Australasian Transport Research Forum 2021 Proceedings
8-10 December, Brisbane, Australia
Publication website: http://www.atrf.info

Multi-Level Trajectory Clustering for Identifying Path Choice Set

Chintan Advani[1], Ashish Bhaskar[2], Md. Mazharul Haque[3]

[1,2,3]School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, Australia

Email for correspondence: ashish.bhaskar@qut.edu.au

1. Introduction

It is well known that the size and composition of the path choice sets play an essential role in the traffic assignment process. Prato (Bovy, 2009) provided an overview of choice set generation algorithms. However, the algorithms failed to capture more than 40% of the chosen route in most cases. Accordingly, no standard method exists in practice to identify all selected routes. Nevertheless, with advanced data collection techniques such as GPS and Bluetooth, vehicle trajectories (Advani et al., 2021) can be extracted and exploited to identify the observed choice set.

Researchers such as Ton et al. (2018), Scott et al. (2021), to name a few, have used trajectories from GPS data observed over several days to model the path choice sets for the bicycle riders empirically. For example, Ton et al. (2018) combined all the unique observed routes for a given OD pair to identify the potential path choice sets. Descriptive statistics of the observed choices are not provided, but the authors have mentioned that potential choices were not observed using the collected data. Further, comparing the observed choices with the labelling (Ben-Akiva et al., 1984) and link elimination methods(Bellman et al., 1960), they observed that these algorithms perfectly capture less than 2% of the chosen route. Similarly, (Scott et al., 2021) utilized a core link signature-based approach to identify the unique paths chosen amongst 5,561 OD pairs, based on one year aggregated GPS data. The authors observed an average of 8 unique routes per OD pair and a range of 2 to 77 for most OD pairs.

Although these studies are based on data-driven path choice set generation, the problem addressed in the present study is fundamentally different as: (a) These studies are conducted on the cyclist’s path, which possesses a much different choice structure than the car travel, and (b) In these studies, all the observed routes are considered a part of the choice set, and they do not emphasize on the path choice reduction process to identify a representative choice set (perhaps they did not observe such problem in the limited data).

Consider an example in Figure 1(b), where 415 unique paths are observed from the Bluetooth MAC scanner (BMS) based trajectory data for a Brisbane city origin-destination (OD) pair, as shown in Figure 1 (a). It is observed that the empirically identified master choice sets (all observed paths) are large, with high mutual spatial overlaps, and only the most feasible paths should be considered in the choice set. Accordingly, this paper clusters spatially similar trajectories to identify a representative set of paths from the large trajectory dataset observed over several days. The contribution of this study is the empirical identification of the path choice sets while minimizing the errors induced due to path omission and aggregation in the clustering process.

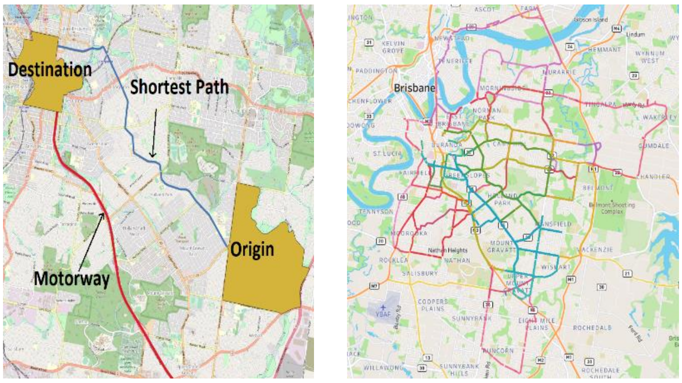


Figure 1 (a) Brisbane OD pair (origin: Mansfield, Destination: Woolloongabba) with shortest and motorway paths (b) Bluetooth data-based observed paths between the OD pairs presented in figure a.

2. Proposed Framework

We use the vehicular trajectories from the Bluetooth MAC Scanners (BMS)(Bhaskar and Chung, 2013) to test our proposed methodology. The objective here is to cluster the trajectories that satisfy the following criteria: a. The paths within the cluster should possess a high spatial similarity among themselves, and b. A clustering algorithm should be able to differentiate the paths which are dissimilar and rarely used.

In this study, a density-based DBSCAN clustering algorithm is used to obtain a representative choice set. It possesses a lower computational complexity of the order O(nlogn) and possesses the advantage of identifying the outliers in the dataset. A generalized framework of obtaining the representative path choice set by clustering the large trajectory dataset is proposed in Figure 2.

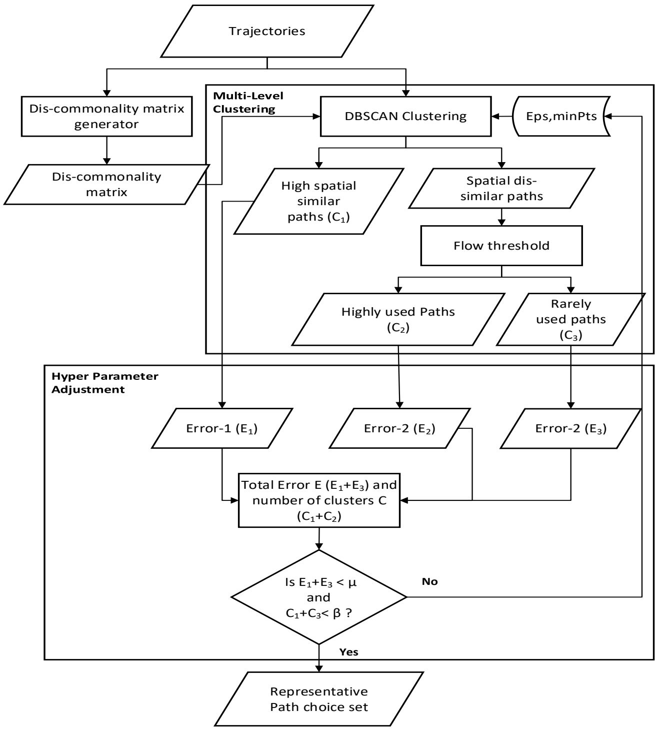

Trajectories
Dis-commonality matrix generator
Dis-commonality matrix
Multi-Level Clustering
DBSCAN Clustering
Eps,minPts
High spatial similar paths (C_1)
Spatial dis-similar paths
Flow threshold
Highly used Paths (C_2)
Rarely used paths (C_3)
Hyper Parameter Adjustment
Error-1 (E_1)
Error-2 (E_2)
Error-2 (E_3)
Total Error E (E_1+E_3) and number of clusters C (C_1+C_2)
Is E_1+E_3 < μ and C_1+C_3< β ?
No
Yes
Representative Path choice set

Figure 2 Generalized Framework to cluster the observed path choice set.

The framework consists of a multi-stage clustering process where the first stage captures spatially similar paths (C_1), and the second stage segregates the dis-similar paths into further two category namely high used paths (C_2) and rarely used paths (C_3) based on thresholds. Further, as the outputs of the clustering process are reliant on its hyper-parameters, the second part of the framework explains the adjustment of hyper-parameters by minimizing the overall choice set as well as the errors induced due to path aggregation, E_1 (from set C_1) and filtered in the clustering process E_3 (from set C_3). A brief explanation of different aspects of the framework is provided in the following section.

3. Representative path choice set identification

3.1 Distance matrix generation

Clustering algorithms require both hyperparameters and distance matrices as inputs. For trajectory clustering, distance is a measure of dissimilarity among the observed trajectories. In this study, the dis-commonality factor (DCF) is used to measure the dissimilarity among the observed trajectories. The DCF among paths is measured as:

$$DCF = 1 - \frac{\delta l_{ij}}{\sqrt{l_i l_j}} \tag{1}$$

Here, δl_{ij} is the sum of the length of the common links between path p_i and p_j, respectively and l_i, l_j are the length of the corresponding paths. The value of DCF ranges on a scale of 0 for completely overlapping paths to 1 for entirely different paths. DCF is used as the distance measure among the observed paths for the DBSCAN clustering.

3.2 First stage clustering

DBSCAN requires two input parameters, i.e., cluster search radius (*Eps*) and minimum objects (*minPts*), to cluster spatially similar paths and segregate dis-similar paths. Here, *Eps* is an upper threshold specifying the level of dissimilarity expected among a group with at least *minPts* paths in it. With the availability of the observed path choices and the corresponding DCF matrix, the DBSCAN algorithm can be initiated with a random set of hyper-parameters. Here, the paths are aggregated based on their spatial attribute. This means each unique path is represented as a single point in the space and the paths satisfying the hyper-parameter criteria are included in the same cluster. As each cluster corresponds to a representative path, the 1st stage of clustering leads to a representative path set (C_1).

3.3 Second stage filtering

The initial clustering process segregates the dis-similar paths that fail to satisfy the hyper-parameter specifications for a given set of hyper-parameters. This is prominent if the paths are isolated in the space with no/few paths in proximity. Several paths are highly preferred in the path choice context (e.g., motorways) but have less similar paths in proximity. Such paths are filtered in the initial clustering and should be included in the representative choice set. This framework adopts a trip proportion threshold-based secondary filter to identify such paths. Accordingly, if the trips on the observed path are beyond a specified threshold φ, the path is added to the representative choice set. These newly identified paths (C_2) are then added to the representative paths obtained from the first stage clustering (C_1) to obtain the overall representative choice set (C), and the paths that fail to satisfy the threshold are identified as the noise and stored in a separate database (C_3).

3.4 Quantifying the errors

Following the first stage of clustering, C_1 representative paths (clusters) will be formed. For each cluster, the non-representative paths are reduced and the trip on these paths is now assigned to the representative path of its corresponding cluster. This is because the representative paths resemble the closest similarity to its non-representative paths in the cluster. The proportion of trips aggregated from the non-representative path is the error of the cluster (E_1). We quantify these errors as the proportion of link proportion incorrectly assigned on the representative paths as expressed in equations 2 and 3.

Further, there are several paths (C_3) filtered as noise in the second stage filtering. These trips are overlooked in the representative path set and should be quantified in the error calculation. The error corresponding to these paths is the sum of trips observed on these paths, as expressed in equation 4. It should be noted that the errors for the filtered paths; C_2 is zero (i.e., E_2=0) as each path is a representative path.

$$\text{Cluster error } (e_i) = \sum_{p_i} DCF_{rp_i} . T_{p_i} \qquad (2)$$

$$\text{Total Clustering error } (E_1) = \frac{1}{T_{OD}} \sum_{i=1}^{C_1} e_i \qquad (3)$$

$$\text{Error due to Noise } (E_3) = \frac{1}{T_{OD}} \sum_{i=1}^{C_3} T_{p_i} \qquad (4)$$

3.5 Objective function for hyper-parameters adjustment

The size and quality of clusters are highly dependent on the hyperparameter inputs. We defined the objective functions that should be minimized to ensure the quality of results. The quality of clusters in this study is evaluated based on the total errors and the cluster size. The parameters *Eps, minPts* are then adjusted to satisfy the following objective function: a. A reasonable number of clusters are formed with an upper threshold of β ; b. Total errors E (E_1 + E_3) should be within the threshold μ. It can be observed that the hyperparameter adjustment depends on the acceptable error (μ) and path choice set size (β) threshold. However, an overall optimized solution can be identified by computing the outputs for various hyper-parameters.

4. Proof of concept on real network

To explain the proof of concept, this section shows the results from 8,269 trajectories for an OD pair obtained by aggregating one-year BMS data. The DCF matrix was computed for the trajectories and the hyper-parameters : *minPts* was varied from 2 to 5, and the *Eps* ranged from 0.05 to 0.3 at an increment of 0.05. As shown in Figure 6, the cluster size (C_1) tends to increase with *Eps* ranging 0.05 - 0.15 for *minPts*=3 to 5, and after that follow a decreasing trend. Thus, the noise (C_3) and the representative paths filtered(C_2) in the second stage of clustering tend to reduce proportionally as more clusters are formed in the first stage of spatial clustering. Correspondingly, the errors tend to increase as more points are accumulated in the clusters, suggesting cluster error (E_1) is independent of *minPts* for a given *Eps*.

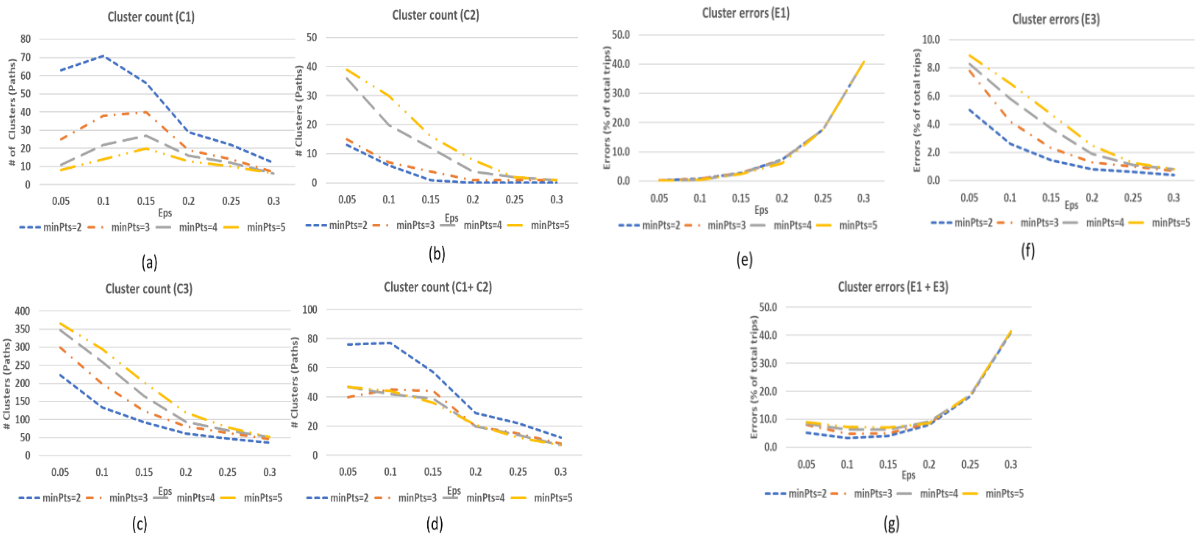


Figure 3 Effect of hyper-parameters on number (#) of clusters for (a) C1; (b) C2; (c) C3 and (d) C1 + C2 and corresponding errors for cluster (e) C1; (f) C3 and (g) C1 + C3

Figure 3 provides an independent understanding of the hyper-parameters effect on various aspects of the clustering. The errors are plotted as a function of the number of clusters to identify the optimal point and the corresponding hyperparameters in Figure 4 (a). The point closest to the origin minimizes both the variables and should be considered optimal. The optimal results are obtained for Eps=0.2 *minPts*=3, resulting in 19 clusters (C_1 = 18, C_2=1) with a total error of 8.9%. Further, the representative paths corresponding to the optimal hyper-parameters are expressed in Figure 4(b).

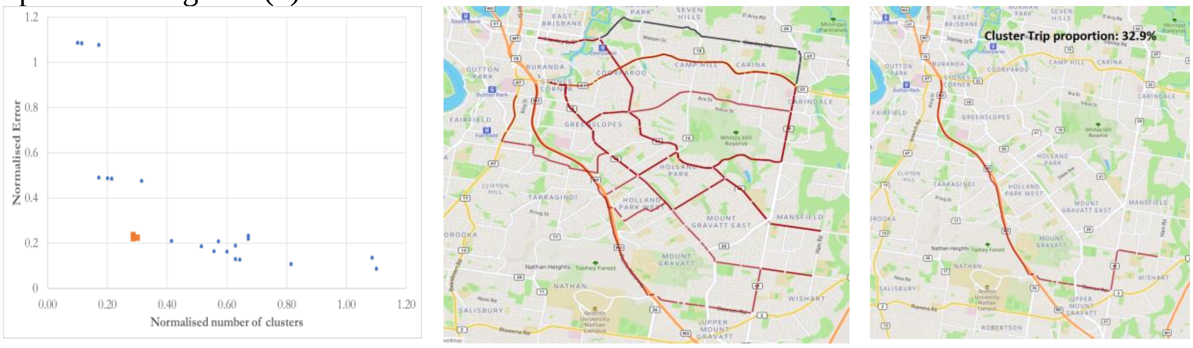
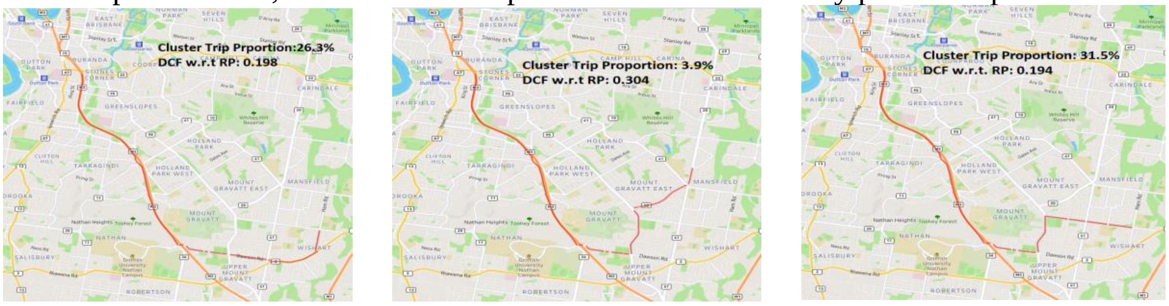


Figure 4 (a) Optimal Hyper-parameter identification (b) Representative paths visualization (c) Representative path of the selected cluster

To present the reliability of method, a cluster of seven paths is presented in Figure 5, with Figure 5 (a) being its representative path. It can be observed that the clustered paths possess high spatial similarity among each other and without any non-essential erroneous path that might affect the overall cluster error, indicating the reliability of the designed framework. In addition, the proposed framework can also be used to identify the nested structure of the observed path choices, as each cluster represents a nest of similarly perceived paths.


Cluster Trip Prportion:26.3%
DCF w.r.t RP: 0.198
Cluster Trip Proportion: 3.9%
DCF w.r.t RP: 0.304
Cluster Trip Proportion: 31.5%
DCF w.r.t. RP: 0.194

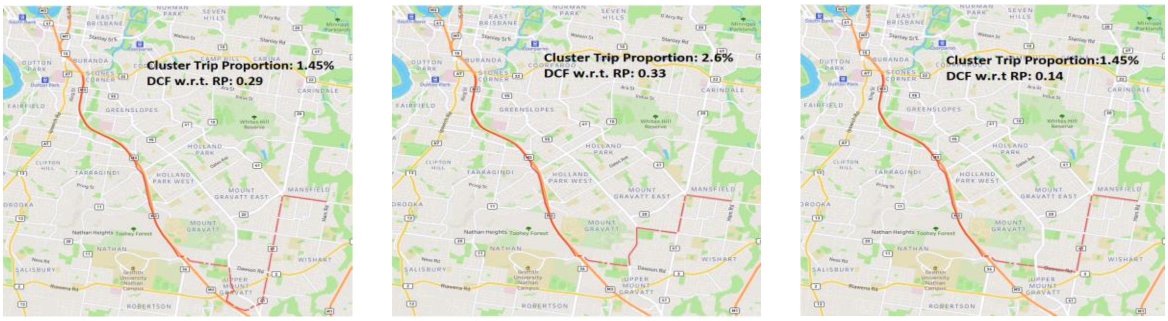

Cluster Trip Proportion: 1.45%
DCF w.r.t. RP: 0.29
Cluster Trip Proportion: 2.6%
DCF w.r.t. RP: 0.33
Cluster Trip Proportion:1.45%
DCF w.r.t RP: 0.14

Figure 5 Visualization of the paths in cluster\

6. Conclusion

Path choice set identification is an essential requirement for route choice modelling and traffic assignment. Availability of detailed vehicle trajectories from large datasets such as Bluetooth MAC Scanners (BMS) unleash opportunities for empirical identification of the path choice set for drivers on the urban road networks. However, the paths observed from trajectory datasets are huge and diverse and require a realistic representation of the observed paths. Addressing the need, this paper proposed a multi-level trajectory clustering framework for a realistic representation of the path choice set. The algorithm minimizes the errors induced due to clustering and identifies the representative path identification from each cluster. As proof of concept, the methodology is applied on 8269 trajectories to reduce 415 paths. The optimal output resulted in 19 representative paths with an error of 8.9% and corresponding hyper-parameters as *Eps* =0.2, *minPts*=3, respectively. Lastly, the visualization of the representative paths and their reduced paths indicates that the obtained results satisfy the spatial similarity and generate suitably distinct clusters, suggesting its practical applicability.

7. References

ADVANI, C., BHASKAR, A., HAQUE, M. M. & CHOLETTE, M. E. 2021. STATER: Slit-Based Trajectory Reconstruction for Dense Urban Network With Overlapping Bluetooth Scanning Zones. *IEEE Transactions on Intelligent Transportation Systems*, 1-11.

BELLMAN, R., KALABA, R. J. J. O. T. S. F. I. & MATHEMATICS, A. 1960. On k th best policies. 8, 582-588.

BEN-AKIVA, M., BERGMAN, M., DALY, A. J. & RAMASWAMY, R. Modelling inter urban route choice behaviour. Papers presented during the Ninth International Symposium on Transportation and Traffic Theory held in Delft the Netherlands, 11-13 July 1984., 1984.

BHASKAR, A. & CHUNG, E. 2013. Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies,* 37, 42-72.

BOVY, P. H. L. 2009. On Modelling Route Choice Sets in Transportation Networks: A Synthesis. *Transport Reviews,* 29, 43-68.

SCOTT, D. M., LU, W. & BROWN, M. J. 2021. Route choice of bike share users: Leveraging GPS data to derive choice sets. *Journal of Transport Geography,* 90, 102903.

TON, D., DUIVES, D., CATS, O. & HOOGENDOORN, S. 2018. Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam. *Travel Behaviour and Society,* 13, 105-117.