Transport mode identification by clustering travel time data

Shen Liu[1] James McGree[2] Gentry White[3]
Wayne Dale[4]

(Received 16 March 2015; revised 14 May 2017)

Abstract

Travel time data of road users collected by Bluetooth scanners are of great value in traffic monitoring and planning. To estimate the travel time of road users over a segment of road, discriminating between different types of travellers is essential, but often overlooked by researchers. This paper explores the feasibility of transport mode identification using clustering methods. The performance of the k-means clustering algorithm and the Gaussian mixture model is examined via an empirical study of travel time data collected from road segments in the north Brisbane region, Queensland, Australia. It is demonstrated that both clustering methods are able to detect multiple transport modes and produce travel time estimates that are close to reality. The methods

DOI:10.21914/anziamj.v56i0.9420, Published May 28, 2017, as part of the Proceedings of the 2014 Mathematics and Statistics in Industry Study Group. ISSN 1445-8810. (Print two pages per sheet of paper.)

and results provide a guideline for transport mode identification, and may contribute to further issues related to traffic monitoring such as forecasting and planning.

Subject class: 82C70; 91C20

Keywords: travel time, transport mode, crisp clustering, fuzzy clustering, Bluetooth data

Contents

1 Introduction

Management of current roadways infrastructure and planning for maintenance and future growth are important projects for local and state governments [15]. The maintenance and construction of new roads are large undertakings, requiring considerable dedication of resources and potential disruption for many users. Making decisions about these activities requires a good understanding

of the patterns of usage for existing roadways and a means of planning and estimating the likely impact of future additions or modification of the existing network. For such purposes, research has been undertaken extensively to estimate travel time on motorways or arterials [2, 3, 4, 10, 17, 18, 19, 23].

The issue of monitoring road use for ongoing planning and efficiency of the road network in the state of Queensland was brought to the Mathematics in Industry Study Group 2014 by the Queensland Department of Transport and Main Roads (TMR). The TMR collects usage data by recording the movement of media access control (MAC) addresses from Bluetooth enabled devices through sensors placed at major intersections throughout Queensland. Tracking Bluetooth MAC addresses has gained much interest as one of the most cost effective ways of recording travel time [2, 3, 4]. MAC IDs of discoverable Bluetooth devices being transported by road users, such as mobile phones or the car systems designed to pair with user devices, are tracked by Bluetooth MAC scanners (BMS), and the travel time are easily recorded by matching the MAC IDs from one BMS to another. These records are referred to as the node-to-node travel time data, which measure the time difference between two BMS at two nodes collecting identical MAC address information. According to the TMR, these data are collected from around 20% of road users, which represent a sample of the actual traffic and are used to make inference about travel times throughout the network. The principles of Bluetooth communication and BMS data acquisition have been studied [2, 3].

The current travel time system of the TMR calculates a single travel time between nodes on the road network and delivers a single average result per period of calculation. This is carried out by computing an average of all observations after removing the outliers, where the outliers include those records below a pre-determined travel speed (usually 5 km/h), those records that appear off route, and those detected by the median absolute deviation (MAE) method [4]. However, travel time estimates obtained in this way may be biased [3]. The primary reason is that there are multiple transport modes of road users such as cars, heavy vehicles, buses, cyclists and pedestrians, but the BMS are not able to identify them as no information about the type of

the mode and number of devices within one mode is available. Although it is evident that means or medians of node-to-node travel time data tend to achieve good performance when concentrating on only one type of road users (e.g., cyclists [19], pedestrians [17], or motor vehicles on a freeway [18]), the reliability of means or medians when multiple transport modes are present is questionable. Since different transport modes are associated with potentially different patterns in travel times, data collected from different types of road users should not be treated as if they were homogeneous. In particular, different road users may travel at varying speeds in the data sample, which will influence the final estimation of travel time. For instance, during peak hours cyclists and pedestrians tend to be much less influenced by congestion compared to cars or buses, whereas during off-peak hours cars are expected to travel faster than cyclists. As a result, it would be more reasonable to estimate travel times separately for various types of road users that utilize the entire network. To achieve this, different transport modes need to be identified from the recorded data in the first place. Since the recorded MAC addresses do not provide any information about the corresponding transport mode, the aim of this paper is to develop an algorithm to model and distinguish multiple travel patterns in node-to-node travel times, with multiple road user groups identified [15].

Numerous methods have been proposed for data grouping purposes, which are categorised into supervised learning (known as classification) and unsupervised learning (known as clustering). The former is applied when the grouping of data is known, while the latter is applicable when the grouping is unknown [13, 14]. As MAC addresses do not show grouping information of road users, cluster analysis of the node-to-node travel time data is carried out in this paper. There are two types of clustering techniques, namely, crisp clustering and fuzzy clustering. Crisp clustering divides data into crisp clusters, where each individual belongs to exactly one cluster. In contrast, fuzzy clustering may determine that an item belongs to more than one cluster, producing degrees of membership that indicate the extent to which such an item belongs to those clusters. D'Urso and Maharaj [6] claimed that crisp clustering may not

be appropriate in practical situations, since in many cases there is no definite boundary between clusters and hence fuzzy clustering appears to be the better option. This claim is relevant to our study, as the boundary of travel time between different types of road users may be vague. For instance, during peak hours cyclists may travel at a speed very close to, or even higher than, that of cars, while during off-peak hours the travel time of a bus over a segment of road might be rather similar to an ordinary motor vehicle. In this study, both crisp and fuzzy clustering techniques are employed, and a comparison between them is carried out to indicate which one is more appropriate in studying BMS travel time data. In particular, we consider the k-means algorithm as a means of crisp clustering and the Gaussian mixture model as a means of fuzzy clustering. Both methods have been widely applied [11]. Section 2 briefly describes these two methods.

2 Methods

To estimate and predict travel time for various transport modes accurately, groups of road users need to be identified in the first place. We propose to use unsupervised learning techniques to study the grouping of travel times for the following reasons.

1. MAC addresses do not indicate which transport modes are in operation during a particular period of time, and transport patterns may show large variability over time. Thus, it is not feasible to pre-determine the groupings.
2. Supervised learning depends heavily on historical information, which may not always be representative in transport research since travel patterns are often influenced by external facts such as weather and incidents.

We consider both crisp and fuzzy clustering in this study. In the following subsections, the k-means algorithm and the Gaussian mixture model are briefly

described, respectively. As both methods are unsupervised non-hierarchical approaches, one needs to determine the number of clusters beforehand. The cluster number selection criterion for each of the two methods is also discussed.

2.1 k-means algorithm

The k-means algorithm, proposed by MacQueen [16], is probably the most widely used non-hierarchical clustering method. Let $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,m})$ be the ith object characterised by m variables, and let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of n objects. Denote U an $n \times k$ partition matrix with binary elements $u_{i,l} = 1$ indicating that object i belongs to cluster l and 0 otherwise. Denote $Z = (Z_1, Z_2, \ldots, Z_k)$ a set of k vectors representing the centroids of the k clusters. Let $d(x_{i,j}, z_{l,j})$ be a distance measure between object i and the centroid of cluster l on the jth variable. For numeric variables, $d(x_{i,j}, z_{l,j})$ is often the L^2-norm; for categorical variables, $d(x_{i,j}, z_{l,j}) = 0$ if $x_{i,j} = z_{l,j}$ and 1 otherwise. The k-means algorithm searches for a partition of X into k clusters which minimises the objective function P with unknown variables U and Z as

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{i,l} d(x_{i,j}, z_{l,j})$$

subject to

$$\sum_{l=1}^{k} u_{i,l} = 1, \quad 1 \leqslant i \leqslant n.$$

To solve the optimisation problem in the k-means algorithm, the following two steps are taken in each of the repeated loops:

1. Fix $Z = \hat{Z}$, and solve the reduced problem $P(U, \hat{Z})$ as follows:

$$u_{i,l} = 1 \quad \text{if } \sum_{j=1}^{m} d(x_{i,j}, z_{l,j}) \leqslant \sum_{j=1}^{m} d(x_{i,j}, z_{t,j}) \quad \text{where } 1 \leqslant t \leqslant k;$$
$$u_{i,t} = 0 \quad \text{for } t \neq l.$$

2. Fix $U = \hat{U}$, and solve the reduced problem $P(\hat{U}, Z)$. For numerical variables, the reduced problem is solved as follows:

$$z_{l,j} = \frac{\sum_{i=1}^{n} u_{i,l} x_{i,j}}{\sum_{i=1}^{n} u_{i,l}} \quad \text{where } 1 \leqslant l \leqslant k,\ 1 \leqslant j \leqslant m.$$

For categorical variables, the reduced problem is solved as follows:

$$z_{l,j} = a_j^r,$$

where a_j^r is the mode of the variable values in cluster l.

In summary, the k-means algorithm partitions n objects into k clusters. Each object is allocated to the cluster with which the dissimilarity measure is the smallest. Each time an object changes clusters, the centroids of both its old and new clusters are updated. This algorithm stops when the optimisation problem is solved.

As the k-means algorithm requires a predetermined number of clusters, we employ the Silhouette coefficient proposed by Rousseeuw [21]. The Silhouette coefficient is a function of both cohesion and separation of clusters, implying that both within-cluster variation and inter-cluster distance are taken into account. For each observation a Silhouette coefficient is computed, and the average value of the Silhouette coefficients of all individuals is used as an overall measure of clustering performance. If the number of clusters is unknown, then the k-means algorithm is subject to a range of possible values, and the one leading to the highest average Silhouette coefficient is considered as the optimal number of clusters.

2.2 Gaussian mixture model

The Gaussian mixture model is applied frequently as a means of clustering [5, 7, 8, 20]. Banfield and Raftery [1] defined the term *model-based cluster analysis* for clustering based on finite mixtures of Gaussian distributions

and related methods. We briefly describe the Gaussian mixture model as follows. Assume that real-valued observations $X_1, \ldots, X_n$ are modelled as independently identically distributed (i.i.d.) with the density

$$f(X; \theta) = \sum_{j=1}^{G} \pi_j \phi(X_i; \mu_j, \sigma_j^2),$$

where G denotes the number of mixture components, $\phi(\cdot; \mu_j, \sigma_j^2)$ is the density of the jth Gaussian distribution with mean μ_j and variance σ_j^2, and π_j is the proportion of the jth mixture component satisfying $\sum_{j=1}^{G} \pi_j = 1$. If G is unknown, then the Bayesian information criterion [22] is employed as a standard estimation method [9]. θ denotes the parameter vector containing all proportions, means and variances, which are estimated by the maximum likelihood estimator defined as

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(X_i; \theta),$$

where $\Theta = (\theta \mid \sigma_j^2 \geqslant s,\ j = 1, \ldots, G,\ \sum_{j=1}^{G} \pi_j = 1)$ for some choice of $s > 0$ that represents the lower bound of variances to avoid degeneracy of the log-likelihood function. The maximum likelihood estimation is usually carried out using the expectation-maximisation (EM) algorithm [20]. Once $\hat{\theta}_n$ is obtained, the posterior probability that the ith observation X_i was generated by the pth component is computed by

$$\hat{\tau}_{i,p} = \frac{\hat{\pi}_p \phi(X_i; \hat{\mu}_p, \hat{\sigma}_p^2)}{f(X_i; \hat{\theta})},$$

and for n observations to be clustered into G groups, an $n \times G$ posterior matrix is

$$\hat{\boldsymbol{\tau}} = \begin{bmatrix} \hat{\tau}_{1,1} & \ldots & \hat{\tau}_{1,G} \\ \vdots & \ddots & \vdots \\ \hat{\tau}_{n,1} & \ldots & \hat{\tau}_{n,G} \end{bmatrix}.$$

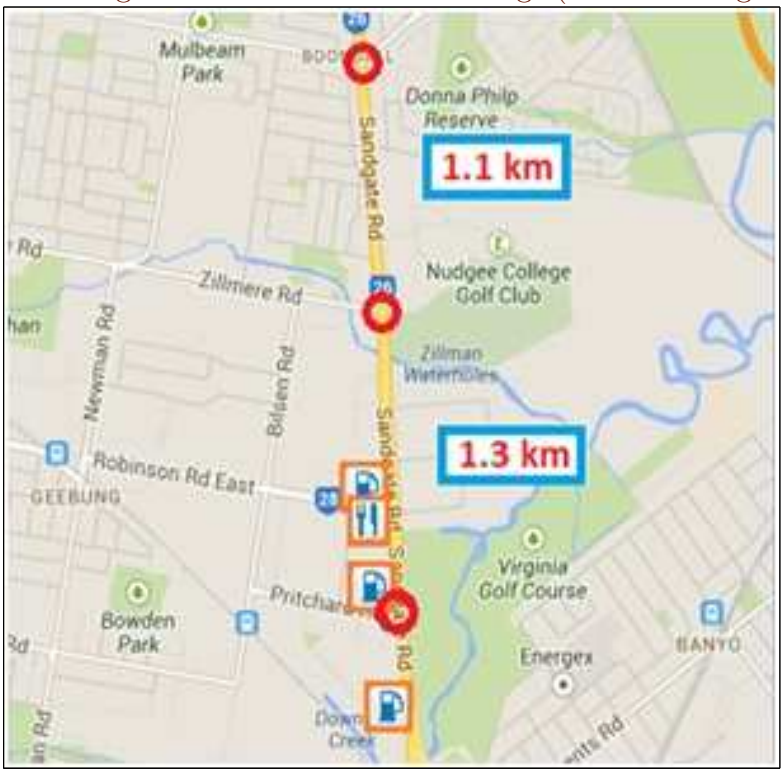
The matrix above is used to determine the degrees of membership, that is, to what extent an individual is believed to belong to each of the clusters. This implies that particular observations are allowed to belong to more than one cluster with different membership degrees, with the associated fuzziness taken into account. D'Urso and Maharaj [6] recommended such a fuzzy clustering approach when the boundary of clusters is not clear-cut, as for those individuals close to the boundary it is more plausible to consider their groupings in terms of membership degrees rather than in terms of total membership versus non-membership. Consequently, the fuzzy clustering exhibits greater adaptivity in defining the prototypes, that is, the representatives of clusters. Such an advantage becomes substantial when two or more groups of data show similar patterns.

3 Data and research design

The data used in this study were observed by the TMR on 12 Nov 2013 (Tuesday) from multiple Bluetooth MAC address sensors, which were located at the intersections on Sandgate Road with Pritchard Road, Zillmere Road and Beams Road, in north Brisbane, Queensland (Figure 1). The link from the Pritchard Road intersection to the Zillmere Road intersection is labelled as Link 1260, whereas the link from the Zillmere Road intersection to the Beams Road intersection is labelled as Link 1262. Both links are outbound from the CBD of Brisbane tracking north, and each link has two bus stops. Link 1260 is 1.3 kilometres long with two intersections that have traffic lights in operation, whereas Link 1262 is 1.1 kilometres long with one intersection that has traffic lights in operation. The speed limit is 70 km/h for both links. Two service stations and a fast food outlet are located along Link 1260. The locations of Bluetooth MAC address sensors, the lengths of the two links and service/food facilities are labelled on the map shown by Figure 1.

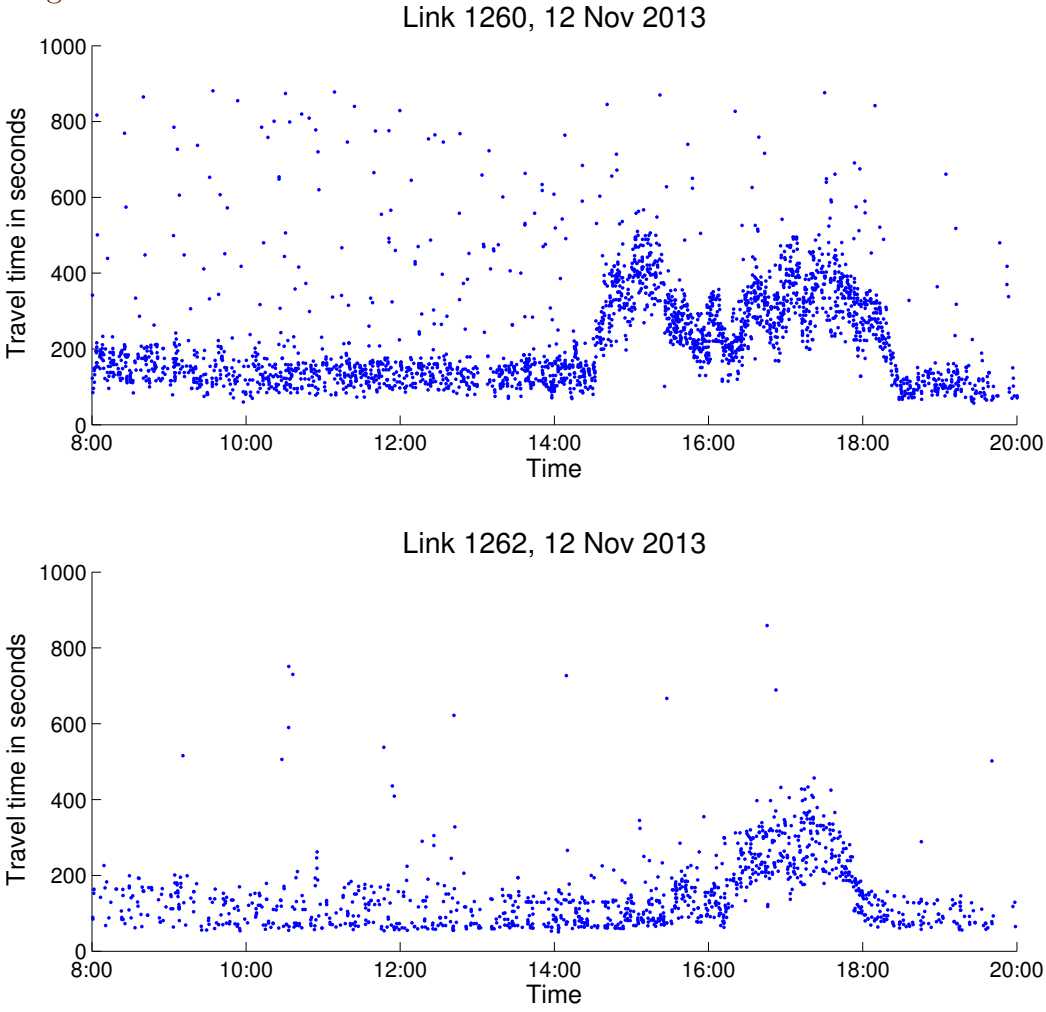
For Link 1260, a total of 3303 valid MAC address pairs were scanned on 12 Nov 2013, and hence 3303 travel times were obtained. For Link 1262, the

Figure 1: Sandgate Road and its surroundings (Source: Google Maps)



sample size is 1424. Figure 2 displays the individual travel times recorded between 8:00 and 20:00 on 12 Nov 2013 for different road users on Link 1260 and 1262. Peak and off-peak hours are displayed by the figure. As both links are outbound from the Brisbane city, peak hours emerge in the afternoon. For Link 1260, the plot shows two peaks. The first peak is between hours from 14:30 to 16:00, where the majority of road users took longer to travel through Link 1260. This is because of the school zone operating times for the Geebung Preschool and St Kevin's Catholic Primary School, which are located around 500 metres away from Sandgate road. During the afternoon school zone time period, motor vehicles picking up students from these two

Figure 2: Individual travel times over Link 1260 and 1262 on 12 Nov 2013



schools may join Sandgate road from Pritchard Road or Robinson Road East (State Route 28), causing delay in travel time over Link 1260. The second peak is mainly because of people getting off work in late afternoon, which starts with an increase in travel time at around 16:30, reaching the maximum at around 17:00–17:30 and ending with a rapid drop at around 18:30. For Link 1262, 16:00–18:00 appeared to be peak hours on that day as the travel time during this period was noticeably higher than the other hours.

The clustering process using the k-means algorithm and Gaussian mixture model is carried out based on 15 minute consecutive time intervals from 8:00 to 20:00. That is, travel time data recorded from 8:00 to 8:15 are clustered first, followed by those recorded from 8:15 to 8:30, and so on. Before fitting the Gaussian mixture model, the natural log of the recorded travel times was taken. For each of the 15 minute time intervals, the optimal number of clusters is determined by the Silhouette coefficient and the Bayesian Information Criterion, and then clusters of data collected over the 15 minute period are determined in an unsupervised manner. The determined clusters are used to estimate the travel time for various transport modes. In particular, we concentrate on the average travel time estimation for cars. During off-peak hours, cars are generally believed to be the quickest group of road users on average, but this is not always the case during peak hours. For instance, in congestion it is not rare to observe cyclists travelling at a faster speed than cars. As a consequence, while cars are identified as the group that has the shortest travel time during off-peak hours, we assume the largest cluster is representative of traffic conditions during peak hours and hence the average travel time of cars is estimated based on this cluster.

To justify the clustering results, two evaluations were conducted. The first evaluation aims at comparing the estimated travel time produced by the clustering methods to some counterpart. To approximate real travel time of vehicles, the TMR also collected spot speed data of Link 1260, namely, data of vehicle speed collected at a single node on the link. Given the length of the link, the spot speed data are converted to travel time in seconds, and then averages are taken over consecutive 5 minute time intervals from 8:00

to 20:00 as approximations of real travel time over the day. The comparison between the results from clustering and the spot speed data helps determine if the two clustering approaches achieve reliable performance.

The second evaluation aims at examining to what extent the produced clusters of road users are consistent over different road segments. If the travel time of a road user is repeatedly observed over two segments of a road, then we expect that data from both segments will imply the same grouping of such a road user. Consequently, a clustering method is said to be relatively consistent if a relatively high proportion of repeatedly observed road users are clustered into the same group over different segments. To carry out this evaluation, repeatedly observed road users on Link 1260 and Link 1262 are identified using the MAC addresses. The groupings of these road users on both links are recorded, and the proportion of consistent groupings is calculated.

4 Empirical results

Figure 3 and Figure 4 display the clusters determined by the two clustering methods for Link 1260 and 1262, respectively. The colour of each scatter point indicates which cluster it should belong to. For the k-means algorithm, a single colour is assigned to the entire cluster as per the crisp clustering principle. For the Gaussian mixture model, the colour is assigned to each individual by its posterior probabilities values, which coincide with the RGB colouring function in MATLAB ([1, 0, 0], [0, 1, 0] and [0, 0, 1] correspond respectively to red, green and blue). For instance, if the Gaussian mixture model determines posterior probabilities [0, 0.01, 0.99], then the colour of the corresponding scatter point is close to pure blue if [0.5, 0.5, 0] are computed, then the colour of that individual is somewhere between red and green. For both clustering methods, the order of the groups is rearranged so that the colouring is consistent with the speed, that is, the red colour always represents the quickest group, with green and blue groups being slower.

Figure 3: Individual travel times over Link 1260 on 12 Nov 2013

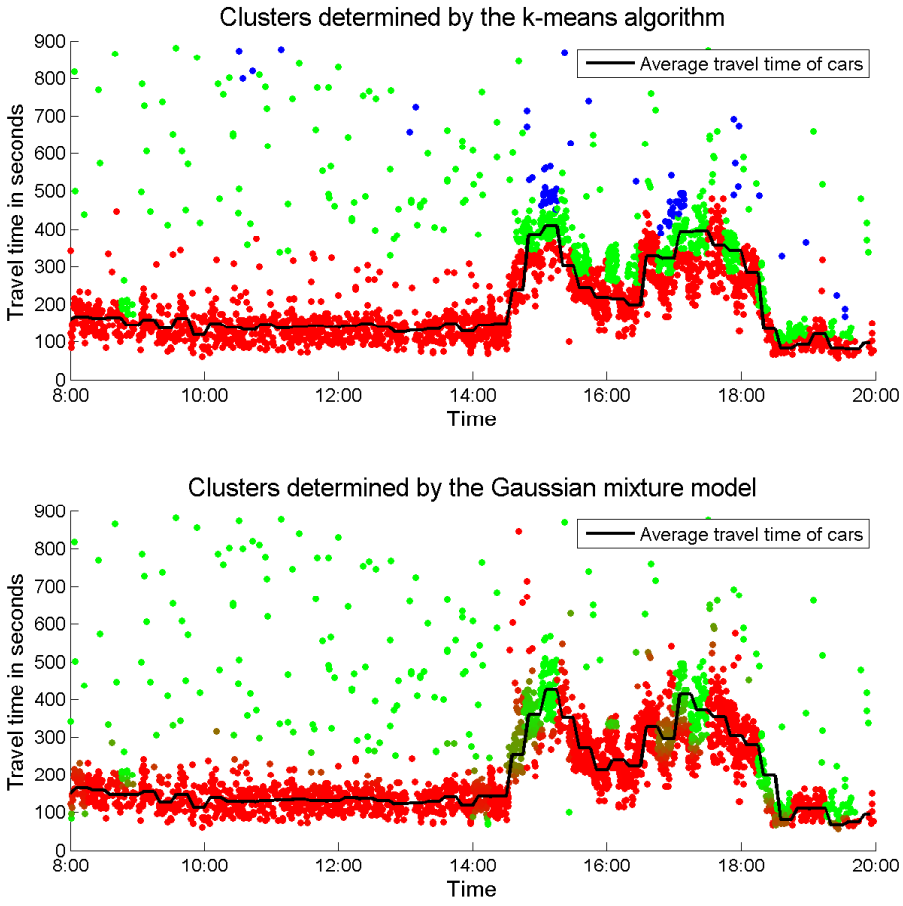


Figure 4: Individual travel times over Link 1262 on 12 Nov 2013

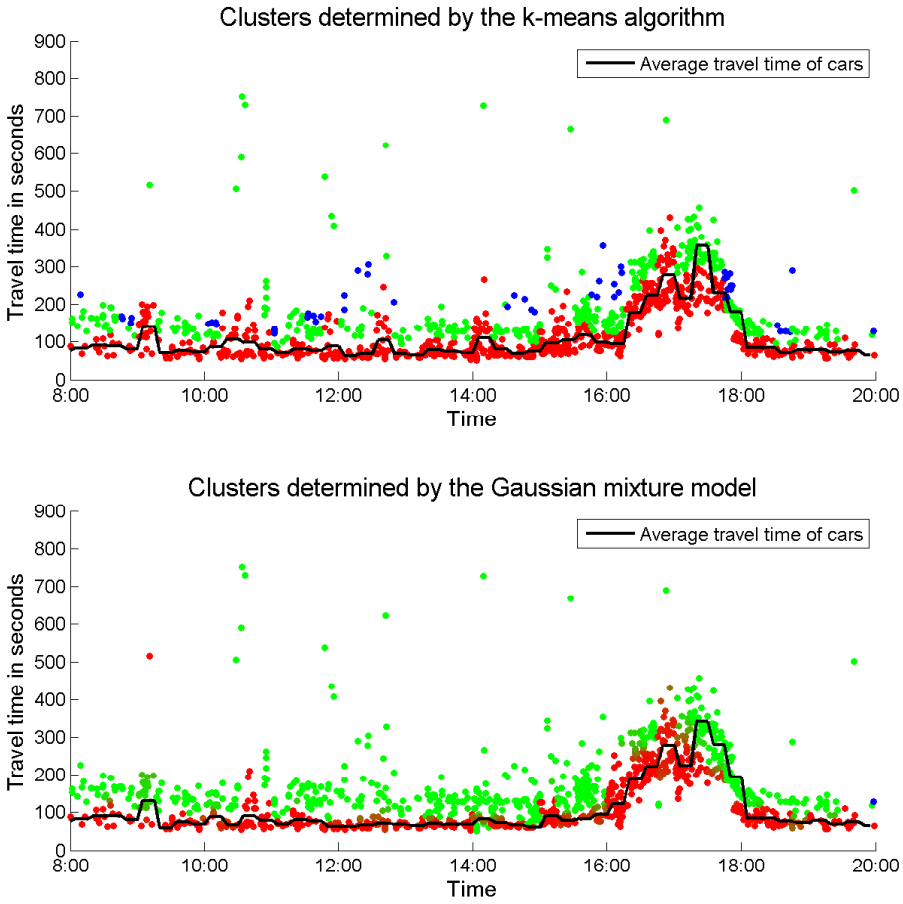
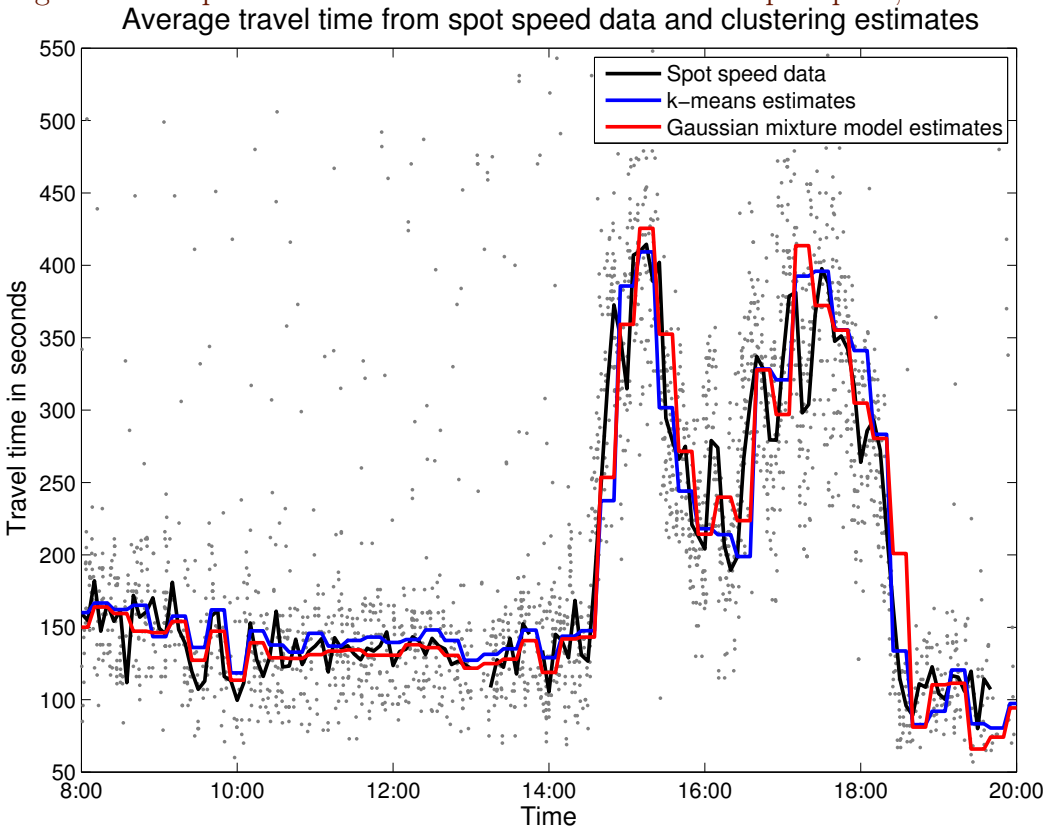


Figure 5: Comparison of the estimated travel time with spot speed, Link 1260



In terms of the number of clusters detected, the Gaussian mixture model consistently prefers two clusters of road users for Link 1260, and for Link 1262 one exception emerged at the last time interval (19:45–20:00) where three clusters were determined based on only a few observations. On the other hand, the k-means algorithm vacillates between two and three cluster solutions. During off-peak hours over Link 1260, the k-means algorithm tends to produce a larger cluster for cars compared to that from the Gaussian mixture model, resulting in slightly higher estimates of the average travel time. During peak hours, while the k-means algorithm produces distinct clusters, the fuzziness

Table 1: Proportions of road users that have the same grouping over the two links.

	Off-peak hours	Peak hours
k-means algorithm	57.96%	51.74%
Gaussian mixture model	75.12%	55.56%
Difference	17.16%***	3.82%

on cluster boundaries is noticeable as indicated by the Gaussian mixture model. Overall, both methods appear to track change points in traffic patterns quite well.

Figure 5 compares the spot speed data to the average travel time estimates obtained by the two clustering methods. The figure indicates that both methods were able to produce travel time estimates that follow the real traffic trend quite well over time. To examine whether the clustering methods are able to group an individual into the same cluster on different road segments, clusters of repeatedly observed road users on Link 1260 and 1262 are recorded and compared. In total, travel times of 647 and 288 road users were repeatedly recorded during off-peak and peak hours, respectively. Table 1 reports the proportions of road users that have the same grouping over the two links. The Gaussian mixture model has higher proportions than the k-means algorithm with the difference over off-peak period being statistically significant at 1% level of significance, as indicated by ***. This implies that the Gaussian mixture model tends to be more consistent in grouping road users, especially during off-peak hours.

5 Discussion

Data collected from the BMS provide information about travel patterns of individual road users, and travel time estimation is carried out using the BMS data. To do so, many past studies take the average or the median of collected

data after removing outliers, overlooking the existence of multiple transport modes which have distinct travel patterns. Insufficient research has been undertaken in relation to identify multiple transport modes from recorded travel times, and we filled this gap by carrying out cluster analysis of the BMS data. The Gaussian mixture model and the k-means algorithm were employed for clustering purposes, and we carried out an empirical study on the BMS travel time data collected from segments of Sandgate Road. Both clustering methods were demonstrated to be competent in discriminating between groups of travellers, producing travel time estimates that are fairly close to the real time data. In addition, the Gaussian mixture model was believed to be more consistent in terms of determining the grouping of repeatedly observed travellers over different road segments.

While the research presented in this paper provides a guideline to categorise transport modes, subsequent studies are worth carrying out to further address the issue of transport monitoring. As the TMR aims at high-quality estimation and prediction of travel time for various transport modes, methods that are capable of modelling travel time data after grouping should be explored. For instance, nonparametric smoothing techniques might be suitable to gather information about distributional properties of travel times at a specific time, whereas functional time series models may contribute to predicting traffic conditions. Furthermore, integrating temporal information with spatial statistics has the potential to improve the efficiency of monitoring road networks [12]. Research related to these methods is planned for the future.

Acknowledgements The authors thank the Queensland Department of Transport and Main Roads for providing the BMS travel time data.

References

[1] Banfield, J. and Raftery, A. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821. doi:10.2307/2532201. M101

[2] Bhaskar, A., and Chung, E. (2013) Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transport Res C-Emer*, 37, 42–72. doi:10.1016/j.trc.2013.09.013. M97

[3] Bhaskar, A., Kieu, L., Qu, M., Nantes, A., Miska, M. and Chung, E. (2014) Is bus overrepresented in Bluetooth MAC scanner data? Is MAC-ID really unique?. *Int J ITS Res*. doi:10.1007/s13177-014-0089-9. M97

[4] Bhaskar, A., Qu, M. and Chung, E. (2014) Bluetooth vehicle trajectories by fusing Bluetooth and loops: motorway travel time statistics. *IEEE T Intell Transp*, 16, 113–122. doi:10.1109/TITS.2014.2328373. M97

[5] Coretto, P. and Hennig, C. (2010) A simulation study to compare robust clustering methods based on mixtures. *Adv Data Anal Classif*, 4, 111–135. doi:10.1007/s11634-010-0065-4. M101

[6] D'Urso, P. and Maharaj, E. (2009) Autocorrelation-based fuzzy clustering of time series. *Fuzzy Set Syst*, 160, 3565–3589. doi:10.1016/j.fss.2009.04.013. M98, M103

[7] Fraley, C. and Raftery, A. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J*, 41, 578–588. M101

[8] Hennig, C. (2004) Breakdown points for maximum likelihood estimators of location-scale mixtures. *Ann Stat*, 32(4), 1313–1340. doi:10.1214/009053604000000571 M101

[9] Hennig, C. (2010) Methods for merging Gaussian mixture components. *Adv Data Anal Classif*, 4, 3–34. doi:10.1007/s11634-010-0058-3 M102

[10] Li, L., Xiqun, C., Zhiheng, L. and Lei, Z. (2013) Freeway travel-time estimation based on temporal & spatial queueing model. *IEEE T Intell Transp*, 14, 1536–1541. doi:10.1109/TITS.2013.2256132 M97

[11] Liao, T. W. (2005) Clustering of time series data - a survey. *Pattern Recogn*, 38, 1857–1874. doi:10.1016/j.patcog.2005.01.025 M99

[12] Liu, S., Anh, V., McGree, J. M., Kozan, E. and Wolff, R. C. (2015) A new approach to spatial data interpolation using higher-order statistics. *Stoch Environ Res Risk Assess*, 29, 1679–1690. doi:10.1007/s00477-014-0985-1 M112

[13] Liu, S. and Maharaj, E. (2013) A hypothesis test using bias-adjusted AR estimators for classifying time series in small samples. *Comput Stat Data An*, 60, 32–49. doi:10.1016/j.csda.2012.11.014 M98

[14] Liu, S., Maharaj, E. and Inder, B. (2014) Polarization of forecast densities: a new approach to time series classification. *Comput Stat Data An*, 70, 345–361. doi:10.1016/j.csda.2013.10.008 M98

[15] Liu, S., McGree, J. M., Ge, Z. and Xie, Y. (2015) *Computational and Statistical Methods for Analysing Big Data with Applications*. Academic Press, London. ISBN: 978-0-12-803732-4. M96, M98

[16] MacQueen, J. (1967) *Some methods for classification and analysis of multivariate observations*. Paper presented at the the 5th Berkeley Symposium on Mathematical Statistics and Probability. M100

[17] Malinovskiy, Y., Saunier, N. and Wang, Y. (2012) Analysis of pedestrian travel with static Bluetooth sensors. *Transport Res Rec*, 2299, 137–149. doi:10.3141/2299-15 M97, M98

[18] Martchouk, M., Mannering, F. and Bullock, D. (2011) Analysis of freeway travel time variability using Bluetooth detection. *J Transp Eng*, 137, 697–704. doi:10.1061/(ASCE)TE.1943-5436.0000253 M97, M98

[19] Mei, Z., Wang, D. and Chen, J. (2012) Investigation with Bluetooth sensors of bicycle travel time estimation on a short corridor. *Int J Distrib Sens N*. doi:10.1155/2012/303521. M97, M98

[20] Peel, D. and McLachlan, G. (2000) Robust mixture modelling using the t-distribution. *Stat Comput*, 10, 339–348. doi:10.1023/A:1008981510081 M101, M102

[21] Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput Appl Math*, 20, 53–65. doi:10.1016/0377-0427(87)90125-7 M101

[22] Schwarz, G. (1978) Estimating the dimension of a model. *Ann Stat*, 6, 461–464. M102

[23] Sun, L., Yang, J. and Mahmassani, H. (2008) Travel time estimation based on piecewise truncated quadratic speed trajectory. *Transport Res A-Pol*, 42, 173–186. doi:10.1016/j.tra.2007.08.004 M97

Author addresses

1. **Shen Liu**, Taylor Fry Analytics and Actuarial Consulting, Level 11, 55 Clarence Street, Sydney, New South Wales 2000, Australia.
mailto:shen.liu@taylorfry.com.au
orcid:0000-0002-7699-0106

2. **James McGree**, School of Mathematical Sciences, Queensland University of Technology, Queensland 4000, Australia.

3. **Gentry White**, School of Mathematical Sciences, Queensland University of Technology, Queensland 4000, Australia.

4. **Wayne Dale**, Department of Transport and Main Roads, Queensland Government, AUSTRALIA.