Please cite this paper as follows:

Behara, K. N., Bhaskar, A., & Chung, E. (2021). A DBSCAN-based framework to mine travel patterns from origin-destination matrices: Proof-of-concept on proxy static OD from Brisbane. *Transportation Research Part C: Emerging Technologies*, 131, 103370.

# A DBSCAN-based framework to mine travel patterns from origin-destination matrices: Proof-of-concept on proxy static OD from Brisbane

Krishna N.S. Behara[a], Ashish Bhaskar[a, 1] and Edward Chung[b]

[a] *School of Civil & Environmental Engineering, Faculty of Engineering, Queensland University of Technology, 2 George St GPO Box 2434 Brisbane QLD-4001, Australia*
[b] *Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong*

## Abstract

Limited studies exist in the literature on demand related travel patterns, the analysis of which requires a rich database of Origin Destination (OD) matrices with appropriate clustering algorithms. This paper develops a methodological framework to explore typical travel patterns from multi-density high dimensional matrices and estimate typical OD corresponding to those patterns. The contributions of the paper are multi-fold. First, to cluster high-dimensional OD matrices, we deploy geographical window-based structural similarity index (GSSI) as proximity measure in the DBSCAN algorithm that captures both OD structure and network related attributes. Second, to address the issue of multi-density data points, we propose clustering on individual subspaces. Third, we develop a simple two-level approach to identify optimum DBSCAN parameters. Finally, as proof-of-concept, the proposed framework is applied on proxy OD matrices from real Bluetooth data (B-OD) from Brisbane City Council region. The OD matrix clusters, typical travel patterns, and typical B-OD matrices are estimated for this study region. The analysis reveals nine typical travel patterns. The methodology was also found to perform better when GSSI was used instead of Euclidian distance as a proximity measure, and two-level DBSCAN instead of K-medoids, Spectral, and Hierarchical methods. The framework is generic and applicable for OD matrices developed from other data sources and any spatiotemporal context. DBSCAN is chosen for this study because it does not require a pre-determined number of clusters, and it identifies outliers as noise.

**Keywords** DBSCAN; typical OD matrices; typical travel patterns; Bluetooth; structural proximity; geographical window

## 1. Introduction

A pattern means the ‘repeated or regular way in which something happens’ (Dictionary, 2018). A travel pattern can be defined as repeated travel behaviour related to various features such as regularity in individual transit passenger boarding and alighting stops (Kieu et al., 2015a, b); mode selection (James, 2020); route selection (Lee and Sohn, 2015); and activity (Yildirimoglu and Kim, 2018). The focus of this paper is on demand (OD) related travel patterns, and in this paper, the term ‘travel pattern’ should be considered for the same.

[1] Corresponding author. Tel.: +61-07-3138-9985; E-mail address: ashish.bhaskar@qut.edu.au

Researchers and practitioners have long tried to understand travel patterns and answer several related research questions, which include: *What are the major travel patterns observed other than weekdays and weekends?*; *How do travel patterns during Saturdays differ from those of Sundays?*; *Are travel patterns during public holidays different from those on weekends?*; *Do school holidays during weekdays having different patterns from regular working weekdays?*; *Are there any temporal trends in travel patterns?*; *What are typical OD matrices for any study region?* To answer these questions, we need: a) A rich database of OD matrices from the same geographical region for several time periods (days); b) A suitable indicator to quantify proximity(similarity) between OD matrices; and c) An appropriate method to cluster multi-density OD matrices and estimate typical OD matrices.

OD matrix database: OD matrices are not directly obtained for large urban networks and are reverse engineered using algorithms such as bi-level optimization (Behara et al., 2020c) or other data driven methods (Behara et al., 2021; Krishnakumari et al., 2020) on the traffic monitoring datasets such as loops and Bluetooth. All signalised intersections in the Brisbane City Council (BCC) region are equipped with Bluetooth MAC Scanners. For the current research, data from these 845 BMS locations over 415 days is available (further details in section 3.1). Due to the unavailability of loop detector data for this research, proxy OD matrices are developed from Brisbane BMS data and referred to as Bluetooth-based origin destination (B-OD) matrices. For reproducibility of the research, these 415 B-OD matrices are publicly shared on https://data.researchdatafinder.qut.edu.au/dataset/bluetooth-od-data.

Indicator to quantify proximity between OD matrices: Behara et al. (2020b) have established the definition of structure in OD matrix context, where OD structure is the skeletal framework of the matrix. The corresponding demand for each OD cell is the mass on the skeleton. Behara et al. (2020b) advocated "two OD matrices have perfect structural similarity if their structures are similar with zero differences in the OD flows. Perfect structural similarity is possible only when the OD matrices are the same." Traditional measures, such as normalised root mean square error, are mere Euclidian distances and do not capture OD structure and any network related attributes. For a holistic comparison of OD matrices, indicators such as geographical window based structural similarity index (GSSI) (Behara et al., 2020a), Levenshtein distance (Behara et al., 2020b), correlation coefficient (Behara et al., 2020a; Djukic et al., 2013), and Wasserstein distance (Ruiz de Villa et al., 2014) are proposed. In this research, we consider GSSI as an indicator to quantify the proximity between OD matrices.

Method to cluster multi-density matrices and estimate typical OD: OD matrices are high dimensional data points. The structural differences among OD matrices of different travel patterns could result in multiple density regimes and the clustering algorithm to be considered for pattern mining should be robust to capture such attributes of the OD database. Clustering algorithms such as K-means (Laharotte et al., 2015), and hierarchical (Friedrich et al., 2010) have been used to cluster OD matrices. However, density-based methods such as density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) has several advantages over most other techniques but never applied for OD clustering process. The advantages of DBSCAN include: a) it does not require any predetermined number of clusters (this is important because we need to explore travel patterns that are otherwise not common or visible); and b) it can handle noise within the database. A traditional DBSCAN cannot handle multi-density data points, and there is no formal method to specify the DBSCAN parameters.

Addressing the above, the paper's objective is to develop a methodological framework to cluster multi-density OD matrices and estimate typical OD matrices for different clusters. As a proof-of-concept, a database of Bluetooth based static daily OD matrices (B-OD) from

Brisbane network is established which acts as a proxy OD for the network. It is assumed the "structure" of B-OD matrix preserves integrity of the actual distribution of travel demand over the network and can be used for travel patterns analysis of the Brisbane network. Estimating accurate OD matrices from the network is outside the scope of this work.

This study does not intend to develop a new data mining technique but tailored existing ones (subspace clustering and two-level DBSCAN) for clustering OD matrices. The major contribution of the paper is the integrated methodological framework developed to gain empirical insight into travel patterns from high dimensional multi-density matrices and estimate typical OD flows.

The remainder of the paper is structured as follows: Section 2 reviews the relevant literature on the analysis of travel patterns; Section 3 presents the methodology adopted in this study; Section 4 is the application of proposed methodology on the earlier mentioned B-OD datasets followed by comparative analysis; Section 5 is the discussion section, and finally Section 6 concludes the paper.

## 2. Literature Review

With advances in emerging technologies, and availability of big traffic data, many studies have analysed travel patterns from the perspective of individual mobility (e.g., Kieu et al. (2015b) and Ma et al. (2013) used trips from smart card, Huang et al. (2018) fused mobile phone data with smart card and taxi data, and Huang et al. (2019) used private car trajectories); travel modes (e.g., Biljecki et al. (2013) and James (2020) used GPS trajectories, and Hussain et al. (2021a) used Bluetooth and smart card); and spatial distribution of activities (e.g., Jiang et al. (2017) used mobile phone data, and Yildirimoglu and Kim (2018) combined bus GPS, smart card, and Bluetooth data). A travel pattern involves both space and time. Some studies investigated patterns across space (e.g., Louail et al. (2015) classified different cities, and Liu et al. (2015) compared suburbs within a city); time (e.g., within-the-day by Jirsa and Susilo (2016), day-to-day by Zhang et al. (2018), and weekly by Zhao et al. (2019); and both space and time (Furno et al., 2017; Laharotte et al., 2015).

Despite the abundance of literature on travel patterns, the volume of studies based on analysing travel patterns directly from OD matrices is very limited. Based on methods to analyse OD related travel patterns these studies are categorised as follows:

1. *Clustering methods*: Algorithms such as k-means (Guo et al., 2012; Liu et al., 2019), and hierarchical (Friedrich et al., 2010) were used to directly cluster OD matrices. Density based algorithms such as DBSCAN has been earlier used to cluster trajectories (Kim and Mahmassani, 2015; Tang et al., 2021); trip ends (Huang et al., 2019; Lu et al., 2015; Tang et al., 2015) and transit stops (Kieu et al., 2015b); but has not been considered to cluster high-dimensional OD matrices.
2. *Graph partitioning methods*: Guo et al. (2012) applied dynamic graph partitioning to identify the clusters of trip ends i.e., origins and destinations based on the distribution taxi trajectories; Luo et al. (2017) proposed a k-means approach to cluster OD pairs based on flows and spatial distance; and a few analysed travel patterns from OD flows using mobility graphs (as by Naveh and Kim (2018) and Zhang et al. (2018)). These methods clustered/classified OD flows and did not cluster OD matrices.
3. *Dimensionality reduction methods*: A few studies proposed dimensionality reduction methods such as principal component analysis (PCA) (Krishnakumari et al., 2020; Yang et al., 2015); singular value decomposition (SVD) (Yang et al., 2017a; Yang et

al., 2015); non-negative tensor factorization methods (Naveh and Kim, 2018); and spatial abstraction methods (Andrienko et al., 2017) to compare OD matrices. While these methods can capture most OD flow information, they might miss subtle differences in the underlying travel patterns. For instance, Steinbach et al. (2004) mentions that "dimensionality reduction approaches based on PCA or SVD may be inappropriate if different clusters lie in different subspaces". Regarding the spatial abstraction methods, discretization of flows and distances might fit different values within the same class (Andrienko et al., 2017).

The authors earlier proposed using a structural proximity measure for travel pattern identification from OD matrices (Behara et al., 2018). The focus of that conference paper was on structural proximity measure, and DBSCAN was directly applied on a small sample of B-OD matrices. That paper did not consider a) the issues related to multi-density OD matrix database; b) an approach for estimating optimum DBSCAN parameters; c) a better representation of typical OD matrices; d) a comparison of clusters resulted from using traditional and structural proximity measures; and e) comparison across different clustering methods.

To summarise, the literature on OD related travel patterns is sparse; and in the era of big data, there is a need to develop a comprehensive methodological framework that employs efficient algorithm to explore latent travel patterns by structurally comparing high-dimensional multi-density OD matrices.

# 3. Methodology

This methodology section is organised as follows. Section 3.1 presents the study location and Bluetooth data required for analysis. Section 3.2 suggests using geographical window-based structural similarity index (GSSI) as a proximity measure for comparing high-dimensional OD matrices. Section 3.3 introduces DBSCAN, and its strengths and limitations. Finally, Section 3.4 presents the proposed framework to cluster high-dimensional and multi-density OD matrices to identify typical travel patterns and typical OD matrices.

## 3.1 Study Area and Descriptive Statistics

The Brisbane City Council (BCC) region is the study area, and the B-OD matrices used for analysis were developed at SA3 (20 zones) level as shown in Figure 1. Raw Bluetooth data from 845 BMS was obtained for 415 days (June, July, August, and December months of 2015 and all months except April of 2016). We classified the day types into six categories as shown in the Figure 2. Here, SATR and SATSH are Saturdays regular and during school holidays; SUNR and SUNSH are Sundays regular and during school holidays; WDR and WDSH are weekdays regular and weekdays during school holidays; PH and LW are Public holidays and Long Weekends, respectively.

The BMS detects the MAC ID when the device is within its scanning range (~100 meters). The raw Bluetooth data includes a record number, *encrypted MAC-ID* of Bluetooth device, *scanner location ID*, *timestamp* (representing the day and time when the device is detected within the communication range of Bluetooth scanner), and *duration* (time period for which the MAC-ID was detected at the location) (Bhaskar and Chung, 2013).

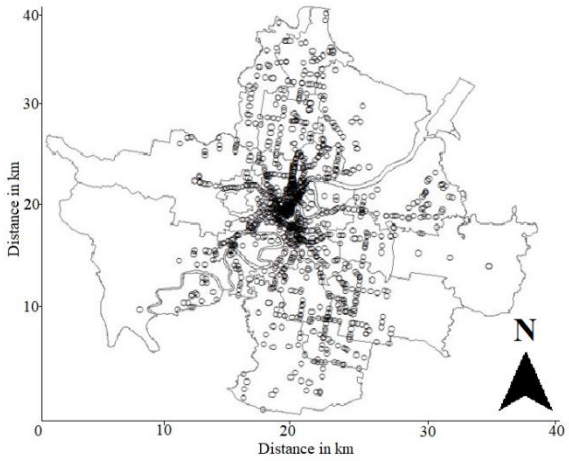


Figure 1: Location of Bluetooth scanners within SA3 zones of the BCC region

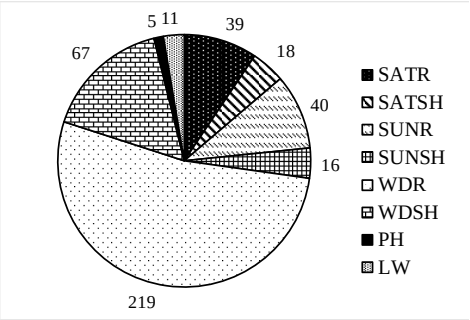


Figure 2: Classification of day types

Most Bluetooth observations are from cars equipped with Bluetooth devices (Bhaskar et al., 2015). The penetration rate of Bluetooth equipped vehicle counts in Australia was around 20% (Laharotte et al., 2014; Michau et al., 2017b). This was confirmed by Behara (2019), where Bluetooth counts with screen line observations were compared and penetration rate reported varies between 15% and 35%. Generally, the penetration rate of Bluetooth trips/trajectories is lower than that of the Bluetooth counts. However, the values are not thoroughly studied in literature primarily due to the unavailability of ground truth for large networks. For a small network, Chitturi et al. (2014) reported 4.4% average detection rate for 12 OD pairs at an interchange level. The OD flows developed from Bluetooth inferred trajectories are only a sample and do not represent actual OD flows. However, the "structure" of Bluetooth based OD can be considered as a proxy for the actual OD despite the flows are only a sample.

To develop a B-OD matrix, raw Bluetooth data from a particular day was spatially and temporally matched to define individual Bluetooth vehicle trajectories that were further split into trips (Michau et al., 2017a). Here, the Bluetooth dataset for the study date was downloaded from the BCC server and unique Device IDs were then identified. Records were retrieved individually for each Device ID and were sorted based on timestamp detections for further analysis. Within the record of each Device ID, difference in timestamps between successive detections, that is, $\delta$, was used to identify unique trips/trajectories. If successive detections were from the same scanner, then the threshold value of $\delta$ chosen to identify a new trip was 10 minutes. On the other hand, if the successive detections were from different scanners, the threshold value of $\delta$ chosen was 30 minutes to identify a new trip. The threshold values were chosen in accordance with a similar study on Brisbane Bluetooth datasets by Michau et al.

(2017b). This way, all individual trips/trajectories of each Device ID were identified and were then used to infer OD trips at a scanner level to form OD matrix of size 845 × 845, which was further transformed into B-OD matrix at SA3 levels. For this, the concordance between BMS location and the SA3 zones were considered from the BCC. The process was repeated over 415 days to generate static B-OD matrices for each day.

### 3.2 Structural Proximity Measure

The OD matrices are high-dimensional data points, and comparison of which require suitable proximity measures. Many studies in the past emphasised on the significance of structural proximity measure in clustering high dimensional data points such as documents (as by Zhang et al. (2011) and Lin et al. (2013)). In the literature, Djukic et al. (2013) proposed mean structural similarity index (MSSIM) for the structural comparison of OD matrices. MSSIM is the average of local window comparisons. Each window consists of a group of OD flows from an OD matrix. This method has a few unaddressed questions including a) what should be the window size?; and b) what is the physical meaning of this window? To address them, Behara et al. (2020a) proposed geographical window-based structural similarity index (GSSI). In the GSSI approach the OD matrix is rearranged so that the lower-level origins ($o_i$) and destinations ($d_j$) can be grouped into respective higher-level origin ($O_k$) and destination ($D_l$) zones. The higher-level boundaries define the geographical windows that capture network related attributes. For instance, the grey coloured rectangle in Figure 3 represents geographical window for the higher-level OD pair $O_1$-$D_1$ and consists of lower zonal level OD pairs – $o_1d_1$ to $o_id_j$. In Australia, the hierarchy levels in statistical area[2] can be used to define the lower and higher zonal levels.

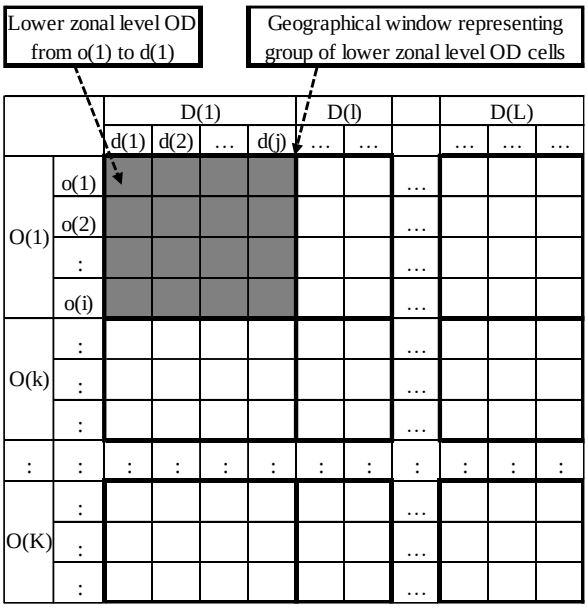


Figure 3: The concept of geographical window

The SSIM is computed on these geographical windows as shown by Equations (1) - (4) and the average of all local SSIM values is referred as GSSI (refer Equation (5))

[2] Australian Statistical Geography Standard (ASGS) defines the hierarchy of geographical areas for the release of statistical information. This includes statistical areas (SA) for four levels: SA1, SA2, SA3, and SA4 (ASGS, 2018).

$$l(\mathbf{x_w}, \mathbf{y_w}) = \frac{(2\mu_{\mathbf{x_w}}\mu_{\mathbf{y_w}} + C_1)}{(\mu_{\mathbf{x_w}}^2 + \mu_{\mathbf{y_w}}^2 + C_1)} \quad (1)$$

$$c(\mathbf{x_w}, \mathbf{y_w}) = \frac{(2\sigma_{\mathbf{x_w}}\sigma_{\mathbf{y_w}} + C_2)}{(\sigma_{\mathbf{x_w}}^2 + \sigma_{\mathbf{y_w}}^2 + C_2)} \quad (2)$$

$$s(\mathbf{x_w}, \mathbf{y_w}) = \frac{(\sigma_{\mathbf{x_w}\mathbf{y_w}} + C_3)}{(\sigma_{\mathbf{x_w}}\sigma_{\mathbf{y_w}} + C_3)} \quad (3)$$

$$SSIM(\mathbf{x_w}, \mathbf{y_w}) = [l(\mathbf{x_w}, \mathbf{y_w})^{\alpha}]\left[c(\mathbf{x_w}, \mathbf{y_w})^{\beta}\right][s(\mathbf{x_w}, \mathbf{y_w})^{\Upsilon}] \quad (4)$$

$$\alpha > 0, \beta > 0 \text{ and } \Upsilon > 0;$$

Assuming $\alpha = \beta = \Upsilon = 1$ and $C_3 = C_2/2$

$$SSIM(\mathbf{x_w}, \mathbf{y_w}) = \frac{(2\mu_{\mathbf{x_w}}\mu_{\mathbf{y_w}} + C_1)(2\sigma_{\mathbf{x_w}\mathbf{y_w}} + C_2)}{(\mu_{\mathbf{x_w}}^2 + \mu_{\mathbf{y_w}}^2 + C_1)(\sigma_{\mathbf{x_w}}^2 + \sigma_{\mathbf{y_w}}^2 + C_2)} \quad (4a)$$

$$GSSI(\mathbf{X}, \mathbf{Y}) = \frac{1}{W}\sum_{w \in W} SSIM(\mathbf{x_w}, \mathbf{y_w}) \quad (5)$$

where, **X** and **Y** represent two OD matrices to be compared; $\mathbf{x_w}$ and $\mathbf{y_w}$ represent the group of OD pairs within the w[th] local geographical windows (total W in number) in both matrices. The individual components of $l(\mathbf{x_w}, \mathbf{y_w})$, $c(\mathbf{x_w}, \mathbf{y_w})$ and $s(\mathbf{x_w}, \mathbf{y_w})$ compare the mean values ($\mu_{\mathbf{x_w}}$ and $\mu_{\mathbf{y_w}}$), the standard deviations ($\sigma_{\mathbf{x_w}}$ and $\sigma_{\mathbf{y_w}}$), and the structure (through covariance) between the group of OD pairs in both matrices. The constants $C_1, C_2$ and $C_3$ are meant to stabilize the result when either mean or standard deviation is close to zero. Generally, $C_3$ is assumed to be $C_2/2$. In the analysis conducted for this study, the OD flows within the geographical window are not all zero, hence the assumption is that both $C_1$ and $C_2$ are zero. The parameters $\alpha, \beta$ and $\Upsilon$ are used to adjust the relative importance of mean, standard deviation, and structural components respectively, and are generally assumed to be unity. The SSIM $(\mathbf{x_w}, \mathbf{y_w})$ is the structural similarity of w[th] geographical windows and $GSSI(\mathbf{X}, \mathbf{Y})$ reports the overall structural similarity of the OD matrices, **X** and **Y**. The values of both SSIM and GSSI lie between -1 and 1. Interested readers can refer Behara et al. (2020a) for more details about the robustness of GSSI technique.

### 3.3 Density based Clustering Algorithm with Noise

The DBSCAN algorithm first marks all the data points (note that data point in the current study should be read as a B-OD matrix) as 'non-visited', starting with an arbitrary selection of a 'non-visited' point and identifying all other data points within the distance threshold, ε (note that GSSI converted to dissimilarity measure is ε in this study). These data points, if any, are termed as neighbourhood points. If the number of neighbourhood points is at least MinPts (size threshold) then the data point under consideration becomes the first point of a new cluster where the neighbourhood points are part of the same cluster; otherwise, the data point is labelled as noise. In either case, the data point is now marked as 'visited'. If a cluster is identified, then the above process for defining neighbourhood points is repeated for all of the new points identified as neighbourhoods in the current cluster and the number of points in the cluster is extended. Thereafter, a new 'non–visited' point is selected, and the process is repeated until all the points are marked as 'visited'. This leads to each point either being defined as a cluster or a marked as noise.

The optimum DBSCAN parameters in the traditional approach are identified using a simple and interactive heuristic proposed by Ester et al. (1996), as follows (refer Figure 4):

***Step 1***: First, a k-dist function is defined that maps each data point, p, to the distance values (k-dist (p)) corresponding to their kth-nearest neighbour.

***Step 2***: For a given value of k, choose the kth neighbourhood of every point in the database and plot the points (x-axis) in the descending order of k-dist values (y-axis). The graph resulting from this distribution is referred to as sorted k-dist graph.

***Step 3***: The shape of the sorted k-dist graph further helps to identify the threshold point. The parameter MinPts is set to k and ε is chosen corresponding to the valley of the sorted k-dist graph. The valley point is identified through a visual observation, and as such, this technique is an interactive approach. All data points on the left side of the threshold point (i.e., higher k-dist value) are considered to be noise and the remaining points (on the right of the threshold point) are assigned to some clusters.

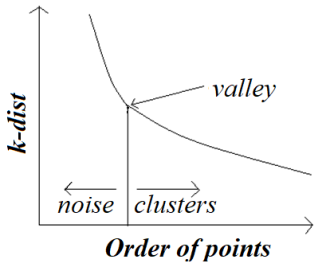


Figure 4: Typical shape of sorted k-dist graph

DBSCAN does not require a-priori specification of number of clusters, and it identifies outliers as noise. These are two major advantages of DBSCAN compared to other methods including Partition (e.g., K-means by Laharotte et al. (2015)), Spectral (Yang et al., 2017b), and Hierarchical (Liu et al., 2019). Table 1 summarised from Rodriguez et al. (2019) and IndiraPriya and Ghosh (2013) compares DBSCAN with the above-mentioned methods.

Table 1: Qualitative comparison of some clustering methods

| **Features** | **Density-based** | **Partition** | **Spectral** | **Hierarchical** |
|---|---|---|---|---|
| Does not require a-priori specification of the number of clusters | Yes | - | - | - |
| Ability to handle noise | Yes | - | - | - |
| Ability to handle arbitrarily sized and arbitrarily shaped clusters. | Yes | - | - | - |
| Ability to handle any form of similarity or distance matrix | Yes | Yes | Yes | Yes |
| Have a logical structure, and easy to read and interpret | - | - | - | Yes |
| Ability to handle varying density clusters | - | - | - | - |
| Computational complexity | Medium O(nlog(n)) (Ester et al., 1996) | Medium O(nlog(n)) (IndiraPriya and Ghosh, 2013) | High O(n3) (IndiraPriya and Ghosh, 2013) | High O(n3) (IndiraPriya and Ghosh, 2013) |

Note: n is number of data points; and O is the order of complexity.

In general, the results of clustering algorithms are sensitive to their hyper parameters. DBSCAN is sensitive to the setting of ε and MinPts and it is reported that DBSCAN does not perform well for multi-density data sets (Huang et al., 2009). Moreover, if the data points are high dimensional matrices, a relevant indicator is required to define the ε.

To address the issue with multi-density data, few researchers suggested dividing datasets based on different density levels prior to the clustering process (Elbatta and Ashour, 2013; Parsons et al., 2004). The initial clusters of these datasets are referred as subspaces, and the method is termed as subspace clustering approach. To explain different subspaces (initial clusters) in a multi-density OD database, consider one subspace that includes all daily OD matrices from weekends and public holidays, and another with all OD matrices from weekdays. The density difference between these two sub-spaces is primarily due to difference in total daily travel demand flows.

The multi-density in the high dimensional dataset can be visualised by k-dist graph (as presented by Louhichi et al. (2019), Mu et al. (2020), and Pradeep and Sowjanya (2015)) where the difference in density levels is observed from the valleys. Data points belonging to the same valley are susceptible to have approximately the same density of data within it. For instance, Figure 5 illustrates two valleys in a typical sorted k-dist plot of two-density database. Thus, the decision to consider subspace clustering should be made based on the density levels following which clustering process needs to be performed within the individual subspaces.

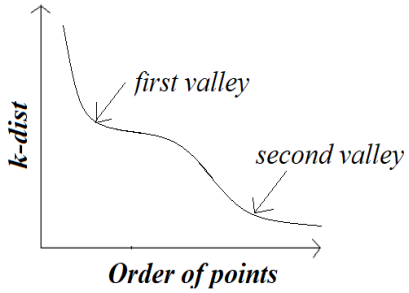


Figure 5: A sorted k-dist plot for two-density regimes

### 3.4 Proposed Framework

This section discusses the entire methodological framework to cluster multi-density high-dimensional OD matrices, identify travel patterns from the resulting clusters, and estimate typical OD matrices. The steps involved in this methodology are shown in Figure 6, and are also presented in the following:

- ***Step-1***: Prior to clustering, the OD matrix database is divided into "S" number of subspaces based on density variations and set s=1 (*see* section 3.4.1). This step addresses the issue related to multi-density database.
- ***Step-2***: Perform two-level DBSCAN clustering on $s^{th}$ subspace using a structural proximity measure (*see* section 3.4.2). This step addresses the issues related to high dimensionality of OD matrices as well as identification of DBSCAN parameters.
- ***Step-3a***: Create a homogeneous database of OD matrices for each resulting cluster of $s^{th}$ subspace. It is homogeneous because OD matrices within a cluster represent similar travel patterns.

- ***Step-3b***: If s <= S, set s=s+1, and repeat the process from Step-2 to Step-4. If all subspaces are analysed (s > S), then proceed to Step-5. This completes identification of typical travel patterns through OD matrix clusters.
- ***Step-4***: Estimate typical OD matrix database for all S subspaces. A typical OD matrix includes mean OD flow values and demand fluctuations for each OD flow (*see* section 3.4.3).

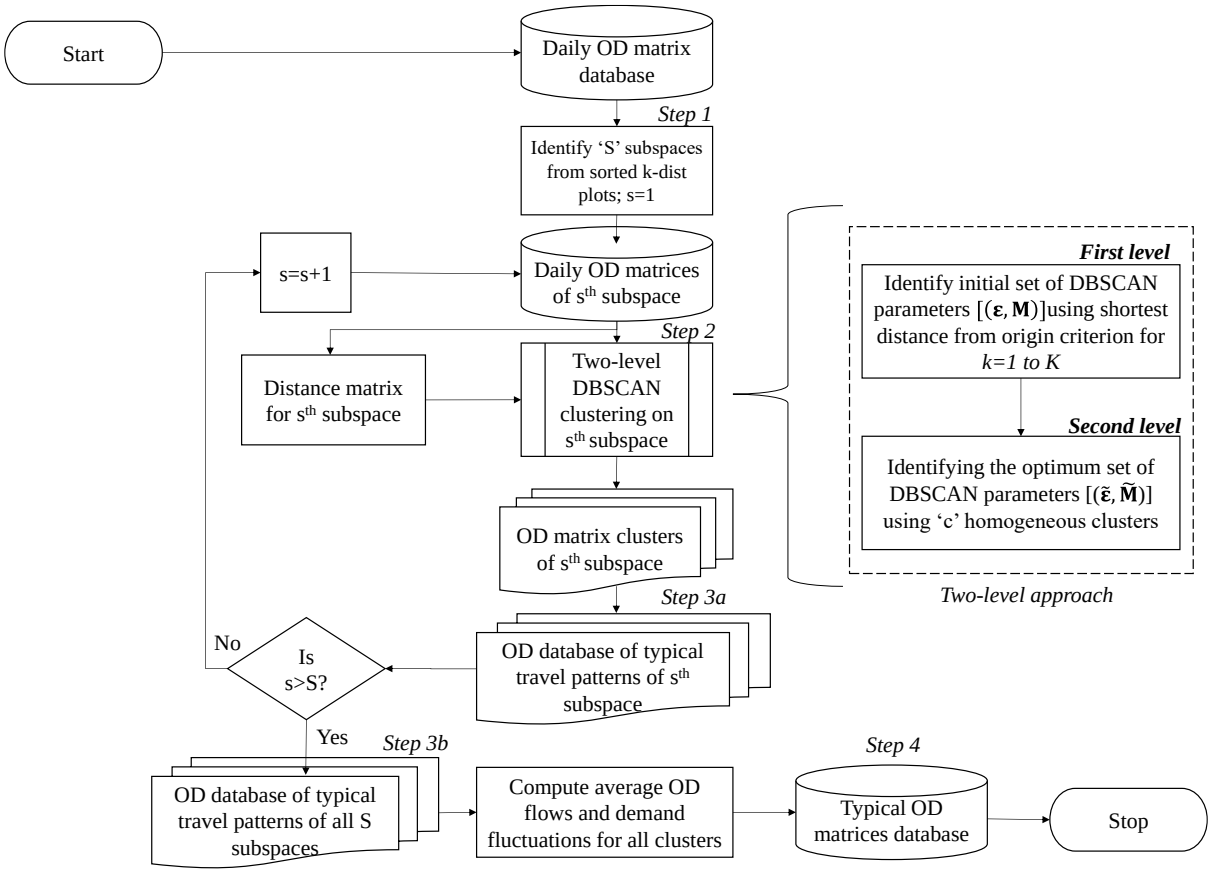


Figure 6: Proposed methodological framework

### *3.4.1 Subspaces and Distance Matrix*

Sorted k-dist plots for a range of k values provide information about the density distribution of data points. If plots show 'S' distinct valleys, then it is a S-density dataset. Thus, the data points are further split into S subspaces for subspace clustering. If the plots represent only one valley then no subspace clustering is undertaken.

The distance between each pair of OD matrices (**X** and **Y**) is computed using a structural proximity measure as shown in Equation (6). The pre-computed GSSI value is multiplied by 1000 so that the distance value is close to one decimal place.

$$d_{GSSI} = 1000 * (1 - GSSI(\mathbf{X}, \mathbf{Y})) \quad (6)$$

The distance matrix for DBSCAN algorithm comprises $d_{GSSI}$ values computed using Equation (6) for all OD matrix combinations within each subspace. Thus, if there are P number of OD matrices in the database of each subspace, the dimensions of the subspace specific distance matrix would be P x P symmetrical matrix.

### *3.4.2 Two-level DBSCAN*

Here, we discuss the proposed two-level approach to identify optimum DBSCAN parameters and then cluster multi-density OD matrix database. The final clusters identified in this step represent the typical travel patterns and are further used to estimate typical OD matrices.

- *First level:* The first level identifies the initial set of DBSCAN parameters, $[(\boldsymbol{\varepsilon}, \mathbf{M})]$. A visual inspection to identify ε is tedious task. Instead, we propose shortest distance from origin criterion to arrive at the optimal ε. According to this criterion, the valley of a sorted k-dist graph corresponds to the shortest distance from the origin of axes formed by k-dist values in the y-axis, and sorted data points (OD matrices) in the x-axis. We apply this criterion on K sorted k-dist plots so that we have the initial set of DBSCAN parameters represented by $[(\boldsymbol{\varepsilon}, \mathbf{M})] = [(\varepsilon_1, 1), \ldots (\varepsilon_k, k) \ldots . (\varepsilon_K, K)]$ where $\varepsilon_k$ is the distance threshold for MinPts = k as shown in Figure 7.

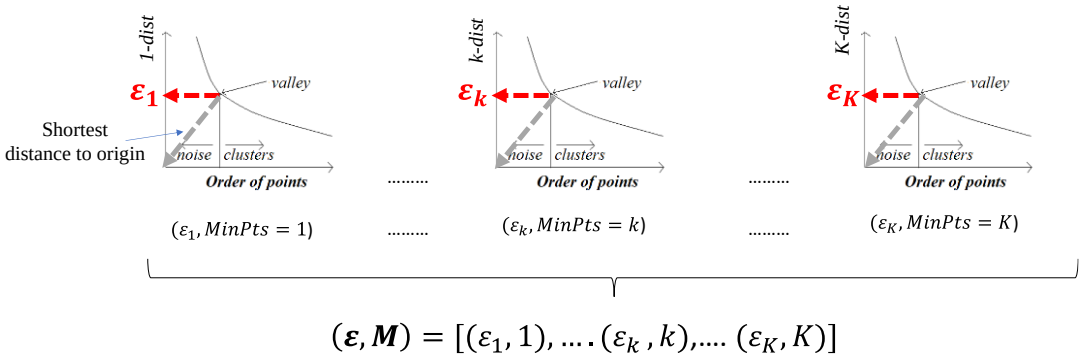


Figure 7: Initial set of parameters from the first level DBSCAN

- *Second level:* The second level identifies optimum set of DBSCAN parameters represented by $[(\tilde{\boldsymbol{\varepsilon}}, \widetilde{\mathbf{M}})]$, and the final clusters of OD matrices. We plot the number of clusters formed against the MinPts. For instance, refer to Figure 8 where x-axis is MinPts, and y-axis is number of clusters. The values of MinPts (and the corresponding ε) for which the number of clusters is less sensitive are selected. This means that there is no significant change in the data points that form clusters, and the proportional share of all such clusters remain almost the same. Generally, this is indicated from the longest plateau section of the plot. These DBSCAN parameters are identified as $[(\tilde{\boldsymbol{\varepsilon}}, \widetilde{\mathbf{M}})]$. For instance, in Figure 8 the values from 2 to 4 has the same number (that is, 5) of clusters, and we define the optimal set as $[(\tilde{\boldsymbol{\varepsilon}}, \widetilde{\mathbf{M}})] = [(\varepsilon_2, 2), (\varepsilon_3, 3), (\varepsilon_4, 4)]$. Since the clusters for $[(\tilde{\boldsymbol{\varepsilon}}, \widetilde{\mathbf{M}})]$ are same they are the final clusters that are representative of typical travel patterns.

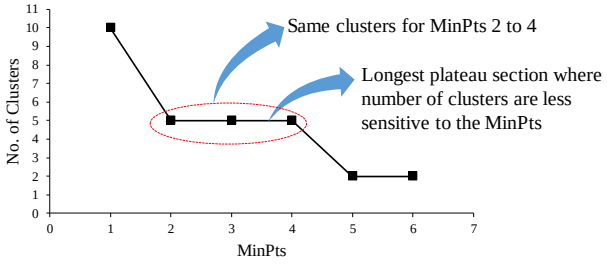


Figure 8: Number of clusters vs MinPts

### *3.4.3 Estimation of Typical OD Matrices*

The typical OD matrix for each cluster is represented by typical OD flows. A typical flow of an OD pair is the sum of mean demand and its corresponding demand fluctuations within the cluster. The mean OD matrix is obtained by computing the average value of OD matrices belonging to the same cluster. While OD matrices within a homogeneous cluster represent similar travel patterns, subtle variations in OD flows do exist. An average OD demand is a deterministic value and may not be a good representation of real-world traffic conditions. The knowledge of this demand fluctuations is important in traffic demand modelling and simulation (Wen et al., 2018). Thus, we have considered both mean flows and demand fluctuations in the estimation of typical OD matrix ($\tilde{\mathbf{X}}_c$) for $c^{th}$ typical travel pattern as shown in Equation (7)

$$\tilde{\mathbf{X}}_c = \boldsymbol{\mu}_{\mathbf{X}_c} \pm \boldsymbol{\sigma}_{\mathbf{X}_c} \quad (7)$$

where $\boldsymbol{\mu}_{\mathbf{X}_c}$ and $\boldsymbol{\sigma}_{\mathbf{X}_c}$ represent the mean ($\mu_{t,c}$) and standard deviations ($\sigma_{t,c}$) of $t^{th}$ OD flows within $c^{th}$ cluster. This method of estimating typical OD from the clusters is simple and easy to represent.

## 4. Analysis and Results

This section presents the results from clustering analysis of B-OD database. It is organised as follows: first, subspaces within multi-density B-OD matrices were identified in Section 4.1; second, application of two-level DBSCAN on this database revealed typical travel patterns as listed in Section 4.2; third, examples of typical OD flows and a typical B-OD matrix were presented in Section 4.3; finally, a comparative analyses demonstrating the benefits of the proposed framework were shown in Section 4.4.

### 4.1 Prior Identification of Subspaces

To identify the subspaces, sorted k-dist graphs for k=1 to k=15 were plotted. We selected an upper limit of k as 15 because for k>15 no more than two clusters were formed. The intial observations from sorted k-dist plots indicated two different density regimes in the datasets as shown in Figure 9. Thus, all data points were first divided into two different subspaces. It was observed that the first 129 points (in the order shown by x-axis in Figure 9) defined subspace-1 and belonged to Saturdays, Sundays, public holidays, and long weekends. The rest of the data points belonged to subspace-2 including regular weekdays and weekday school holidays.

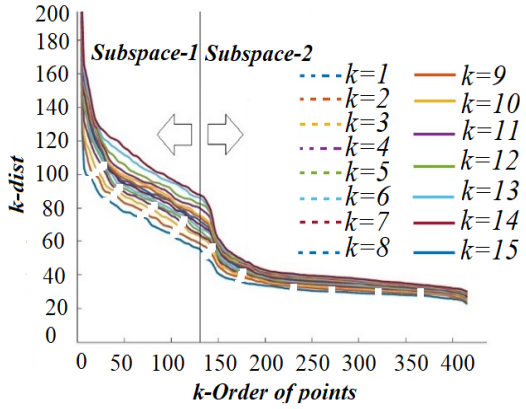


Figure 9: Identification of subspace from sorted k-dist plots

The major difference between two subspaces was the difference in total daily demand flows. This is illustrated in Figure 10 where x-axis refers to OD data points and y-axis is total daily demand flows. Since the direct application of simple DBSCAN resulted in only two clusters, we proposed to apply two-level DBSCAN on each individual sub-space as shown in the following Section 4.2. If we apply a simple DBSCAN (involves visual identification of ε from the elbow of sorted k-graph) instead of the two-level DBSCAN (auto-selection of ε) the resulting clusters will not be different. However, we claim the proposed two-level technique is a major contribution in this methodological framework because of the following reasons. First, it is an automated selection of DBSCAN parameters instead of cumbersome visual extraction. Second, the selection of optimum parameters has a practical relevance associated with it and is based on sensitivity towards the number of clusters; that is, there should be no significant change in the data points that form clusters, and the proportional share of all such clusters remain almost the same.

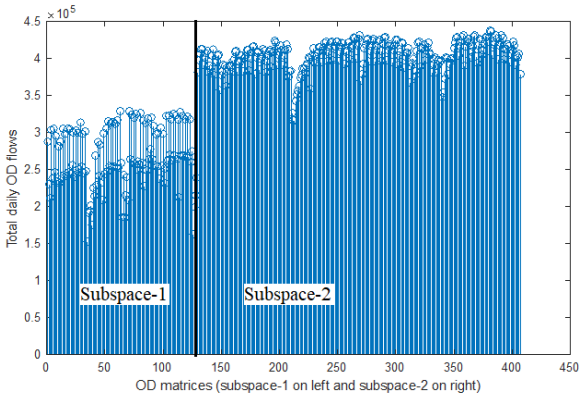


Figure 10: Difference between both subspaces in terms of total daily demand

## 4.2 Typical Travel Patterns

The typical daily travel patterns from each of the subspaces are discussed in this section.

*Travel patterns from subspace-1:* Here, the analysis was performed on 129 data points of subspace-1. The initial set of DBSCAN parameters; that is, $[(\boldsymbol{\varepsilon}, \mathbf{M})]$ were identified based on the *shortest distance from origin* criterion. Figure 11 presents the number of clusters formed for different MinPts. The number of final clusters is 5 (as represented in the pie-chart) and is least sensitive to MinPts = 4 to MinPts = 9 (refer to the longest plateau region in Figure 11). Figure 12 illustrates the proportional share of the clusters in subspace-1 for different MinPts. The percentage of noise was nearly 14% for subspace-1.

The clusters of subspace-1 that correspond to unique travel patterns are:

- Cluster-1 (C1) included Sundays and Saturdays – both regular and during school holidays periods from 2015 and 2016. Public holidays during the school holiday season such as 28th Dec 2015 (Boxing Day), 26th Jan 2016 (Australia Day), 26th March 2016 (The day after Good Friday), 28th March 2016 (Easter Monday), and 2nd May 2016 (Labour Day). It constitutes 41% of total subspace-1.
- Cluster-2 (C2) included Sundays of 2016 and constituted 13% of subspace-1.
- Cluster-3 (C3) included Saturdays of 2016 and constituted 11% of subspace-1.
- Cluster-4 (C4) included Sundays of 2015 and constituted 11% of subspace-1.

- Cluster-5 (C5) included Saturdays of 2015 and constituted 10% of subspace-1.

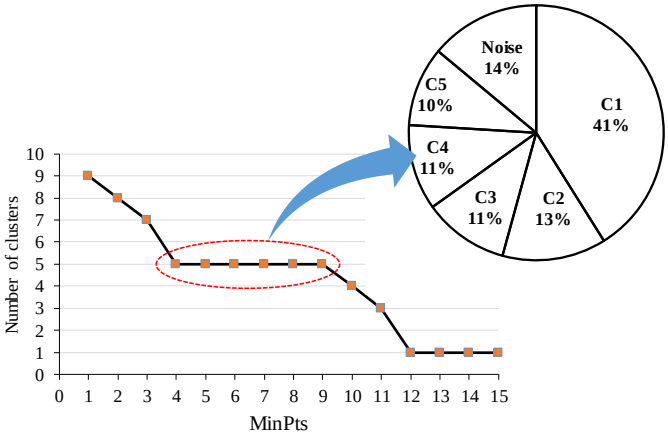


Figure 11: Number of clusters vs MinPts for subspace-1

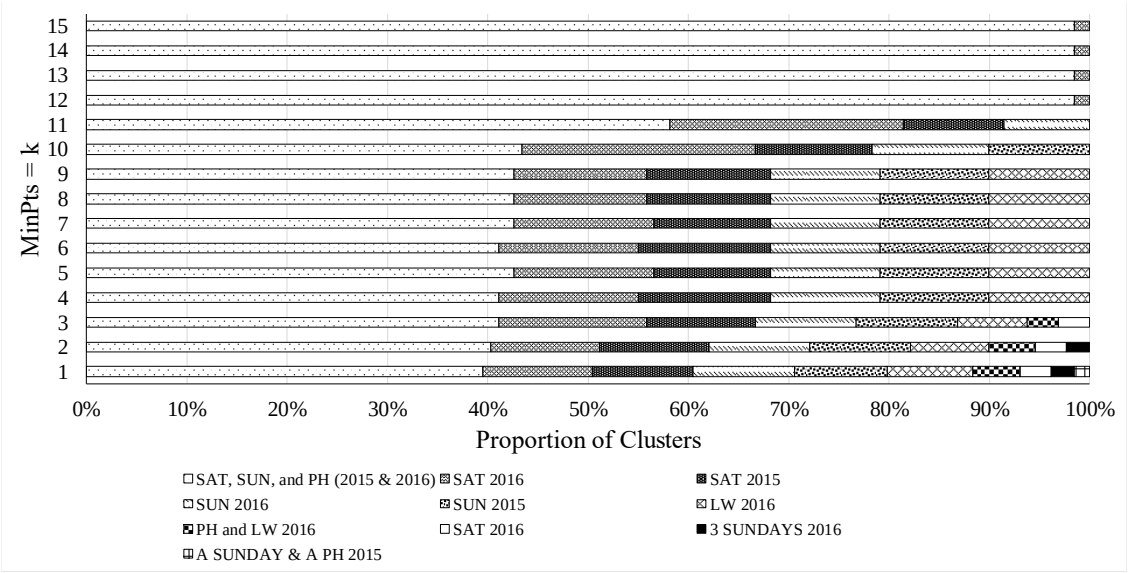


Figure 12: Proportion of subspace-1 clusters for different selection of MinPts

Figure 12 clearly illustrated that the proportional share is almost the same for MinPts ranging from 4 to 9, and for these values the clusters are less sensitive to MinPts.

*Travel patterns from subspace-2:* Similar to the last analysis, the graph presented in Figure 13 indicates the number of clusters formed for different MinPts. The number of final clusters is 4 (as represented in the pie-chart) and is least sensitive to MinPts = 3 to MinPts = 12 (refer to the longest plateau region in Figure 13). Figure 14 illustrates the proportional share of the clusters in subspace-2 for different MinPts. The percentage of noise was nearly 5% for subspace-2.

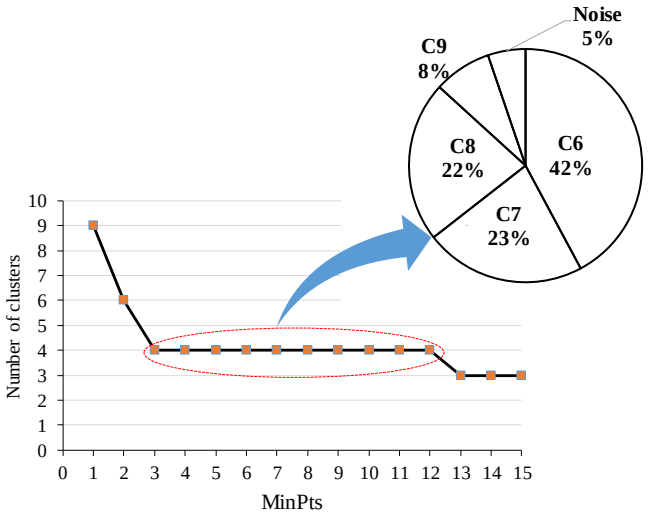


Figure 13: Number of clusters vs MinPts for subspace-2


Figure 14: Proportion of subspace-2 clusters for different selection of MinPts

Similarly, it is evident from Figure 14 that the clusters are less sensitive for MinPts ranging from 3 to 12.

The following are the observed clusters that corresponded to unique travel patterns:

- Cluster-6 (C6) included WDR of 2016 except summer and constituted 44% of subspace-2
- Cluster-7 (C7) included WDR, 2015 and constituted 24% of subspace-2.
- Cluster-8 (C8) included WDSH, 2015 and 2016 and constituted 24% of subspace-2.
- Cluster-9 (C9) included WDR of November 2016 and constituted 8% of subspace-2.

## 4.3 Typical B-OD matrices

To demonstrate typical flows (that is, combination of mean flows and the demand fluctuations) for an OD pair for nine typical patterns, refer to the box plot shown in Figure 15 for the OD pair- Mt. Gravatt and Brisbane CBD. The x-axis and y-axis of the box plot shows typical clusters and OD flows, respectively.

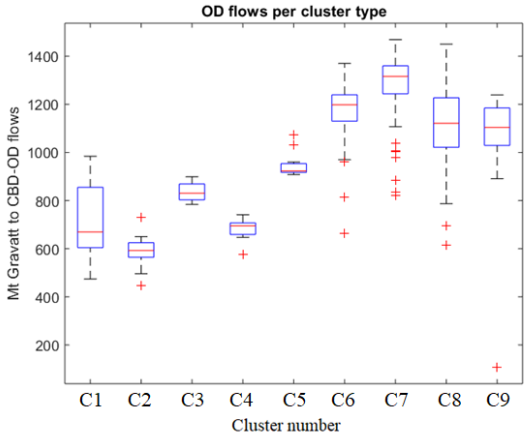


Figure 15: Typical OD flows between Mt. Gravatt and Brisbane CBD.

The inferences made from Figure 15 are:

- The magnitude of travel demand was less for clusters of subspace-1 (C1 to C5) compared to that of subspace-2 (C6 to C9). This was because the number of trips during non-working days (subspace-1) were generally less compared to trips during working days (subspace-2).
- From subspace-1 we observed that the mean OD demand was higher during Saturdays (C3 and C5) compared to Public holidays (C1) and Sundays (C2 and C4).
- From subspace-2 we observed that the mean OD demand was higher during weekdays, 2015 (C7) compared to weekdays of 2016 and other cluster types. The mean demand was almost same during weekday school holidays (C8) and WDR, November 2016 (C9).

An example of a typical weekday B-OD matrix is presented in Figure 16. The inner level column and row headers represent zonal IDs of SA3s from the BCC region. The outer level column and row headers; that is, East, North, South, West, and Inner identify the corresponding SA4 zones. Each cell in the matrix shown in Figure 16 represent a mean ± standard deviation for flows from the same OD pair.

| Typical Weekday OD | | East | | North | | | | South | | | | | | West | | | | Inner | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *30101* | *30103* | *30201* | *30202* | *30203* | *30204* | *30301* | *30302* | *30303* | *30304* | *30305* | *30306* | *30401* | *30402* | *30403* | *30404* | *30501* | *30502* | *30503* | *30504* |
| **East** | *30101* | 7005 ± 137 | 10575 ± 226 | 35 ± 3 | 95 ± 5 | 443 ± 11 | 172 ± 8 | 872 ± 22 | 309 ± 8 | 230 ± 7 | 74 ± 4 | 72 ± 4 | 43 ± 3 | 2 ± 1 | 1 ± 0 | 67 ± 4 | 34 ± 2 | 693 ± 61 | 292 ± 8 | 505 ± 13 | 132 ± 6 |
| | *30103* | 10321 ± 220 | 84476 ± 1436 | 211 ± 9 | 492 ± 12 | 2080 ± 37 | 1295 ± 39 | 8175 ± 196 | 2503 ± 48 | 2393 ± 59 | 1000 ± 22 | 1072 ± 26 | 695 ± 17 | 52 ± 4 | 4 ± 1 | 447 ± 15 | 249 ± 7 | 4200 ± 484 | 1386 ± 53 | 2465 ± 55 | 756 ± 17 |
| **North** | *30201* | 32 ± 3 | 240 ± 9 | 23079 ± 706 | 8556 ± 146 | 2858 ± 65 | 4246 ± 90 | 103 ± 5 | 383 ± 13 | 120 ± 6 | 147 ± 9 | 101 ± 6 | 40 ± 3 | 38 ± 3 | 6 ± 1 | 473 ± 41 | 1191 ± 49 | 864 ± 28 | 86 ± 4 | 1827 ± 56 | 570 ± 25 |
| | *30202* | 90 ± 5 | 532 ± 12 | 8030 ± 156 | 54665 ± 846 | 4959 ± 103 | 5067 ± 89 | 256 ± 8 | 1166 ± 25 | 337 ± 9 | 457 ± 13 | 231 ± 8 | 136 ± 5 | 64 ± 4 | 7 ± 1 | 899 ± 46 | 2502 ± 43 | 3043 ± 48 | 257 ± 8 | 9120 ± 124 | 1520 ± 37 |
| | *30203* | 426 ± 13 | 2593 ± 62 | 2518 ± 59 | 5479 ± 115 | 59827 ± 763 | 9128 ± 199 | 887 ± 29 | 1513 ± 44 | 818 ± 20 | 921 ± 19 | 955 ± 24 | 429 ± 11 | 125 ± 6 | 14 ± 2 | 1663 ± 114 | 651 ± 17 | 4068 ± 153 | 371 ± 10 | 9924 ± 164 | 1567 ± 48 |
| | *30204* | 171 ± 6 | 2036 ± 53 | 3835 ± 70 | 5096 ± 98 | 9466 ± 154 | 82212 ± 1407 | 550 ± 22 | 549 ± 15 | 402 ± 12 | 323 ± 9 | 367 ± 16 | 197 ± 7 | 36 ± 3 | 3 ± 1 | 480 ± 31 | 546 ± 17 | 1557 ± 127 | 173 ± 6 | 3197 ± 46 | 533 ± 15 |
| **South** | *30301* | 904 ± 24 | 8585 ± 185 | 102 ± 5 | 291 ± 9 | 983 ± 26 | 612 ± 23 | 45849 ± 744 | 8482 ± 117 | 5808 ± 113 | 1510 ± 29 | 621 ± 16 | 489 ± 14 | 71 ± 3 | 8 ± 1 | 654 ± 16 | 257 ± 8 | 5389 ± 111 | 3863 ± 69 | 2288 ± 46 | 1166 ± 23 |
| | *30302* | 295 ± 9 | 2459 ± 53 | 371 ± 13 | 1103 ± 30 | 1434 ± 34 | 552 ± 16 | 8628 ± 144 | 85436 ± 1211 | 9196 ± 163 | 11193 ± 185 | 4655 ± 79 | 1545 ± 31 | 429 ± 13 | 31 ± 3 | 3381 ± 59 | 1231 ± 37 | 16375 ± 351 | 4249 ± 77 | 10097 ± 323 | 4019 ± 96 |
| | *30303* | 266 ± 8 | 2271 ± 48 | 150 ± 7 | 392 ± 10 | 908 ± 19 | 431 ± 14 | 5406 ± 109 | 9840 ± 143 | 106806 ± 1492 | 7511 ± 109 | 3487 ± 64 | 8136 ± 192 | 192 ± 8 | 15 ± 2 | 1110 ± 23 | 594 ± 13 | 6102 ± 104 | 897 ± 20 | 3428 ± 81 | 1604 ± 27 |
| | *30304* | 87 ± 5 | 943 ± 20 | 224 ± 14 | 648 ± 25 | 1100 ± 27 | 413 ± 15 | 1540 ± 30 | 14089 ± 327 | 7459 ± 126 | 48582 ± 652 | 6858 ± 113 | 7557 ± 139 | 266 ± 11 | 30 ± 2 | 2236 ± 49 | 689 ± 36 | 8804 ± 317 | 1350 ± 32 | 5789 ± 197 | 1926 ± 84 |
| | *30305* | 70 ± 4 | 959 ± 22 | 108 ± 5 | 250 ± 8 | 944 ± 23 | 382 ± 16 | 653 ± 17 | 4868 ± 82 | 3369 ± 66 | 7016 ± 107 | 85267 ± 1222 | 9119 ± 155 | 624 ± 17 | 47 ± 3 | 2583 ± 44 | 313 ± 11 | 2973 ± 47 | 378 ± 10 | 1933 ± 60 | 669 ± 14 |
| | *30306* | 48 ± 3 | 649 ± 15 | 40 ± 3 | 146 ± 5 | 545 ± 11 | 220 ± 10 | 505 ± 15 | 1605 ± 27 | 7911 ± 170 | 7506 ± 124 | 9819 ± 155 | 38694 ± 738 | 97 ± 5 | 3 ± 1 | 402 ± 11 | 128 ± 5 | 2290 ± 37 | 189 ± 6 | 1397 ± 28 | 436 ± 10 |
| **West** | *30401* | 2 ± 1 | 56 ± 3 | 41 ± 4 | 66 ± 4 | 154 ± 6 | 44 ± 3 | 60 ± 4 | 449 ± 13 | 177 ± 8 | 263 ± 8 | 616 ± 17 | 92 ± 5 | 19813 ± 436 | 118 ± 6 | 1438 ± 49 | 151 ± 6 | 648 ± 14 | 60 ± 4 | 700 ± 18 | 617 ± 15 |
| | *30402* | 1 ± 0 | 4 ± 1 | 4 ± 1 | 5 ± 1 | 18 ± 2 | 3 ± 1 | 9 ± 1 | 37 ± 3 | 14 ± 2 | 33 ± 2 | 67 ± 4 | 4 ± 1 | 158 ± 7 | 2348 ± 68 | 241 ± 11 | 23 ± 2 | 87 ± 5 | 7 ± 1 | 83 ± 4 | 80 ± 4 |
| | *30403* | 59 ± 4 | 444 ± 15 | 464 ± 39 | 833 ± 43 | 1640 ± 114 | 508 ± 34 | 667 ± 18 | 3457 ± 58 | 1020 ± 26 | 1824 ± 39 | 2520 ± 37 | 393 ± 12 | 1308 ± 37 | 219 ± 11 | 59181 ± 920 | 1779 ± 102 | 6367 ± 174 | 609 ± 17 | 7803 ± 426 | 9465 ± 258 |
| | *30404* | 33 ± 3 | 280 ± 8 | 1358 ± 55 | 2651 ± 52 | 635 ± 17 | 537 ± 12 | 261 ± 7 | 1335 ± 30 | 435 ± 13 | 450 ± 14 | 332 ± 10 | 143 ± 6 | 130 ± 6 | 27 ± 3 | 1726 ± 100 | 32377 ± 498 | 3526 ± 53 | 262 ± 8 | 5526 ± 106 | 5269 ± 75 |
| **Inner** | *30501* | 680 ± 41 | 3639 ± 255 | 898 ± 25 | 3093 ± 66 | 3525 ± 66 | 1431 ± 46 | 5063 ± 104 | 17349 ± 393 | 5382 ± 118 | 5860 ± 128 | 2998 ± 66 | 2138 ± 49 | 641 ± 17 | 75 ± 4 | 6310 ± 193 | 3459 ± 71 | 196399 ± 3004 | 5167 ± 88 | 26803 ± 433 | 11282 ± 202 |
| | *30502* | 295 ± 9 | 3311 ± 89 | 98 ± 5 | 277 ± 8 | 382 ± 10 | 173 ± 7 | 3819 ± 70 | 4708 ± 62 | 918 ± 23 | 1332 ± 27 | 405 ± 11 | 217 ± 7 | 67 ± 4 | 7 ± 1 | 677 ± 16 | 274 ± 8 | 5756 ± 98 | 18732 ± 304 | 2855 ± 50 | 1209 ± 20 |
| | *30503* | 530 ± 15 | 2769 ± 74 | 2002 ± 54 | 9211 ± 152 | 9770 ± 177 | 3565 ± 63 | 2319 ± 53 | 11194 ± 329 | 3121 ± 79 | 4408 ± 105 | 2134 ± 72 | 1344 ± 28 | 654 ± 18 | 69 ± 4 | 7939 ± 435 | 5630 ± 115 | 28513 ± 381 | 2967 ± 56 | 143431 ± 1866 | 10505 ± 188 |
| | *30504* | 148 ± 6 | 787 ± 18 | 569 ± 24 | 1518 ± 36 | 1560 ± 52 | 545 ± 16 | 1168 ± 26 | 4281 ± 83 | 1485 ± 32 | 1228 ± 29 | 682 ± 14 | 424 ± 12 | 642 ± 16 | 80 ± 5 | 9518 ± 302 | 5519 ± 110 | 11643 ± 185 | 1139 ± 21 | 11129 ± 192 | 59268 ± 1283 |

Figure 16: Typical B-OD matrix for weekdays regular of 2016, expressed as $\boldsymbol{\mu_{X_c} \pm \sigma_{X_c}}$

## 4.4 Comparative Analysis

In the proposed framework we propose to use DBSCAN as a clustering algorithm, for which the parameters - epsilon (distance threshold) and MinPts – need to be known *a-priori*. We have explained how these two parameters can be optimally identified for multi-density OD database. From the past research (Behara et al., 2020a) we have proposed GSSI as an indicator to compare OD matrices. In the proposed framework, we further advocate using GSSI as the indicator for the distance measure in the clustering algorithm. To further support our methodology, we first highlighted the importance of GSSI over RMSN in Section 4.4.1; and thereafter compared our approach with other clustering algorithms in Section 4.4.2.

### *4.4.1 Structural versus Traditional Proximity Measures*

We compare the clusters resulted from two-level DBSCAN using GSSI and RMSN as proximity measures. The formulation of RMSN is same as the one used by Antoniou et al. (2004) and is expressed as shown in Equation (8)

$$\text{RMSN}\ (\mathbf{X}, \mathbf{Y}) = \frac{\sqrt{N \sum_N (X_n - Y_n)^2}}{\sum_N X_n} \tag{8}$$

where, $X_n$ and $Y_n$ are the OD flows of $n^{th}$ OD pair, and N represents number of OD pairs in both **X** and **Y**.

To maintain a fair comparison with $d_{GSSI}$ and constrain the distance values close to one decimal place, Equation (8) was multiplied with 1000 as shown in Equation (9).

$$d_{RMSN} = 1000 * \text{RMSN}\ (\mathbf{X}, \mathbf{Y}) \tag{9}$$

The clusters for subspace-1 and subspace-2 resulted from RMSN-based DBSCAN are as follows:

*Travel patterns from subspace-1:* Only one major cluster was formed as shown in Figure 17. It included all Saturdays, Sundays, Public Holidays of 2015 and 2016 except Saturdays of spring and summer, 2016 that was considered noise, and the percentage of noise was 9%.

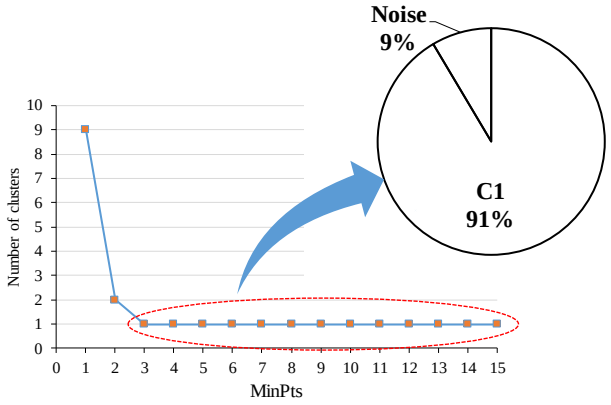


Figure 17: Number of clusters vs MinPts for subspace-1, RMSN-based experiment

*Travel patterns from subspace-2:* The graph presented in Figure 18 indicates the number of clusters formed for different MinPts. The number of final clusters is 4 (as represented in the

pie-chart) and is least sensitive to MinPts = 4 to MinPts = 13 (refer to the longest plateau region in Figure 18). The percentage of noise was nearly 9% for subspace-2.

- Cluster-1 (C1) included WDR of 2016 except summer, and constituted 37% of subspace-2
- Cluster-2 (C2) included WDR, 2015, and constituted 22% of subspace-2
- Cluster-3 (C3) included WDSH, 2015 and 2016, and constituted 21% of subspace-2
- Cluster-4 (C4) included WDR of November 2016, and constituted 11% of subspace-2

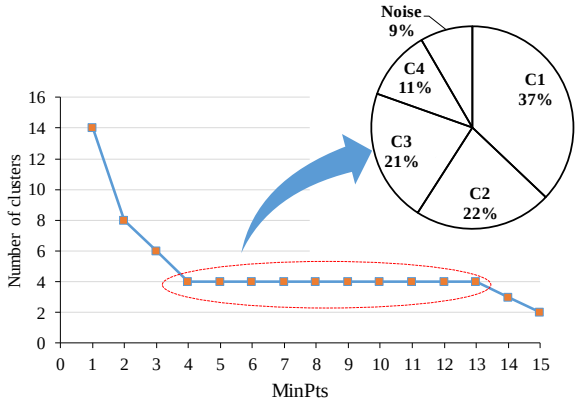


Figure 18: Number of clusters vs MinPts for subspace-2, RMSN-based experiment

Since the ground truth is unknown, one way to compare results from GSSI- and RMSN-based DBSCAN is to see how good they can reproduce pre-classified day-types described in Figure 2. The comparison in Figure 19 shows that PH (Public holidays), LW (Long weekends), and School holidays during Saturdays and Sundays could not form standalone clusters. Nonetheless, GSSI (9 clusters) could better represent the pre-classification than RMSN (5 clusters). The values in the orange-coloured boxes refer to the number of OD matrices from a particular day type that were part of a specific cluster.

### *4.4.2 Two-level DBSCAN versus other Clustering Methods*

Table 2 compares clusters from the proposed two-level DBSCAN against spectral, k-medoids, and hierarchical clustering techniques. Note that to maintain consistency same proximity measure (GSSI) was employed across all methods. Key findings from this comparison are as follows:

1. Compared to others DBSCAN is the only algorithm that can identify noise.
2. K-medoids clusters in Subspace-1(weekends, school holidays, and public holidays) are very close to DBSCAN. However, the clusters from Spectral and Hierarchical are different; and could not differentiate between Saturday and Sunday travel patterns.
3. K-medoids and hierarchical algorithms failed to identify better clusters in Subspace-2 (weekdays and weekday school holidays). On the contrary, spectral and DBCAN could identify prominent clusters in this subspace.
4. In short, some methods performed better for subspace-1 and other for subspace-2; and DBSCAN is the only method that performed better for both subspaces.

| | | | Weekdays | | | | Public Holidays | | Weekends | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Regular Weekdays | | School Holidays during weekdays | | Normal Public Holidays (PH) | Long Weekends | Saturdays | | | | Sundays | | | |
| | | | | | | | | | During School | | Regular | | During School | | Regular (SUNR) | |
| **GSSI-based experiment** | | | 2015 | 2016 | 2015 | 2016 | 2015,16 | 2015,16 | 2015 | 2016 | 2015 | 2016 | 2015 | 2016 | 2015 | 2016 |
| Subspace-1 | 1 | Weekends, PH and LW, Jan-Jun 2016 | | | | | 2 | 3 | | 8 | | 16 | | 8 | | 17 |
| | 2 | Sundays of 2016 | | | | | 1 | | | 2 | | | | 5 | | 8 |
| | 3 | Saturdays of 2016 | | | | | | | | 5 | | 9 | | | | |
| | 4 | Sundays of 2015 | | | | | 1 | | | | | | 3 | | 10 | |
| | 5 | Saturdays of 2015 | | | | | | | 3 | | 10 | | | | | |
| Subspace-2 | 6 | WDR, 2016 except summer | | 119 | | | | | | | | | | | | |
| | 7 | WDR, 2015 | 63 | 1 | | | | | | | | | | | | |
| | 8 | WDSH, 2015 and 2016 | | 3 | 22 | 40 | | | | | | | | | | |
| | 9 | WDR, November 2016 | | 23 | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | |
| **RMSN-based experiment** | | | 2015 | 2016 | 2015 | 2016 | 2015,16 | 2015,16 | 2015 | 2016 | 2015 | 2016 | 2015 | 2016 | 2015 | 2016 |
| Subspace-1 | 1 | Weekends, PH and LW, 2015 and 2016 | | | | | 5 | 11 | 6 | 10 | 11 | 18 | 6 | 10 | 11 | 29 |
| Subspace-2 | 2 | WDR, 2015 | 61 | 1 | | | | | | | | | | | | |
| | 3 | WDSH, 2015 and 2016 | | 1 | 22 | 39 | | | | | | | | | | |
| | 4 | WDR, November 2016 | | 24 | | | | | | | | | | | | |
| | 5 | WDR, 2016 except summer | | 109 | | | | | | | | | | | | |

Figure 19: Comparison of clusters resulted from the experiments based on both GSSI and RMSN (excluding noise)

Table 2: Comparison of different clustering techniques

| DBSCAN | | | Spectral | | | K-medoids | | | Hierarchical | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **C. No** | **Subspace-1** | **ODs** | **C. No** | **Subspace-1** | **ODs** | **C. No** | **Subspace-1** | **ODs** | **C. No** | **Subspace-1** | **ODs** |
| 1 | Sundays, Saturdays, and Public Holidays | 53 | 1 | Sundays, Saturdays, and Public Holidays | 32 | 1 | Sundays and Saturdays 2015 and 2016 | 36 | 1 | Sundays of 2015 and 2016 | 54 |
| 2 | Sundays of 2016 | 17 | 2 | Sunday, and Saturday School Holidays of 2016 | 28 | 2 | Sundays of 2016 | 18 | 2 | Saturdays of 2016 | 26 |
| 3 | Saturdays of 2016 | 14 | 3 | Saturdays of 2016 | 21 | 3 | Saturdays of 2016 | 14 | 3 | Saturdays of 2015 and 2016 | 27 |
| 4 | Sundays of 2015 | 14 | 4 | Saturdays and Sundays of 2016 | 9 | 4 | Sundays of 2015 | 24 | 4 | Christmas 2015 and 2016; and 1st Jan 2016 | 3 |
| 5 | Saturdays of 2015 | 13 | 5 | Sunday, and Saturday School Holidays of 2015 | 28 | 5 | Sunday, and Saturday School Holidays of 2016 | 26 | 5 | Long Weekends & Weekend School Holidays | 6 |
| | Noise | 18 | 6 | Public Holidays | 11 | 6 | Public Holidays | 11 | 6 | Public Holidays | 13 |
| Total OD matrices | | 129 | Total OD matrices | | 129 | Total OD matrices | | 129 | Total OD matrices | | 129 |

| **C. No** | **Subspace-2** | **ODs** | **C. No** | **Subspace-2** | **ODs** | **C. No** | **Subspace-2** | **ODs** | **C. No** | **Subspace-2** | **ODs** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | WDR of 2016 except summer | 121 | 7 | WDR of 2016 except summer | 107 | 7 | WDR of 2016 except summer; and WDR, 2015 | 198 | 7 | WDR of 2016 except summer; WDR, 2015; and WDSH, 2015 and 2016 | 265 |
| 7 | WDR, 2015 | 64 | 8 | WDR, 2015 | 64 | 8 | WDSH, 2015 and 2016 | 80 | 8 | WDR of Nov 2016 | 13 |
| 8 | WDSH, 2015 and 2016 | 64 | 9 | WDSH, 2015 and 2016 | 65 | 9 | Three days from Feb and Dec 2016 each | 6 | 9 | 3 WDR of Feb 2016 and 3 WDSH of Dec | 6 |
| 9 | WDR of November 2016 | 23 | 10 | WDR of Nov 2016 + WDR of 2016 except summer | 35 | 10 | 1 WDSH of Jan 2016 | 1 | 10 | 1 WDR of Nov | 1 |
| | Noise | 15 | 11 | WDR of Feb and Dec 2016; WDSH of Dec 2016 | 15 | 11 | 1 WDR of Nov 2016 | 1 | 11 | 1 WDSH of Jan 2016 | 1 |
| Total OD matrices | | 286 | Total OD matrices | | 286 | Total OD matrices | | 286 | Total OD matrices | | 286 |

# 5. Discussion

This section is divided into three sub-sections. The first sub-section provided insights into typical travel patterns for the BCC region. The second sub-section summarises the comparative analysis presented in the previous section. The third sub-section discusses the practical application of typical travel patterns in transport planning.

## 5.1 Insights into BCC Travel Patterns

Some inferences from the analysis presented in the previous section are as follows:

- Public holidays (2015 and 2016), weekdays during November (2016), and weekday school holidays (2015 and 2016) were other travel patterns besides regular weekdays and weekends.
- The GSSI-based structural proximity measure was able to differentiate weekday and weekend patterns. However, there was no typical weekend travel pattern because travel patterns during Saturday and Sunday differed from each other. This finding is in line with another study by Naveh and Kim (2018) for Brisbane, Australia, O'Fallon and Sullivan (2003) in New Zealand, and Lockwood et al. (2005) in California, United States. There are multiple underlying reasons for the differences in activity-travel patterns during these two days including:
    - The average number of trips made by a person during Sundays are less than those observed on Saturdays (O'Fallon and Sullivan, 2003)).
    - The average duration of physically active social and/or recreation activities undertaken on Saturdays are much greater than the duration of similar activities undertaken during weekdays and Sundays. This is because time-constraints during weekdays prevent individuals from undertaking long physically active recreational activities and Sundays are regarded as “days of rest” or “relaxation days” (Lockwood et al., 2005).
    - The frequency and duration of physically inactive social/recreational, meals, non-maintenance shopping, and participation in community/religious activities are greater during Sundays than Saturdays (Lockwood et al., 2005).
- Some regular and school holiday weekends were assigned to the cluster of public holidays, namely Easter holidays, Labour Day and Australia Day. This showed that some weekends were similar to public holidays in their travel patterns. However, long weekends and regular weekends had different patterns. This is mainly because trips during long weekends are usually longer than usual. Some studies found different social behaviour and poor driving patterns during long weekends often resulting in traffic congestions (Government of South Australia, 2006).
- We observed temporal trends in travel patterns. For instance, Saturdays in 2015 were found to have different travel patterns as compared to the Saturdays in 2016. A similar observation was noted for Sundays. The travel patterns of weekdays in 2015 and 2016 were also different. This is primarily because we have noticed that (car) travel demand during the year 2015 was higher than in 2016. This finding is in line with the car travel demands estimated for 2015 and 2016 by BITRE (2016).
- The travel patterns during weekday school holidays were the same during both 2015 and 2016 and were different from those of regular working weekdays. The difference was

mainly because most school trips were local and the intrazonal trips during weekday school holidays were less by approximately 80%.

- The regular weekday travel patterns during November 2016 differed from that of other regular working weekdays. This difference in travel patterns could be attributed to major events held in that month. The annual report published by Royal National Agricultural and Industrial Association of Queensland (RNA, 2016) estimated that, in 2016, Brisbane Showgrounds attracted almost a million people by hosting more than 250 events, with an increase of 20% as compared to 2015. The month of November was the busiest month of 2016 due to hosting a total of 35 events.

### 5.2 Summary of Comparative Analysis

The clusters produced from RMSN-based experiment demonstrated temporal trends in subspace-2 travel patterns. However, it resulted Saturdays, Sundays, and Public Holidays into one major cluster, and thus failed to distinguish the differences among daily travel patterns during those days. This was mainly because RMSN is based on deviations of individual OD flows and could not account for the subtle structural differences within the respective B-OD matrices.

Different clustering techniques can lead to similar results, if their hyperparameters are properly calibrated. For the current application, DBSCAN has identified most of the long weekends as noise, and holidays are part of Cluster-1 which includes Public Holidays, Saturdays, and Sundays. Whereas hierarchical clustering with 6 number of pre-specified clusters have grouped long weekends and weekend school holidays together. Should the long weekends have a distinct cluster is hard to know. It is worth mentioning that the results of hierarchical clustering (and k-medoids and spectral) will be sensitive to the number of clusters considered. However, DBSCAN has benefit over this as it does not need pre-specified number of clusters and is robust to the noise. Therefore, we propose to use DBSCAN over other clustering algorithm and in absence of any further information we are more confident with clusters from DBSCAN than other algorithms.

### 5.3 Practical Application

The knowledge of travel patterns and typical ODs identified from this approach has the following practical applications:

- Although the study demonstrated the application using static B-OD matrices, the methodology is generic and is applicable for OD matrices developed from other data sources, and for any spatiotemporal context. The proposed methodology can also be used to compare multi-modal travel patterns; for instance, comparing clusters of smartcard and Bluetooth OD matrices can help in identifying the differences in travel patterns between transit and car users, respectively.

- Right selection of typical prior OD minimises the search space of an OD estimation problem (Mo et al., 2020). Choosing a typical seed OD, traffic counts, and any other traffic information such as sub-path/partial-path flows for the same typical travel pattern reduces the scope of optimisation algorithm and improves the quality of final OD estimate.

- Understanding typical travel patterns of any city can assist in effective and rational policy developments. From our findings, we observed that some of the public holidays on Mondays (Labour Day and Easter Monday) had travel patterns similar to that of weekends (in cluster C1). Shifting public holidays towards weekends has many practical benefits

because they can form a greater number of long weekends, and this encourages public to spend more via excursions, short-stay holiday trips etc., boosting the nation's economy. A similar strategic move was implemented in Japan to improve nation's ailing economy (Chung, 2003).

- The knowledge of travel patterns can help transport planners to appropriately conduct travel surveys across the study network within a year. For instance, the Household Travel Survey (HTS) for South East Queensland (SEQTS, 2010) was conducted for over 10 weeks from mid-April through late-June and in July in 2009. The survey period avoided the days during school/university holidays. However, our study added value to those days by recognising unique travel patterns during weekday school holidays irrespective of the years. This could enable better capture of travel patterns for any large-scale study region.

## 6. Conclusion

Limited studies are available in the literature on OD related travel patterns. This is primarily due to lack of a rich database of OD matrices from the same geographical region for several time periods. The analysis also needs, a suitable method to structurally compare high dimensional OD matrices; and an appropriate method to cluster multi-density matrices and estimate typical OD matrices. To this end, the paper develops a systematic methodological framework to explore typical travel patterns from multi-density high dimensional OD matrix database and estimate typical OD matrices for large-scale networks. The practicality of the proposed framework was demonstrated with a proof-of-concept application using a proxy demand from 415 Bluetooth OD matrices from BCC region for the years 2015 and 2016. For the proposed framework: GSSI is deployed as an appropriate structural proximity measure to cluster high-dimensional OD matrices; individual subspaces are identified before clustering to address the issue of multi-density OD matrices; a simple two-level approach is developed to identify optimum DBSCAN parameters for OD matrix clustering; and estimate typical OD matrices from the resulting clusters. A comparative analysis has revealed that proposed methodology can produce meaningful clusters which a traditional measure and other clustering methods have failed to achieve.

We would like to acknowledge that different clustering algorithms if properly calibrated should provide reasonable results. The analyst must make appropriate selection of the algorithm based on its requirement. We recommend DBSCAN because it does not require prior specification of clusters, and is robust with the noise in the dataset.

The proof-of-concept application identified nine typical patterns for the BCC region. The travel patterns and typical OD matrices for Saturdays (2015), Saturdays (2016), Sundays (2015), Sundays (2016), public holidays (2015 and 2016), regular weekdays (2015), regular weekdays (2016), weekdays during November (2016), and weekday school holidays (2015 and 2016) were different.

The study can be further extended in the following ways:

- For the current proof-of-concept, static Bluetooth-based OD (B-OD) at statistical area-3 (SA3) level for the entire day was used. This is primarily because the sample size of daily B-OD is higher than the hourly-based one. Exploring more detailed OD in both temporal and spatial context will help better understand the patterns where temporal

context is for different time periods of the day (morning peak, business hours, evening peak, and off peak), and spatial is for different statistical area levels.

- Bluetooth based OD matrix from a network, highly equipped with Bluetooth sensors, such as Brisbane has the potential to provide a proxy OD matrix for longitudinal travel pattern analysis. However, if the analysis period is for larger time such as over different years, it is recommended to analyse the penetration rate of the Bluetooth over different OD pairs. Significant changes in the penetration rate over different OD pairs can significantly impact the travel pattern analysis.
- The findings of the analysis presented in the paper are limited to the availability of the datasets for 415 days (from June-Aug 2015, Dec 2015 and all months of 2016 except April 2016). To study the seasonality and weather impacts, it is suggested to apply the proposed methodology on a larger period.
- OD matrices from multiple modes can be investigated to obtain a holistic picture of network wide travel patterns. For instance, transit OD from smart card (Hussain et al., 2021b) can be compared with Bluetooth OD patterns identified in this study.

## Acknowledgements

The authors are thankful to the Brisbane City Council (BCC) for providing the Bluetooth data and the Queensland University of Technology (QUT) for supporting this research. The conclusions of this paper reflect the understandings of the authors, who are responsible for the accuracy of the findings.

## References

Andrienko, G., Andrienko, N., Fuchs, G., Wood, J., 2017. Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. *IEEE Transactions on Visualization & Computer Graphics*(1), 1-1.

Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2004. Incorporating automated vehicle identification data into origin-destination estimation, *Transportation Research Record*, pp. 37-44.

ASGS, 2018. Australian Statistical Geography Standard (ASGS). Commonwealth of Australia.

Behara, K.N., Bhaskar, A., Chung, E., 2018. Classification of typical Bluetooth OD matrices based on structural similarity of travel patterns-Case study on Brisbane city, *Transportation Research Board 97th Annual Meeting*, Washington DC, United States.

Behara, K.N., Bhaskar, A., Chung, E., 2020a. Geographical window based structural similarity index for origin-destination matrices comparison. *Journal of Intelligent Transportation Systems*, 1-22.

Behara, K.N., Bhaskar, A., Chung, E., 2020b. A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. *Transportation Research Part C: Emerging Technologies* 111, 513-530.

Behara, K.N., Bhaskar, A., Chung, E., 2020c. A novel methodology to assimilate sub-path flows in bi-level OD matrix estimation process. *IEEE Transactions on Intelligent Transportation Systems*, 1-11.

Behara, K.N., Bhaskar, A., Chung, E., 2021. Single-level approach to estimate origin-destination matrix: exploiting turning proportions and partial OD flows. *Transportation Letters*, 1-12.

Behara, K.N.S., 2019. Origin-destination matrix estimation using big traffic data: A structural perspective. PhD Thesis, Queensland University of Technology, Brisbane.

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.
Bhaskar, A., Qu, M., Nantes, A., Miska, M., Chung, E., 2015. Is bus overrepresented in Bluetooth MAC scanner data? Is MAC-ID really unique? *International Journal of Intelligent Transportation Systems Research* 13(2), 119-130.
Biljecki, F., Ledoux, H., Van Oosterom, P., 2013. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science* 27(2), 385-407.
BITRE, 2016. 2016 SoE Built environment baseline modal share projections for total urban travel to 2030 (BITRE), State of the Environment, Australia.
Chung, E., 2003. Classification of traffic pattern, *Proc. of the 11th World Congress on ITS*, pp. 687-694.
Dictionary, 2018. *Cambridge online dictionary*, Cambridge, UK.
Djukic, T., Hoogendoorn, S., Van Lint, H., 2013. Reliability assessment of dynamic OD estimation methods based on structural similarity index, *Transportation Research Board 92nd Annual Meeting*, Washington DC.
Elbatta, M.T., Ashour, W.M., 2013. A dynamic method for discovering density varied clusters. *Int. Journal of Signal Processing, Image Processing, and Pattern Recognition* 6(1), 123-134.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, *Kdd*, pp. 226-231.
Friedrich, M., Immisch, K., Jehlicka, P., Otterstätter, T., Schlaich, J., 2010. Generating origin-destination matrices from mobile phone trajectories. *Transportation Research Record: Journal of the Transportation Research Board*(2196), 93-101.
Furno, A., Fiore, M., Stanica, R., 2017. Joint spatial and temporal classification of mobile traffic demands, *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, Atlanta, United States, pp. 1-9.
Government of South Australia, 2006. Long weekend 2005-06 archive. Department of Infrastructure and Transport, South Australia.
Guo, D., Zhu, X., Jin, H., Gao, P., Andris, C., 2012. Discovering spatial patterns in origin-destination mobility data. *Transactions in GIS* 16(3), 411-429.
Huang, T.-q., Yu, Y.-q., Li, K., Zeng, W.-f., 2009. Reckon the parameter of DBSCAN for multi-density data sets with constraints, *Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on*. IEEE, pp. 375-379.
Huang, Y., Xiao, Z., Wang, D., Jiang, H., Wu, D., 2019. Exploring individual travel patterns across private car trajectory data. *IEEE Transactions on Intelligent Transportation Systems* 21(12), 5036-5050.
Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., Wang, F.-Y., 2018. Modeling real-time human mobility based on mobile phone and transportation data fusion. *Transportation Research Part C: Emerging Technologies* 96, 251-269.
Hussain, E., Behara, K.N., Bhaskar, A., Chung, E., 2021a. A Framework for the Comparative Analysis of Multi-Modal Travel Demand: Case Study on Brisbane Network. *IEEE Transactions on Intelligent Transportation Systems*.
Hussain, E., Bhaskar, A., Chung, E., 2021b. Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research Part C: Emerging Technologies* 125.
IndiraPriya, P., Ghosh, D., 2013. A survey on different clustering algorithms in data mining technique. *International Journal of Modern Engineering Research (IJMER)* 3(1), 267-274.

James, J., 2020. Semi-supervised deep ensemble learning for travel mode identification. *Transportation Research Part C: Emerging Technologies* 112, 120-135.
Jiang, S., Ferreira, J., González, M.C., 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* 3(2), 208-219.
Jirsa, V., Susilo, Y.O., 2016. Estimating the hourly variability of bicycle trip patterns and characteristics from automatic bicycle counters: Case study in Prague, Czech Republic, *ICTTE 2016: proceedings of the 3rd International Conference on Traffic and Transport Engineering*. City Net Scientific Research Center.
Kieu, L.-M., Bhaskar, A., Chung, E., 2015a. A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. *Transportation Research Part C: Emerging Technologies* 58, 193-207.
Kieu, L.-M., Bhaskar, A., Chung, E., 2015b. Passenger segmentation using smart card data. *IEEE Transactions on Intelligent Transportation Systems* 16(3), 1537-1548.
Kim, J., Mahmassani, H.S., 2015. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Procedia* 9, 164-184.
Krishnakumari, P., van Lint, H., Djukic, T., Cats, O., 2020. A data driven method for OD matrix estimation. *Transportation Research Part C: Emerging Technologies* 113, 38-56.
Laharotte, P.-A., Billot, R., Come, E., Oukhellou, L., Nantes, A., El Faouzi, N.-E., 2014. Spatiotemporal analysis of bluetooth data: Application to a large urban network. *IEEE Transactions on Intelligent Transportation Systems* 16(3), 1439-1448.
Laharotte, P.-A., Billot, R., Come, E., Oukhellou, L., Nantes, A., El Faouzi, N.-E., 2015. Spatiotemporal analysis of Bluetooth data: Application to a large urban network. *IEEE Transactions on Intelligent Transportation Systems* 16(3), 1439-1448.
Lee, M., Sohn, K., 2015. Inferring the route-use patterns of metro passengers based only on travel-time data within a Bayesian framework using a reversible-jump Markov chain Monte Carlo (MCMC) simulation. *Transportation Research Part B: Methodological* 81, 1-17.
Lin, Y.-S., Jiang, J.-Y., Lee, S.-J., 2013. A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering* 26(7), 1575-1590.
Liu, T., Krishnakumari, P., Cats, O., 2019. Exploring demand patterns of a ride-sourcing service using spatial and temporal clustering, *2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, pp. 1-9.
Liu, X., Gong, L., Gong, Y., Liu, Y., 2015. Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography* 43, 78-90.
Lockwood, A.M., Srinivasan, S., Bhat, C.R., 2005. Exploratory analysis of weekend activity patterns in the San Francisco Bay Area, California. *Transportation Research Record* 1926(1), 70-78.
Louail, T., Lenormand, M., Picornell, M., Cantú, O.G., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2015. Uncovering the spatial structure of mobility networks. *Nature Communications* 6, 6007.
Louhichi, S., Gzara, M., Ben-Abdallah, H., 2019. MDCUT 2: a multi-density clustering algorithm with automatic detection of density variation in data with noise. *Distributed and Parallel Databases* 37(1), 73-99.
Lu, M., Wang, Z., Liang, J., Yuan, X., 2015. OD-Wheel: Visual design to explore OD patterns of a central region, *2015 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, pp. 87-91.
Luo, D., Cats, O., van Lint, H., 2017. Constructing transit origin–destination matrices with spatial clustering. *Transportation Research Record* 2652(1), 39-49.

Michau, G., Nantes, A., Bhaskar, A., Chung, E., Abry, P., Borgnat, P., 2017a. Bluetooth data in an urban context: Retrieving vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems* 18(9), 2377-2386.
Michau, G., Pustelnik, N., Borgnat, P., Abry, P., Nantes, A., Bhaskar, A., Chung, E., 2017b. A primal-dual algorithm for link dependent origin destination matrix estimation. *IEEE Transactions on Signal and Information Processing over Networks* 3(1), 104-113.
Mo, B., Li, R., Dai, J., 2020. Estimating dynamic origin–destination demand: A hybrid framework using license plate recognition data. *Computer-Aided Civil and Infrastructure Engineering* 35(7), 734-752.
Mu, B., Dai, M., Yuan, S., 2020. DBSCAN-KNN-GA: a multi Density-Level Parameter-Free clustering algorithm, *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, p. 012023.
Naveh, K.S., Kim, J., 2018. Urban Trajectory Analytics: Day-of-Week Movement Pattern Mining Using Tensor Factorization. *IEEE Transactions on Intelligent Transportation Systems*.
O'Fallon, C., Sullivan, C., 2003. Understanding and managing weekend traffic congestion, *at 26th ATRF Conference*.
Parsons, L., Haque, E., Liu, H., 2004. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter* 6(1), 90-105.
Pradeep, L., Sowjanya, A., 2015. Multi-density based incremental clustering. *International Journal of Computer Applications* 116(17).
RNA, 2016. The Royal National Agricultural and Industrial Association (RNA) of Queensland Annual Report, Albion, Queensland Australia.
Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.d.F., Rodrigues, F.A., 2019. Clustering algorithms: A comparative approach. *PloS one* 14(1), e0210236.
Ruiz de Villa, A., Casas, J., Breen, M., 2014. OD matrix structural similarity: Wasserstein metric, *Transportation Research Board 93rd Annual Meeting*.
SEQTS, 2010. South-East Queensland Travel Survey 2009, *Queensland Transport and Main Roads*, Brisbane.
Steinbach, M., Ertöz, L., Kumar, V., 2004. The challenges of clustering high dimensional data, *New Directions in Statistical Physics*. Springer, pp. 273-309.
Tang, J., Bi, W., Liu, F., Zhang, W., 2021. Exploring urban travel patterns using density-based clustering with multi-attributes from large-scaled vehicle trajectories. *Physica A: Statistical Mechanics and its Applications* 561, 125301.
Tang, J., Liu, F., Wang, Y., Wang, H., 2015. Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and its Applications* 438, 140-153.
Wen, T., Gardner, L., Dixit, V., Waller, S.T., Cai, C., Chen, F., 2018. Two methods to calibrate the total travel demand and variability for a regional traffic network. *Computer-Aided Civil and Infrastructure Engineering* 33(4), 282-299.
Yang, C., Yan, F., Xu, X., 2017a. Daily metro origin-destination pattern recognition using dimensionality reduction and clustering methods, *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*. IEEE, pp. 548-553.
Yang, C., Yan, F.F., Xu, X.D., 2015. Clustering Daily Metro Origin-Destination Matrix in Shenzhen China, *Applied Mechanics and Materials*. Trans Tech Publ, pp. 422-432.
Yang, S., Wu, J., Qi, G., Tian, K., 2017b. Analysis of traffic state variation patterns for urban road network based on spectral clustering. *Advances in Mechanical Engineering* 9(9), 1687814017723790.

Yildirimoglu, M., Kim, J., 2018. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. *Transportation Research Part C: Emerging Technologies* 89, 254-267.
Zhang, A., Kang, J.E., Axhausen, K., Kwon, C., 2018. Multi-day activity-travel pattern sampling based on single-day data. *Transportation Research Part C: Emerging Technologies* 89, 96-112.
Zhang, T., Tang, Y.Y., Fang, B., Xiang, Y., 2011. Document clustering in correlation similarity measure space. *IEEE Transactions on Knowledge and Data Engineering* 24(6), 1002-1013.
Zhao, Y., Zhu, X., Guo, W., She, B., Yue, H., Li, M., 2019. Exploring the Weekly Travel Patterns of Private Vehicles Using Automatic Vehicle Identification Data: A Case Study of Wuhan, China. *Sustainability* 11(21), 6152.