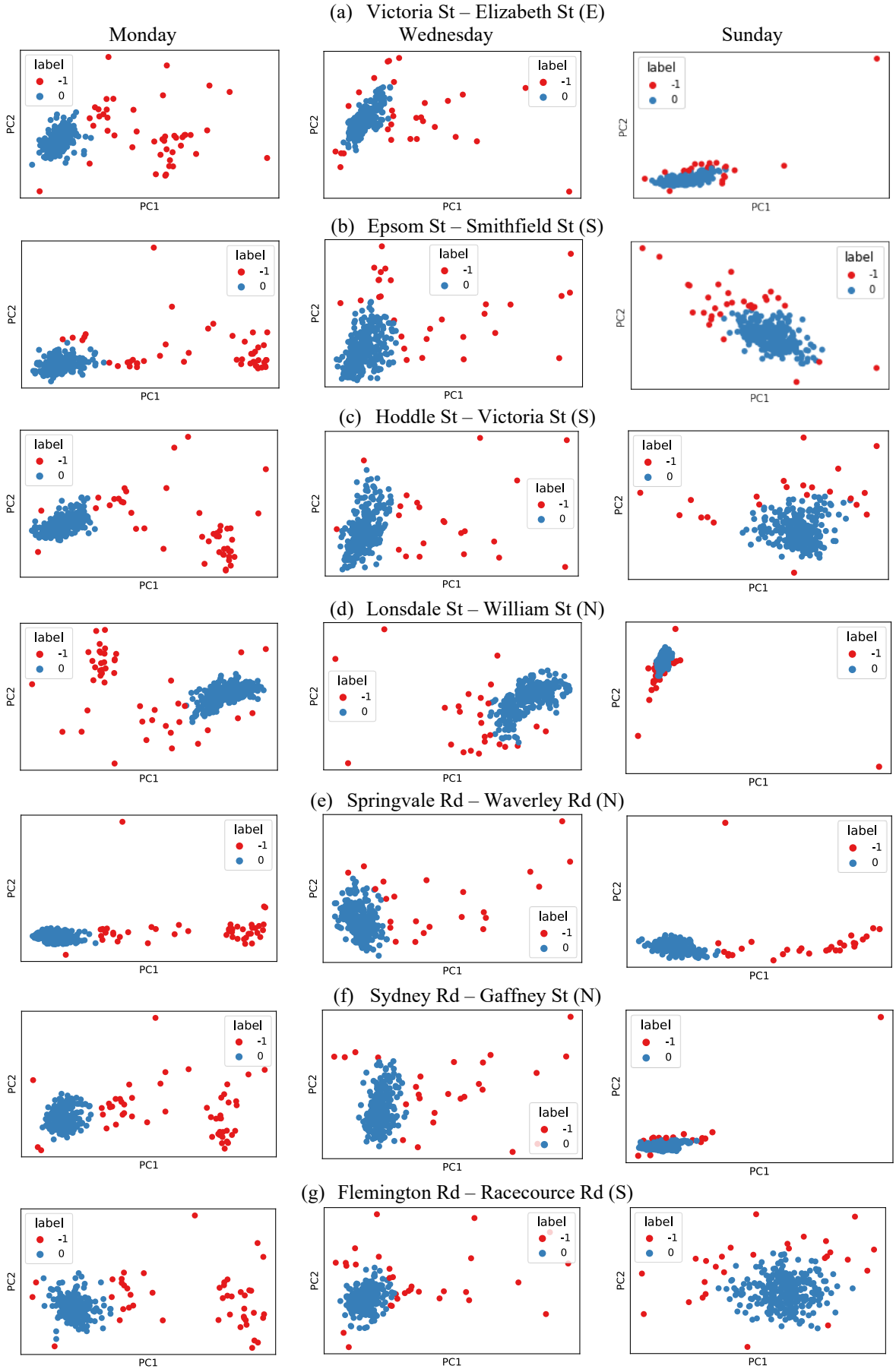


Figure 3 continued

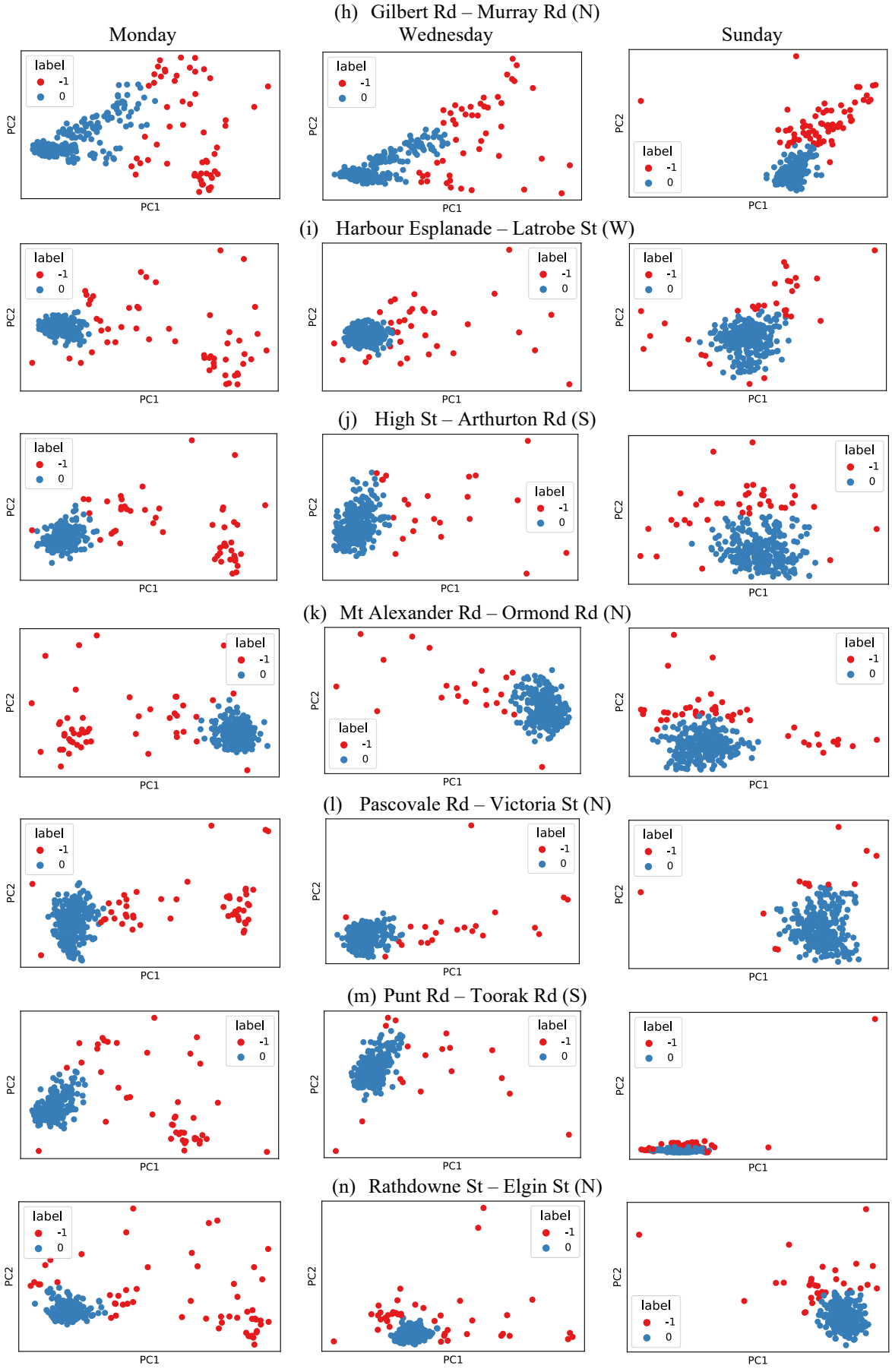
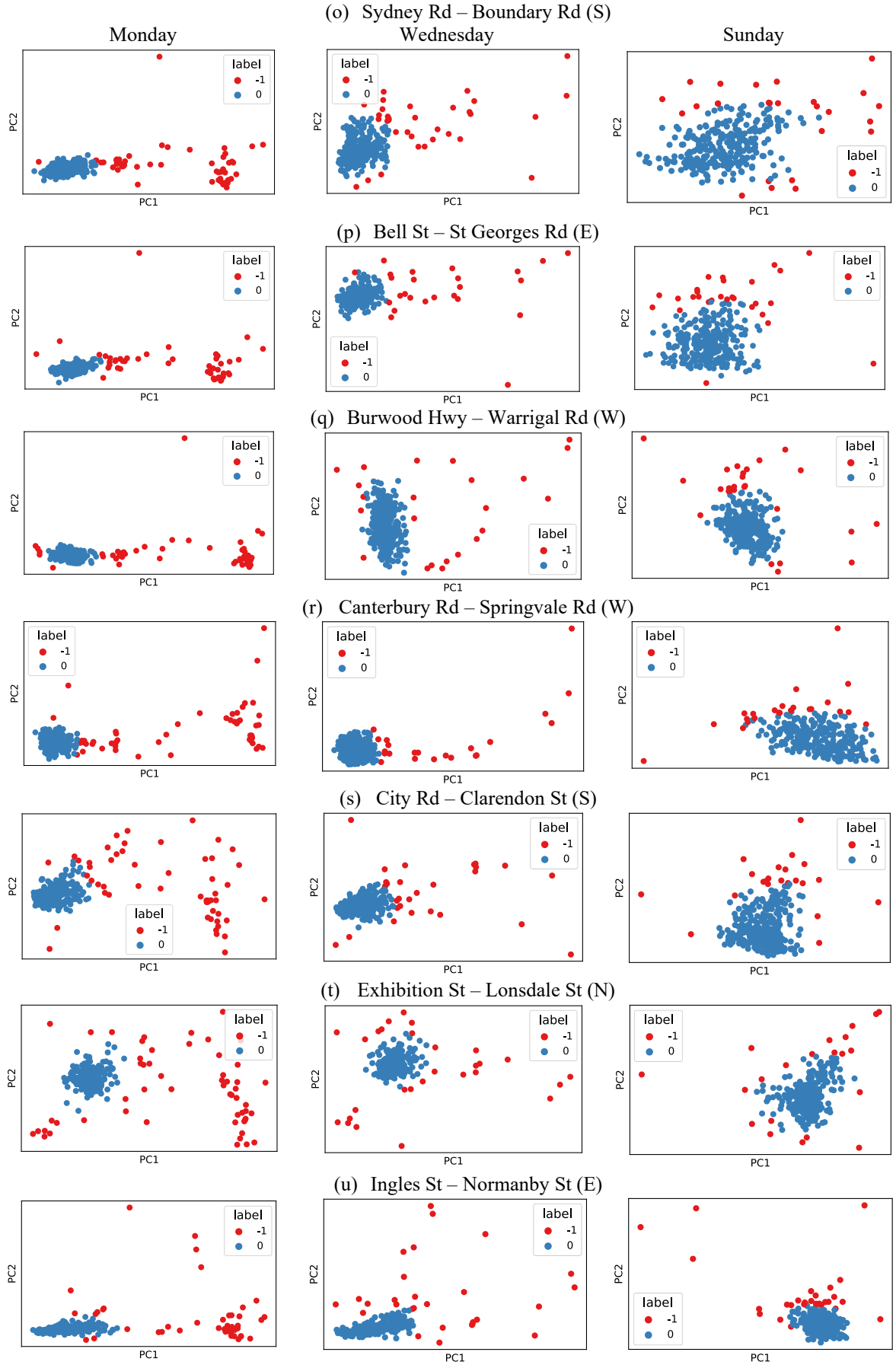
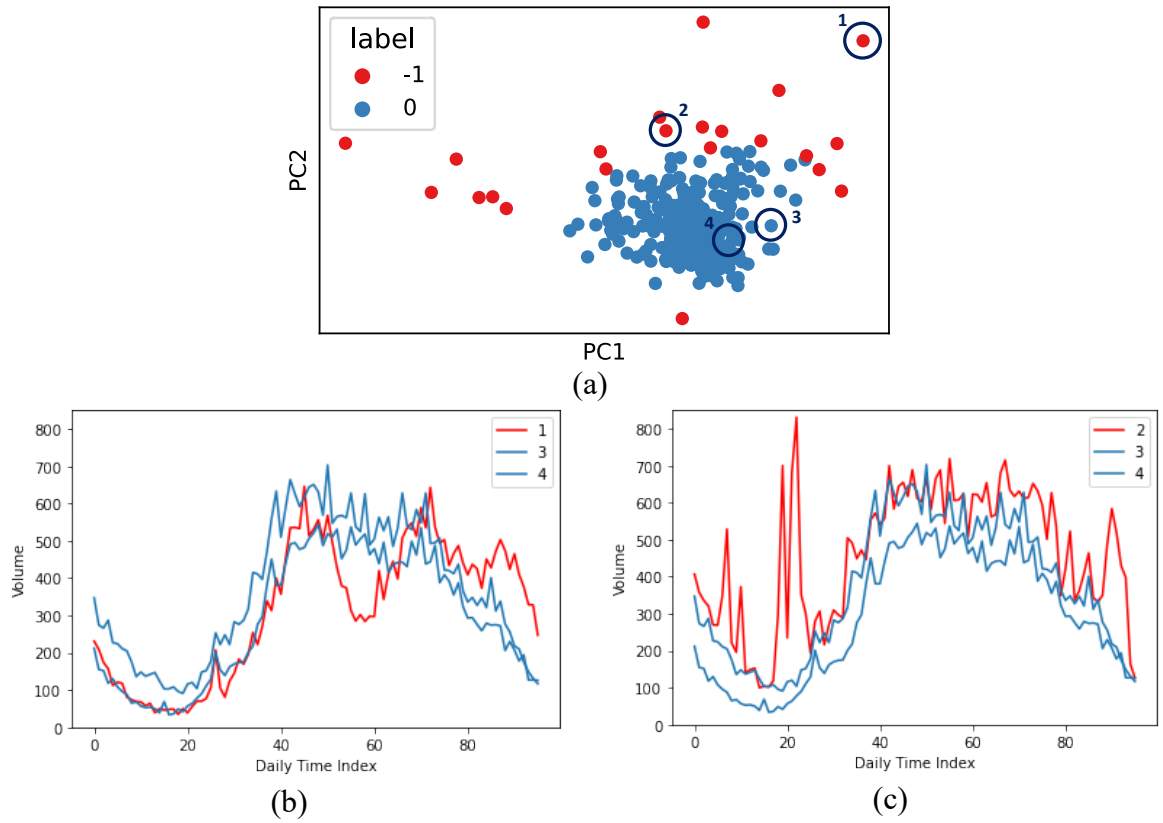


Figure 3 continued



In order to grasp an understanding of the PCA performance, Figure 4 is also provided as an example. We have spotted four different points on one arbitrary sample of a 2-D PCA latent space and visualized their original daily volume profiles separately for comparison. The difference between normal and anomalous profiles can be clearly observed in this figure, and it confirms the theoretical idea of PCA which is considerable different principal components for anomalous high dimensional data. Abnormal datapoints number 1 and 2 are fully demonstrated and anomalous intervals within those days can be easily recognised.

Figure 4: A close inspection of daily volume profiles mapped by the PCA into a 2-D space. Complete volume profiles related to the points spotted on (a) are demonstrated in (b) and (c). (Data derived from Hoddle St – Victoria St (S), Sunday)



5. Conclusion and future works

We propose a lightweight and fast anomaly detection approach tailored to historical and live traffic count data (known as SCATS traffic data). As a case study, we focused on anomalous daily volume profiles and represented the whole daily pattern of data with only two features with the help of the PCA. Then, we applied an adjusted parameter-free version of the DBSCAN to data related to each specific day of the week and detected the anomalies. The performance of this approach is verified with the data from 21 different locations in Melbourne. It was found that the performance is perfect when it comes to severe anomalous profiles, and it can also detect the boarders of a normal region with a reasonable accuracy. We concluded that this framework detects anomalies in big traffic data precisely in an offline manner and showed that it can be also used for live applications. For future work, testing the performance of the model with different data size scenarios is suggested.

References

Agarwal, S., Kachroo, P., & Regentova, E. (2016). A hybrid model using logistic regression

and wavelet transformation to detect traffic incidents. *IATSS Research*, *40*(1), 56–63. https://doi.org/10.1016/j.iatssr.2016.06.001

Emami, A., Sarvi, M., & Asadi Bagloee, S. (2019). Using Kalman filter algorithm for short-term traffic flow prediction in a connected vehicle environment. *Journal of Modern Transportation*, *27*(3), 222–232. https://doi.org/10.1007/S40534-019-0193-2/FIGURES/5

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Kdd (Vol. 96, No. 34, Pp. 226-231)*. www.aaai.org

Fang, Y. F., Yang, Q., Zheng, L., Zhou, X., Peng, B., & Nakano-Miyatake, M. (2020). A Deep Cycle Limit Learning Machine Method for Urban Expressway Traffic Incident Detection. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/5965089

Gao, J., Zheng, D., & Yang, S. (2021). Sensing the disturbed rhythm of city mobility with chaotic measures: anomaly awareness from traffic flows. *Journal of Ambient Intelligence and Humanized Computing*, *12*(4), 4347–4362. https://doi.org/10.1007/s12652-019-01338-7

Goswami, M., Challu, C., Callot, L., Minorics, L., & Kan, A. (2022). *Unsupervised Model Selection for Time-series Anomaly Detection*. http://arxiv.org/abs/2210.01078

He, D., Kim, J., Shi, H., & Ruan, B. (2023). Autonomous anomaly detection on traffic flow time series with reinforcement learning. *Transportation Research Part C: Emerging Technologies*, *150*, 104089. https://doi.org/10.1016/j.trc.2023.104089

Hwang, W. S., Yun, J. H., Kim, J., & Min, B. G. (2022). "Do you know existing accuracy metrics overrate time-series anomaly detections?" *Proceedings of the ACM Symposium on Applied Computing*, 403–412. https://doi.org/10.1145/3477314.3507024

Jiang, H., & Deng, H. (2020). Traffic incident detection method based on factor analysis and weighted random forest. *IEEE Access*, *8*, 168394–168404. https://doi.org/10.1109/ACCESS.2020.3023961

Kalair, K., & Connaughton, C. (2021). Anomaly detection and classification in traffic flow data from fluctuations in the flow–density relationship. *Transportation Research Part C: Emerging Technologies*, *127*(July 2020), 103178. https://doi.org/10.1016/j.trc.2021.103178

Kazemeini, A., Taheri, I., & Samimi, A. (2022). A GPS-based Algorithm for Brake and Turn Detection. *International Journal of Intelligent Transportation Systems Research*, *20*(2), 433–445. https://doi.org/10.1007/s13177-022-00301-9

Nakamura, T., Mercer, R., Imamura, M., & Keogh, E. (2023). MERLIN++: parameter-free discovery of time series anomalies. *Data Mining and Knowledge Discovery*, *37*(2), 670–709. https://doi.org/10.1007/s10618-022-00876-7

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, *136*, 105405. https://doi.org/10.1016/J.AAP.2019.105405

Sarteshnizi, I. T., Tavakkoli Khomeini, F., Khedri, B., & Samimi, A. (2022). Sensitivity analysis of driving event classification using smartphone motion data: case of classifier type, sensor bundling, and data acquisition rate. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. https://doi.org/10.1080/15472450.2022.2140048

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How Should (Still) Use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3). https://doi.org/10.1145/3068335

Shang, Q., Feng, L., & Gao, S. (2021). A Hybrid Method for Traffic Incident Detection Using

Random Forest-Recursive Feature Elimination and Long Short-Term Memory Network with Bayesian Optimization Algorithm. *IEEE Access*, *9*, 1219–1232. https://doi.org/10.1109/ACCESS.2020.3047340

Taheri Sarteshnizi, I., Sarvi, M., Bagloee, S. A., & Nassir, N. (2022). Abnormality Detection in Urban Traffic Data: A Review. *Australasian Transport Research Forum*. http://www.atrf.info

Taheri Sarteshnizi, I., Sarvi, M., Bagloee, S. A., & Nassir, N. (2023). Temporal pattern mining of urban traffic volume data: a pairwise hybrid clustering method. *Transportmetrica B*, *11*(1). https://doi.org/10.1080/21680566.2023.2185496

Wang, X., & Sun, L. (2021). Diagnosing Spatiotemporal Traffic Anomalies With Low-Rank Tensor Autoregression. *IEEE Transactions on Intelligent Transportation Systems*, 1–10. https://doi.org/10.1109/TITS.2020.3044466

Wu, R., & Keogh, E. J. (2023). Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *IEEE Transactions on Knowledge and Data Engineering*, *35*(3), 2421–2429. https://doi.org/10.1109/TKDE.2021.3112126

Xiao, J. (2019). SVM and KNN ensemble learning for traffic incident detection. *Physica A: Statistical Mechanics and Its Applications*, *517*, 29–35. https://doi.org/10.1016/j.physa.2018.10.060

Yazdani, M., Sarvi, M., Asadi Bagloee, S., Nassir, N., Price, J., & Parineh, H. (2023). Intelligent vehicle pedestrian light (IVPL): A deep reinforcement learning approach for traffic signal control. *Transportation Research Part C: Emerging Technologies*, *149*, 103991. https://doi.org/10.1016/J.TRC.2022.103991

Zarei Yazd, M., Taheri Sarteshnizi, I., Samimi, A., & Sarvi, M. (2022). A robust machine learning structure for driving events recognition using smartphone motion sensors. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. https://doi.org/10.1080/15472450.2022.2101109