Paper submitted for publication and presentation at:
94th Annual Meeting of the Transportation Research Board

Using Farecard Data to Identify and Exploit Transit Passenger Flows in a Spatial-temporal Context

Siddharth Gupta, sid1.gupta1@gmail.com
Undergraduate student, Infrastructural Civil Engineering, Indian Institute of Technology- Madras, India
Occupational Trainee, University of Queensland, Australia

Mark Hickman, m.hickman1@uq.edu.au
Professor and ASTRA Chair of Transport Engineering, University of Queensland, Australia

Submission Date: August 1, 2014
Word Count: 5238 words + 9 Figures + 1 Table = 7738 words

Abstract

In this paper, we consider ways of representing passenger journeys and more aggregate flows in public transport networks. Since the objective of travel is never to move specifically from one transit stop to another, there is a need to understand these flows as being associated with geographic regions (as perceived by the travelers), at certain times of the day, and/or as fitting in with the transit system's scheduled services. Using a spatial-temporal representation of flows common to transport geography, we show how large data sets, such as those associated with automated fare collection systems (smart cards), can be used to identify major flow patterns. Moreover, these flows can be further analyzed using additional information, such as vehicle occupancy, frequency of travel, or socio-economic variables associated with the farecard type. Based on these flow patterns, we also identify ways in which these emergent patterns can be exploited to improve transit service planning and to promote travel demand management.

1. Introduction

A transit network consists of passengers performing journeys on transit vehicles. These journeys and the characteristics associated with each of them give rise to a wide range of flows in the network. The most commonly analyzed flows are the flows of the passengers themselves. However, the multitude of characteristics attached to the passenger journeys can lead to a proliferation in the types of flows that occur in the network as well. This typology can be analyzed to understand the network better. Some examples of these can be journeys that are infrequent versus regular travel, the journeys associated with a particular type of travel experience or level of service (LOS) such as travel time or occupancy, journeys constituted by people of a particular age group/occupation, etc. Since the objective of travel is never to move specifically from one transit stop to another, there is a need to understand these flows as being associated with geographic regions (as perceived by the travelers) or as fitting in with the transit system's operations network.

To understand passenger flows, we usually count the number of passengers moving from one stop or a defined geographical region to another within a predetermined time window. We represent these movements as origin-destination (OD) matrices, often for specific time intervals in the day. However, such a representation of flows has the following drawbacks (*1-6*):

1. The origin regions and destination regions an analyst may define might not be how these regions are perceived by travelers;
2. Travelers commuting in different directions might have different perceptions of attributes of their journey, depending on trip purpose, access and egress characteristics, etc.;
3. Travelers might consider different sets of stops at different times of the day, in order to complete their journey; and,
4. The time of day at which the transition between different perceptions occurs is not easily captured and may vary depending on the traveler and his/her activity schedule.

Due to these variations in traveler perceptions, we need an approach that identifies the primary channels of flow in space and also captures the time variation of these channels, while falling back only on basic assumptions regarding what constitutes a journey and how people perceive the transport system. Some examples of these assumptions are:

1. In the downtown area, 500 people traveling in the same direction might constitute a "significant" flow, while a "significant flow" might only be 100 people in the suburbs.
2. Five transit vehicles with a sub-par occupancy in a particular region of space-time might constitute a passenger flow worth remediating through changes in service.
3. People might consider a 15-minute window to consider alternative paths for traveling to work in the morning or a 30-minute window for returning from work in the evening.

This paper discusses a method of identifying such flows, while also identifying regions of flow as observed in the network, rather than these flows being predefined.

The paper begins with a discussion of spatial-temporal passenger flows, from the concepts of space-time in transport geography. This represents an important extension of these concepts of spatial-temporal flows to "big data" from public transit farecards. The latter sections discuss and demonstrate how a variety of flow types can be identified and analyzed to support transit service planning. Specifically, identifying flows of "similar" journeys while considering different characteristics can be beneficial towards:

1. Providing services for journeys occurring in a corridor
2. Managing the flows themselves to facilitate more efficient movement

At the core of the analysis is the view that journeys are multi-dimensional vectors in a 3-dimensional space-time (*7-9*). After describing the dataset used for analysis, we describe methods to generate clusters of flows in space-time, and then describe several examples to illustrate how these clusters can be used for service planning applications.

This paper uses data from the Automated Fare Collection (AFC) system implemented in Brisbane's public transit network in the form of smart cards called Go Cards. The AFC system in Brisbane is a closed system, i.e. passengers are required to tap-on and tap-off during their journeys. The dataset is therefore able to provide information regarding boarding and alighting locations and timestamps. It also provides encrypted but consistent IDs of the cards used on the trips. The penetration rate of the farecards is 85-90% of all journeys in Brisbane, thus generating a nearly complete universe of journeys within the transit network.

2. Multi-dimensional Journey Vectors

The location of transit stops in Brisbane is shown in Figure 1. The journeys observed in a transit system occur between any two of these stops. They can be considered as vectors starting with the boarding time at the origin stop and ending with the alighting time at the alighting stop.

The time period under consideration for the purpose of this analysis is 6am to 9am on March 1 2013. During this period, a total of 110,504 traceable journeys occurred in the network. These journeys have been traced at 99% transparency in Figure 2. Thus, to represent a journey vector in the spatial temporal space, we need to have four properties of the vector: the boarding and alighting locations and timestamps.

The perception and comparison of journey vectors usually stop here. One tends to say that two journeys are similar to each other if they occur between the same origin and destination, on the same mode and at a reasonable separation of time. This, however, leads to neglecting a lot of additional information related to a journey and a comparison between two journeys. For example, consider the four journeys in Figure 3 and the two journeys in Figure 4.

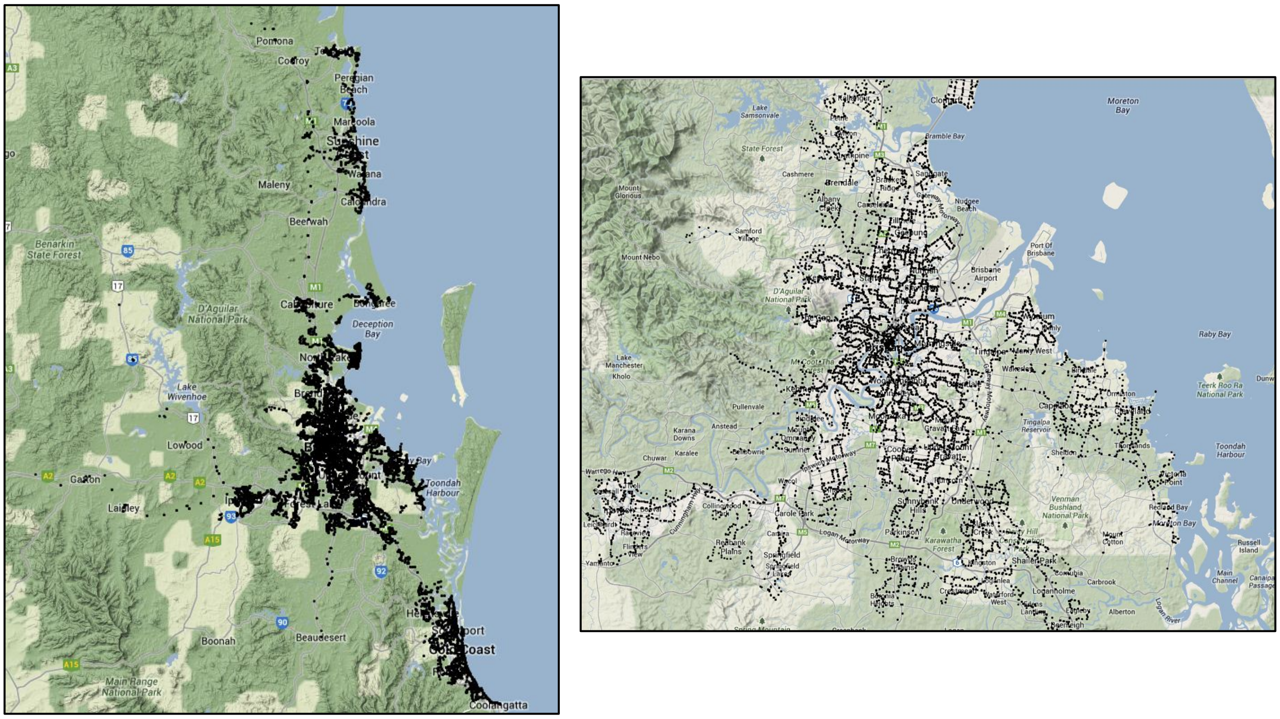


Figure 1: Location of transit stops in Brisbane

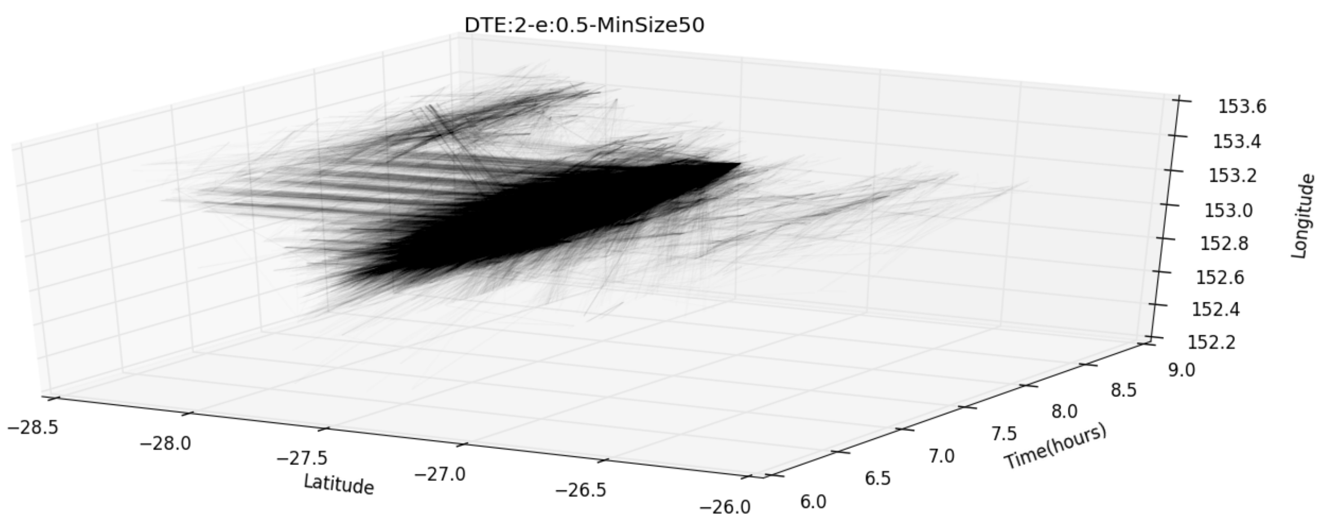


Figure 2: Spatial-temporal representation of journeys in Brisbane from 6am to 9am on March 1 2013

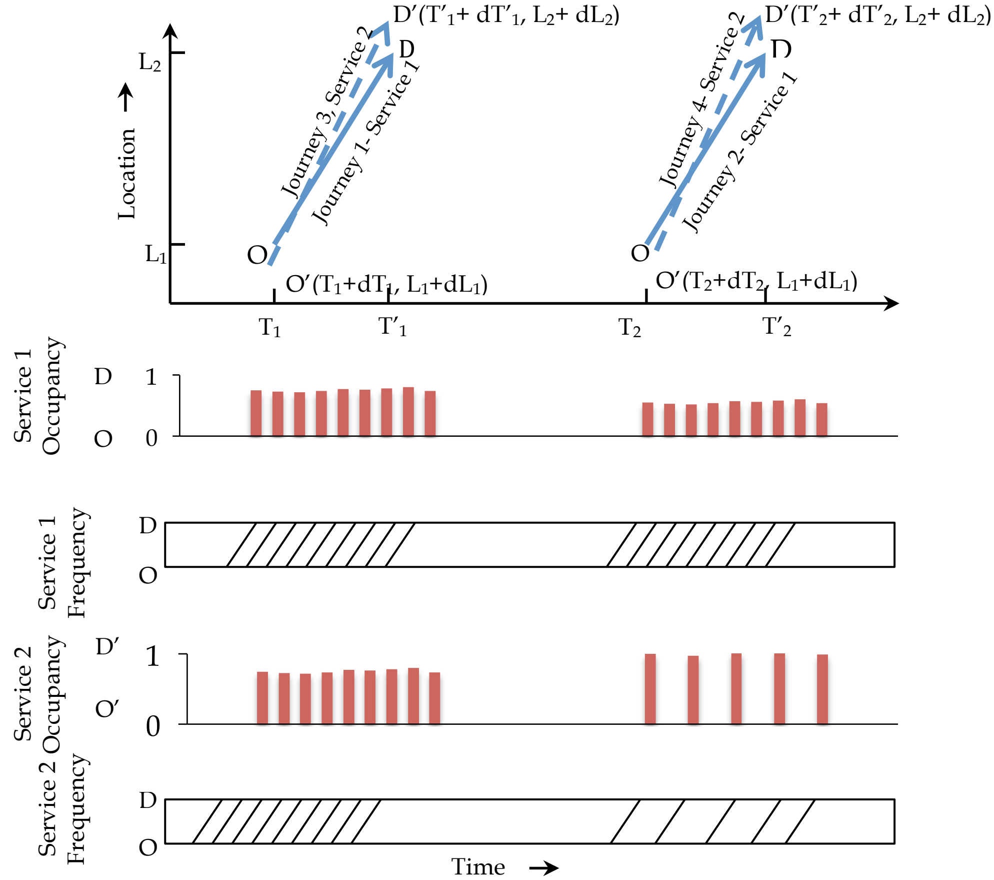


Figure 3: An illustration of multiple dimensions of journey vectors

From Figure 3, the following might be justifiable conclusions:

1. From a traveler's perspective, the similarity between Journey 1 and Journey 3, and Journey 2 and Journey 4, might be higher than any other combination. This is despite the fact that neither of these two pairs of journeys occurs between the same origin and destination stop or even at the same time.
2. For the network operator, from the perspective of passenger LOS, Journeys 1 and 3 are quite similar. Journeys 2 and 4, however, are significantly different. Providing information to passengers performing journeys of type 4 on service 2 regarding the availability of service 1 would be an important focus. In other words, the operator would be interested in the combined performance of two sets of journeys.
3. Apparently similar journeys can also be quite different in terms of the waiting times, reliability of the service on which they are performed, accessibility of stops, availability of bike racks, vehicle occupancy, etc.

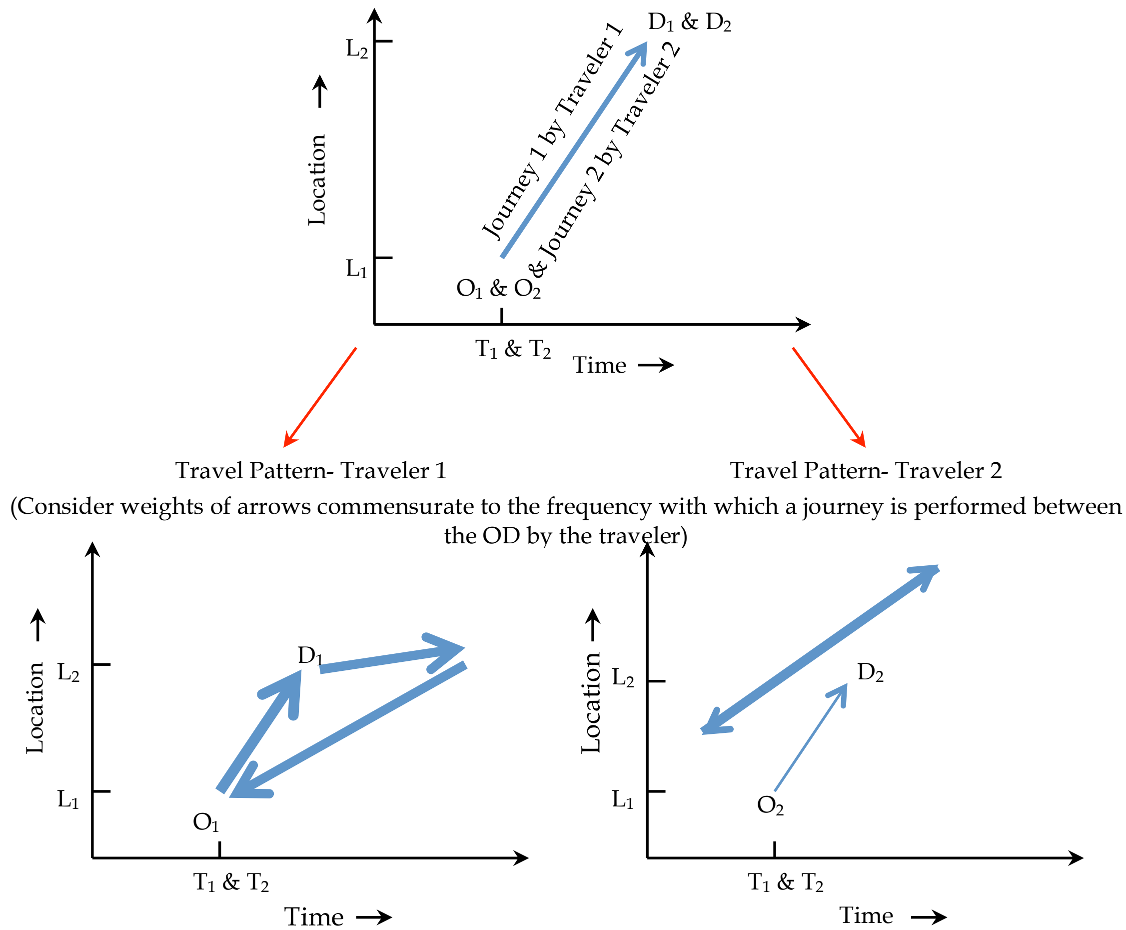


Figure 4: Journeys that have very dissimilar impacts on the transit network

In Figure 4, two travelers perform a journey together between the same OD. The travel pattern of traveler 1, shown in the bottom left graph, indicate that he frequently travels between the OD pair under consideration. The travel pattern of traveler 2 (bottom right), however, indicates that the journey between the OD pair under consideration was a one-time (or infrequent) journey for him. From the point of view of the network, journey 2 levies an infrequent load on the service on which it was performed and may, in turn, relieve another service from an expected journey.

Hence, any journey in a network is a unique event and has a large number of attributes that can be attached to it. While analyzing and comparing journeys in a network, with the availability of rich datasets, it is possible to look at several attributes attached to journeys. For some applications, attaching these attributes might even be mandatory. The choice of attributes and the importance of each depend on the application for which the comparison is performed.

3. Comparing Journey Vectors

Our objective in this part of the analysis is to identify the important spatial-temporal corridors of passenger flow from the suburbs to the city, from 6am to 9am on a typical weekday, in Brisbane. For this, let us consider the simplest instance when journey vectors have only the four basic spatial-temporal attributes attached to them: boarding time, boarding location, alighting time and alighting location. The location dimensions of these vectors actually consist of a pair of latitude and longitude. Hence, the vector in effect has 6 dimensions: the origin of a journey, at coordinates (Lat_1, $Long_1$); the destination, at (Lat_2, $Long_2$); the tap-on time associated with the journey (T_1); and, the tap-off time (T_2). The journey vector ($\hat{j}$) can then be represented as

$$\begin{Bmatrix} Lat_1 \\ Long_1 \\ T_1 \\ Lat_2 \\ Long_2 \\ T_2 \end{Bmatrix}$$

A journey vector with additional dimensions can look something like-

$$\begin{Bmatrix} Lat_1 \\ Long_1 \\ T_1 \\ Lat_2 \\ Long_2 \\ T_2 \\ 1 \\ O \end{Bmatrix}$$

Here, the 7th dimension of the vector is a binary variable, which has a value of 1 if the journey is in the set of frequent journeys of the user and 0 otherwise. The 8th dimension, 'O' represents the occupancy of the vehicle on which the journey was performed.
At this point, there are two attributes of a generalized journey vector hindering meaningful comparison between the vectors:

1. The presence of both discrete and qualitative variables for which arithmetic operations might not be meaningful
2. Different variables possessing different units, preventing direct inter-dimensional comparisons

Normalization for meaningful intra and inter-dimensional comparison
To effectuate meaningful comparison, we need to initiate some kind of normalization. The normalization process could involve the following processes:

1. Setting a reference level/unit with respect to which normalization shall be initiated
2. Deciding the threshold of acceptable dissimilarity with respect to each dimension, if all other significant dimensions were identical between the two journeys
3. Choosing the normalization factors

An instance of the normalization process using journey vectors with the eight dimensions above is illustrated in the following paragraph. This encompasses the normalization process for the 6-dimensional vector that we shall use in further analysis.

Suppose we are interested in determining if two spatially and temporally closely spaced journeys are similar in terms of the travel experience (quantified only in terms of occupancy). Since the quantification is only in terms of occupancy, journeys occurring on the same transit vehicle shall be considered equivalent. Suppose we are comparing two journey vectors $\hat{J}_1$and $\hat{J}_2$. For meaningful analysis let us consider that the two journeys are of the type J_2 and J_4 in Figure 3. Based on the process outlined above, the comparison could be performed as follows, as only one example among several possible variations:

1. The reference level is set as a quantity similar to distance (say distance equivalents) measured in km-equivalents, with the normalizing factor for distance being equal to 1km-equivalent/km
2. We know that dissimilarity shall be measured in terms of distance, time and occupancy. For example, an acceptable level of dissimilarity in terms of distance can be assumed to be 1km at the origin and destination each (hence 2km total), a 15-minute window at each end and 10% difference in occupancy. [Note that we can have different normalization factors for access and egress distances. Thus, we can, for example, set 1km access = 1km-equivalent and 1.4km egress = 1km-equivalent.]
3. Since a journey for which the difference between the origin and destination location with respect to the origin and destination of a reference journey is greater than 2km (=2km-equivalent) is not likely to serve as an alternative to the reference journey, we can use the 2km-equivalent level as the threshold cutoff for dissimilarity.

   Based on this, the normalizing factors would be:

   a. 2km-equivalent/30 min = 4km-equivalents/hour
   b. 2km-equivalent/(10% Occupancy) = 20km-equivalents/(max. occupancy)

   Hence, we have the 3 normalizing factors for the purpose of the analysis. Note that the dimension corresponding to the 'frequent trip flag' is inconsequential to the analysis (as we choose to neglect familiarity of passengers with each other) and will therefore have a normalizing factor of 0 km-equivalents.

The dimensions are multiplied by the corresponding normalization factor to obtain common units for each of them. Now, inter and intra-dimensional comparison of the vectors should be possible.

Deciding upon a distance metric

A wide variety of distance metrics can be employed to compare two journey vectors. Different metrics (such as Manhattan distance/ Bays-Curtis dissimilarity/variations of Minkowski distance) can be suitable for different applications and normalization approaches. The distance metric used in the example illustrated here is the Euclidean distance.

Intra-dimension variability in normalizing factor
In several applications, the weight given to a particular dimension of a journey might differ under different conditions. For example, while considering walkability, we might want to consider a distance of 500m to a stop as an acceptable walk in case of suburban areas, and a greater distance if the stop is located downtown. By appropriately programming the algorithm for computing dissimilarity, we can capture the essence of intra-dimensional variability of the normalizing factor.

4. Identification of Flows from Aggregation of Similar Journeys

Once we have decided upon an approach to compute the level of similarity or dissimilarity between two journey vectors which best mimics the purpose of our application, we can proceed towards finding aggregations of "similar" journeys. One approach towards finding these aggregations is using the DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm (*10*).

Suppose we have performed the normalization of all 6-dimensional vectors as outlined in the previous section. To proceed with the aggregation, we need to decide the minimum size of aggregation that we are looking for; i.e., if we want every pair of journeys within the similarity bounds to be represented, or if we want only denser aggregations. Thus, the choice of the minimum cluster size does depend to a certain extent on the number of aggregate clusters we want. An array of minimum sizes within an intuitive range can be tested. In this case, 5 sizes - 50, 100, 300, 500 and 1000 journey vectors - were considered. The DBSCAN algorithm was then implemented on the normalized vectors for each of these thresholds and is described below.

Suppose the minimum size of clusters being considered is M. For each vector, the number of other vectors within a Euclidean distance of 2km-equivalents was determined. If a journey had a minimum of M neighbors within the threshold Euclidean distance, it was identified as a core journey. If a journey was not a core journey but was within a 2km-equivalent radius of a core journey, it was labeled as a non-core journey. If a journey was neither a core journey nor a non-core journey, it was considered as noise.

After this labeling process was complete, the process of creating clusters was initiated. If two or more 'core' journeys were within the threshold radius of each other, they, along with the non-core journeys associated with them, were coalesced into the same cluster. The result of this process can vary from having each core journey forming its own cluster to all core journeys being coalesced into a single cluster.

In this way, the number of clusters, along with a set of metrics for each cluster and for the overall performance of the clustering, was determined for each of the five values of M listed above. Summary statistics are provided in Table 1. The number of clusters, along with the size of the clusters, and other metrics related to the identified clusters, can be used to determine the most suitable threshold cluster size for the required objective.

Table 1: Summary Results of Clustering for the Five Values of M

Min. Size (M)	Clusters	% of all journeys in clusters
50	41	67
100	26	59.5
300	9	35.4
500	3	20.9
1000	1	1.02

From Table 1, for the purpose of identifying channels of flow from the suburbs to the city center, a choice between 100 and 300 seems appropriate. Such a value of M seems to balance the number of clusters, but at the same time ensuring that the clusters have sufficient journeys to represent major movements. At this range, the clusters capture around half of the total journeys in the Brisbane region during this time interval. The discussion and visualization of the results follow in the next section.

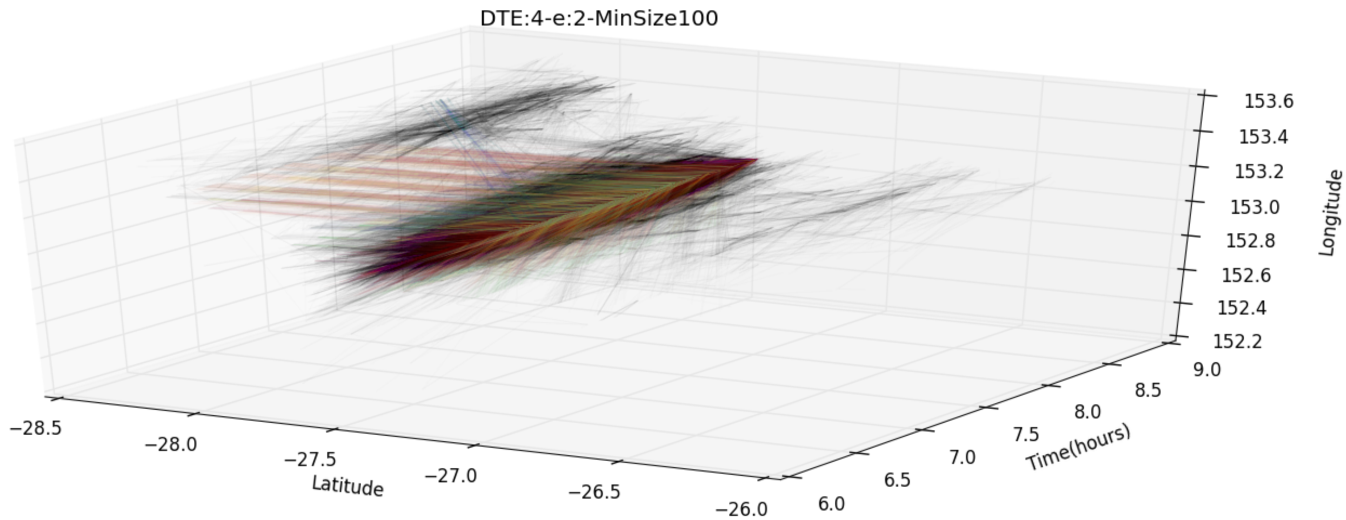
In this case, the process of performing aggregation initiates with an intuitive sense of the problem size. Alternately, looking at the results of aggregation using a broader range of normalization coefficients and minimum cluster sizes could instead provide a deeper understanding of the mechanics of the clustering phenomenon. Also, it may be possible that different minimum cluster sizes might actually be the best fit for different parts of the city.

5. Analyzing Results of Multi-dimensional Aggregation

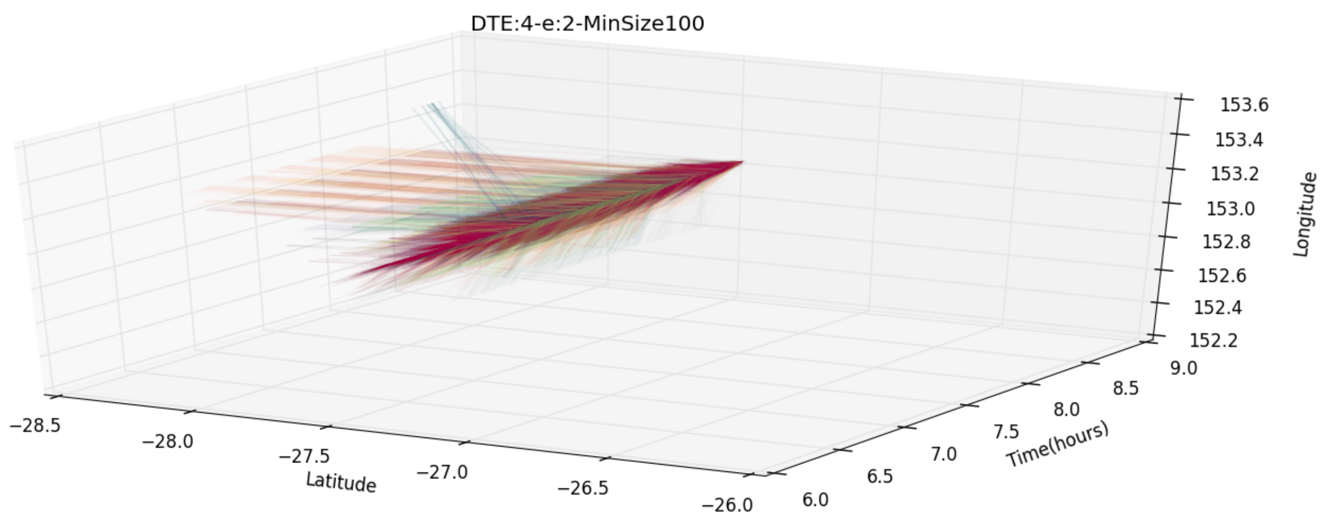
In this section, we shall provide and analyze results corresponding to the chosen threshold M (around 200) and also some similar thresholds (100 & 300). For the purpose of detecting significant flows from suburbs to the city, identifying somewhere between 9 and 26 major flows seems reasonable.

Figures 5 through 7 illustrate the results obtained from performing the clustering with the three chosen levels of minimum cluster size: 100, 200 and 300 respectively. Each of the figures show the same image repeated: once with the journeys classified as noise and then without them for clearer illustration. In this section, we highlight some key features that stand out and can be useful for understanding the results in some detail.

A threshold cluster size of 100 provided us with 26 clusters. As is clear from Figure 5, several of these clusters are associated with journeys that start from remote suburbs, probably made by people coming into the city for work. Also, the central downtown cluster is the largest and encompasses most of the inner suburbs of the city. As noted earlier, this threshold captures 59.45% of all journeys.

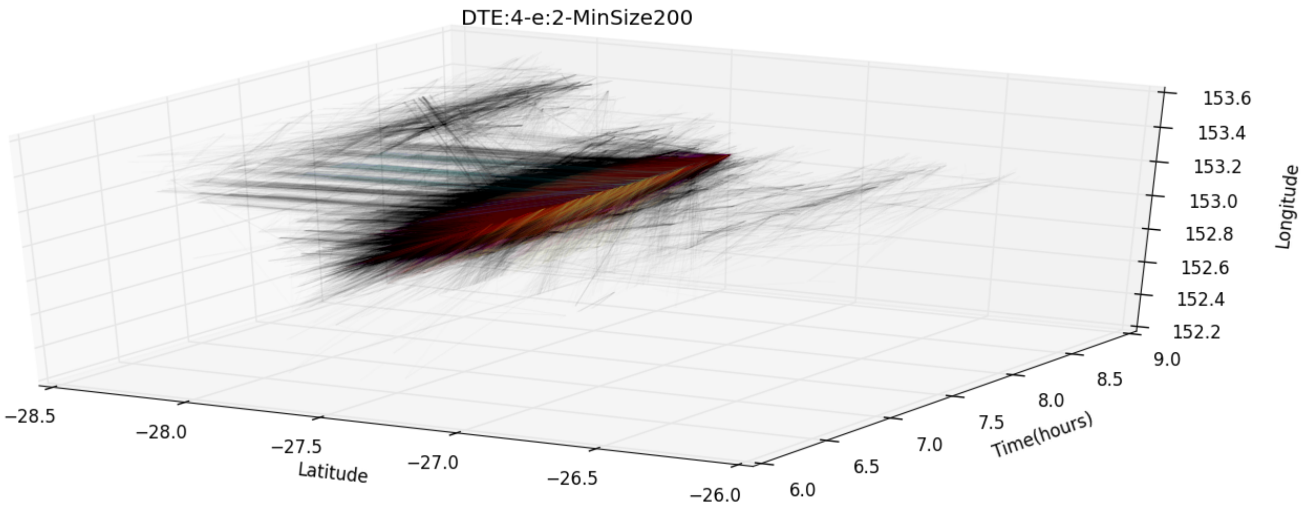


(a) Representation of all journeys

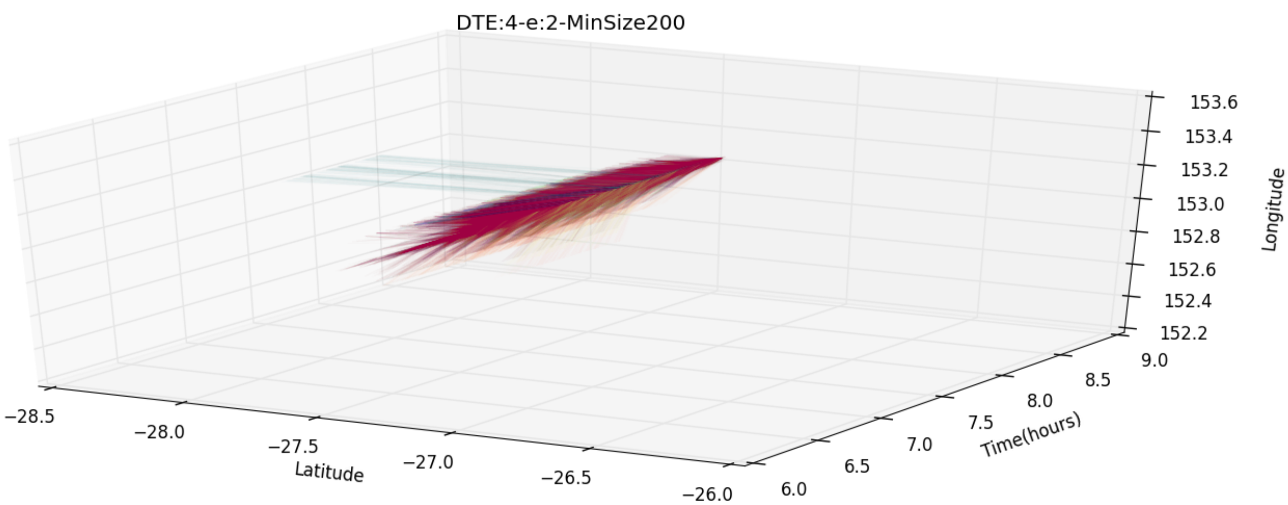


(b) Representation of clustered journeys without noise

Figure 5: Observed flows with minimum flow size of 100



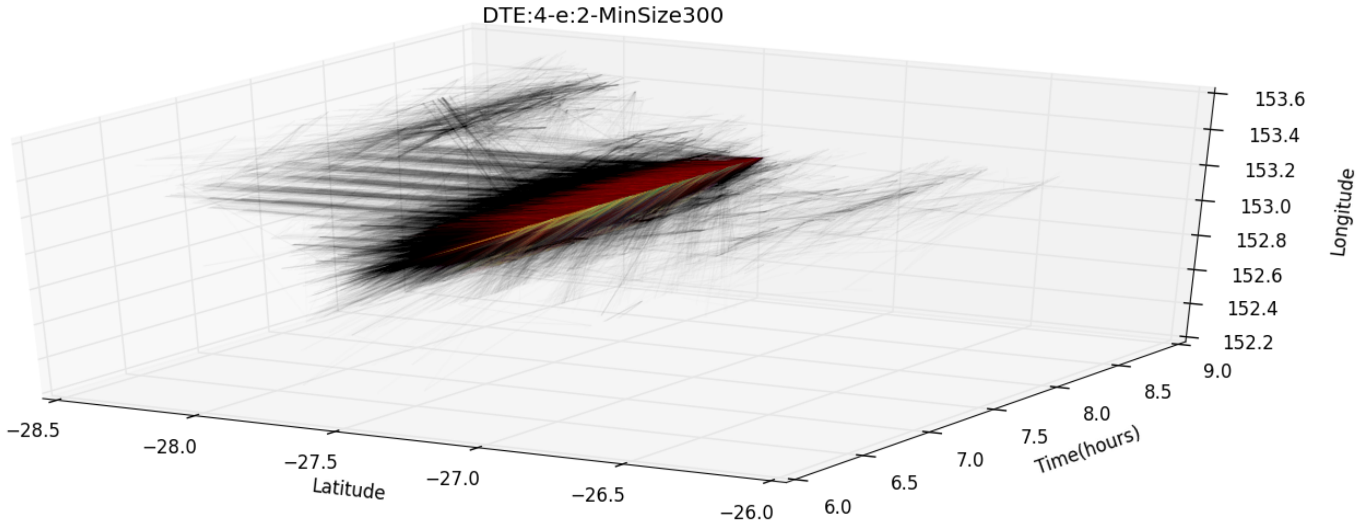
(a) Representation of all journeys



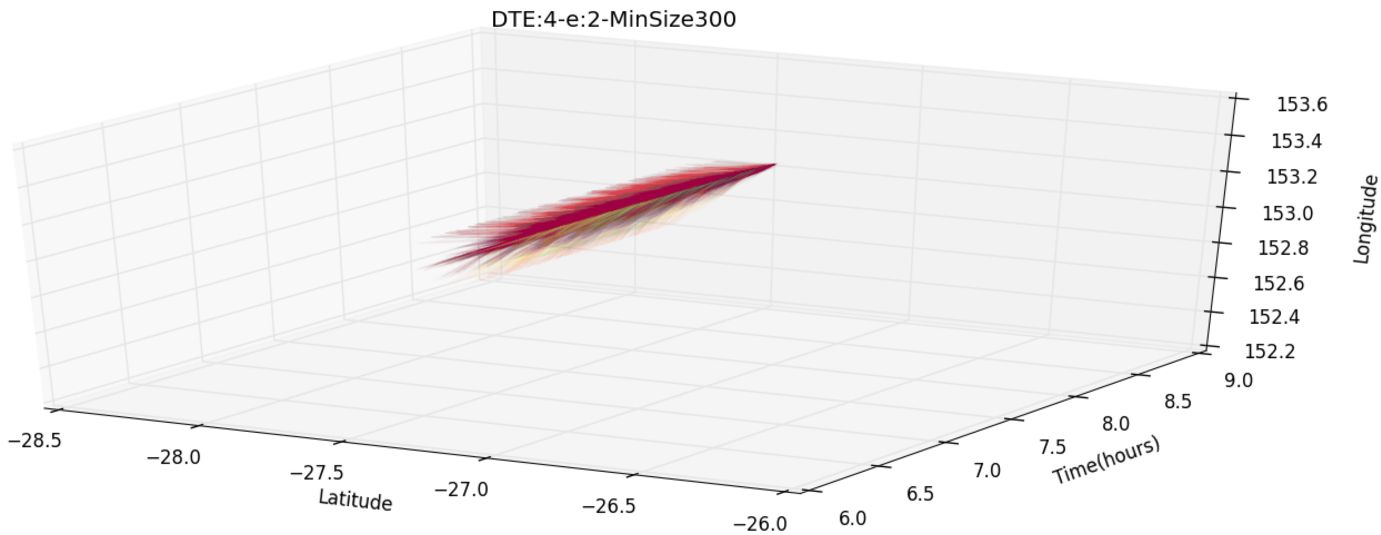
(b) Representation of clustered journeys without noise

Figure 6: Observed flows with minimum flow size of 200

The solution with M=200, shown in Figure 6, resulted in 14 clusters and accounted for 45.55% of all journeys. With this threshold, most of the flows from the outer suburbs seem to have disappeared. The reduction of 12 clusters, when compared with M=100, is associated with breaking up of a large (probably downtown) cluster into multiple smaller clusters with a loss of some journeys to noise. This also helps explain the lower fraction (45.44%) of journeys being included at this threshold.



(a) Representation of all journeys



(b) Representation of clustered journeys without noise

Figure 7: Observed flows with minimum flow size of 300

The threshold of 300, shown in Figure 7, sees the disappearance of all the clusters from the outer suburbs. The downtown cluster is the smallest and more flows from the inner suburbs have been detected. The fraction of journeys explained by this clustering, as expected, is the smallest (35.42%).

6. Interpreting Shapes of Clusters: Studying the Colossal Downtown Cluster

Journeys in the downtown area are quite numerous and closely packed. As a result, a large number of journey aggregations appear in the downtown area. Being relatively close to each other both spatially and temporally, several of these are combined together to produce a cluster expanding both spatially and temporally. This cluster (the maroon cluster in Figures 5, 6 and 7), often the largest, represents the burst of activity occurring in the downtown area throughout the day. Also, as the minimum size of aggregation (M) increases, the large downtown cluster tends to diminish, as the number of core journeys over which chaining can occur decreases. We are then able to obtain a clearer image of the major flows from the inner suburbs, which are otherwise aggregated in the bigger cluster. The remaining large downtown cluster therefore represents journeys with an origin and destination within downtown itself.

This downtown cluster may not always be necessarily relevant to the analysis. For example, while looking for flows from suburbs to the city center, a large downtown cluster is not too helpful. A more disaggregate analysis can be used in cases where we are interested in tapping flows from the inner and outer suburbs together. This could be an added utility of using intra-dimensional variability of normalizing coefficients.

One should also remember that dissimilarity, measured as the distance between two journeys in the reference units (km-equivalents in the case here), is not a transitive property. Therefore, if we consider a core journey, J, and two non-core journeys, $\hat{\jmath}_1$ and $\hat{\jmath}_2$, associated with J, $\hat{\jmath}_1$ is within the threshold dissimilarity of J and $\hat{\jmath}_2$ is also within the threshold dissimilarity of J, but j_1 and j_2 can be more dissimilar than the threshold dissimilarity. This non-transitivity becomes more prominent as multiple core journeys are coupled together in a cluster. This is because so long as two journeys are associated with the same core journey, the maximum level of dissimilarity that they can have is twice the threshold level. Once the cluster starts expanding, the dissimilarity between the most dissimilar core and non-core journeys within the same cluster can increase significantly.

The mechanics of formation of colossal clusters are a combination of blobbing and chaining of core journeys. The extent of these phenomena can somewhat be captured by the core to non-core journey ratio in the cluster. Blobbing, or the phenomenon of core journeys being very close to each other along a particular dimension, leads to coherence along that dimension. Chaining, or the occurrence of core journeys at a separation near to the threshold separation, on the other hand, leads to more dispersed clusters.

Coherence and dispersion can be studied individually for each dimension of the journey vectors. For example, a highly spatially dispersed cluster can temporally coherent. Such a cluster would represent the presence of transit vehicles at multiple locations at the same time. Similarly, if a large cluster is spatially coherent but temporally dispersed, it indicates that consecutive transit vehicles at the same or similar location are able to observe high demand for a long period of time.

From the viewpoint of transit operations, it is not usually good to observe high spatial and temporal coherence concurrently as this might suggest services are heavily utilized (or even oversaturated) at those spatial-temporal locations, while preceding and succeeding services may be substantially under-utilized.

Hence the shape of clusters along any particular combination of directions can help us to identify potential measures to achieve better transit operation: increasing service frequency to provide more capacity, decreasing frequency to reduce capacity, or provide demand management strategies to shift passengers to off-peak periods.

Finally, on major drawback of this approach is that it is unable to capture the variation of appropriate spatial thresholds for aggregation from different parts of the city. Though it might be possible to capture this through intra-dimensional variation through the normalization factor or by adding dimensions in particular geographic regions, the option has not yet been explored.

7. Contribution and Applications of the Method

The proposed flow clustering approach provides an example of how we can use rich data being produced in modern transportation networks to understand movements based on the observed perception of travelers in the system. It also shows how different levels and types of aggregations can be useful in analyzing different flows.

So far, we have illustrated only the simplest application of the outlined approach – identifying spatial-temporal flows with no additional attributes attached to the journey vectors. However, some attributes can be easily added by mining smart card data include vehicle occupancy information, the nature of the journey (infrequent or frequent), and characteristics of the traveler (child/student/adult/retired, among several others).

In an analysis on determining the type of travelers in a transportation system, Kieu et al. (*11*) consider temporal and spatial regularity as the basis for classification. The analysis provided here can be used to classify travelers not just in terms of regularity in the spatial and temporal domains but also specificity in these domains. For example, if we assume that the important flows in a city from the period from 6am to 9am can be classified into two main groups, the first constituted by flows directed from the outer suburbs to the city and the second by those from the inner suburbs to the city and within the city, we can analyze the travel patterns of individual travelers and determine the flows within which his/her journeys fall.

The identification of flows of the first type (outer vs. inner suburbs) has already been illustrated. To capture flows under the second category, we need to change the parameters used for clustering. We should be looking for larger clusters with a lower radius within the spatial domain and more flexibility in the temporal space. Significant flow patterns are seen to emerge from the inner suburbs when we use a minimum cluster size of 750 with a reduced radius of 1.5km-equivalents and a higher leverage along the temporal domain with a normalization factor of 0.5km-equivalents per hour. Using these parameters, we capture 32% of the journeys and restrict the size of the downtown cluster (which can be broken down further, if required). The results for this are shown in Figure 8.

Gordon et al. (*12*) discuss the possibility of identifying journeys in smart card data that involve latent short-term activities. The analysis here, by finding common journey durations between specific origins and destinations, should be able to detect if certain journeys between the same OD regions during similar time intervals differ significantly in travel time. Such journeys can be contenders to having activities hidden within them.

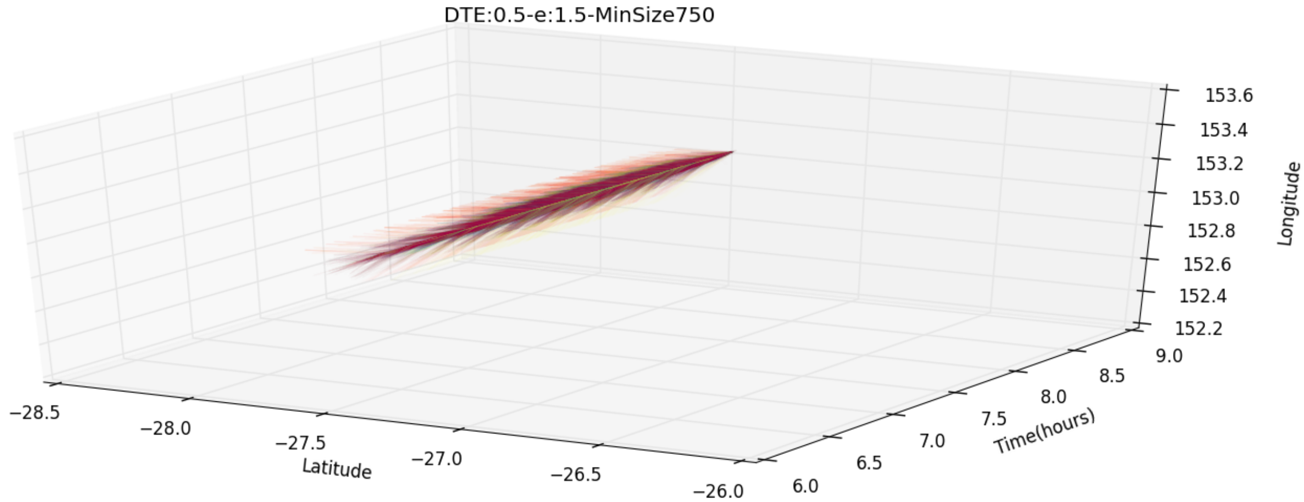


Figure 8: Flows occurring in the inner suburbs

This approach also opens up the possibility of providing broader route choice to passengers. In the following example, we consider the case of a particular cluster in an iteration of the spatial temporal clustering with time equivalence of 0.5km, radius of 0.5km and minimum cluster size of 50. This is done to find clusters with tight spatial distribution. The duration of the journeys in the cluster are plotted against time and colored by the route on which they were performed. The results are shown in Figure 9.

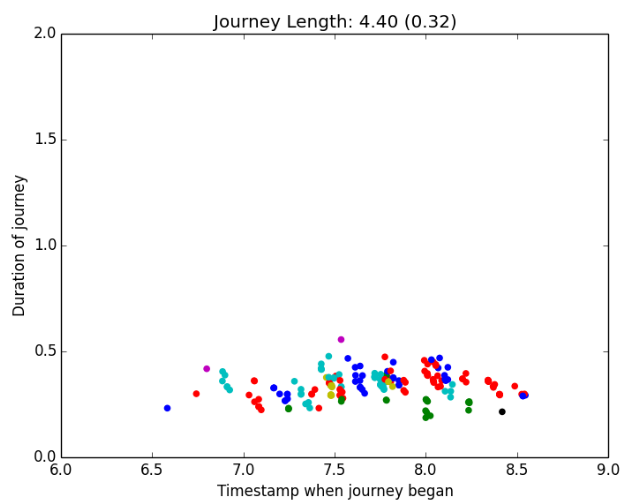


Figure 9: Clustered journeys with tight spatial distribution

The journeys are of length 4.4km on an average with a standard deviation of 320m. Concentrating on a time when services are available on multiple routes, say 8am in Figure 9, we can see that even though the travel time for the route in green is significantly lower, more passengers seem to be utilizing the routes in red and blue.

For services operating at or near capacity, or which exceed capacity occasionally, it might be useful to determine if all the observed services in the spatial-temporal interval are over-utilized or if there is a possibility of distributing the load to other existing services. Further, it might be a useful exercise to determine the proportion of infrequent vs. frequent journeys occurring on such vehicles and to find whether it is infrequent journeys that are causing the overcrowding. Comparing these proportions and results with other services can help us to design incentives for certain travelers to switch services. Or, we can alternatively explore the possibility of changing the stopping pattern of different services to attain more uniform utilization.

8. Conclusion

Journeys in a transit network and the characteristics attached to them can be used to identify channels of flow of passengers and the locations in space-time where various phenomena related to journey occur: poor level of service, heavy vehicle occupancy, etc. This paper has proposed a way of detecting these flows and phenomena. Most importantly, the analysis is able to aggregate the characteristics associated with journeys in a diffuse region of space and time without an explicit reference to the region of space-time to be considered or the routes and modes to be analyzed. This allows us to look more specifically at spatial and temporal patterns in passenger flows, and to devise solutions to improve service quality for passengers.

References

[1] Geurs, K. T., & Van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography*, *12*(2), 127-140.

[2] Neutens, T., Schwanen, T., & Witlox, F. (2011). The prism of everyday life: towards a new research agenda for time geography. *Transport Reviews*, *31*(1), 25-47.

[3] Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: positive and normative implementations of various accessibility indicators. *Journal of Transport Geography*, *25*, 141-153.

[4] Kwan, M. P. (1998). Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geographical Analysis*, *30*(3), 191-216.

[5] Wu, Y. H., & Miller, H. J. (2001). Computational tools for measuring space-time accessibility within dynamic flow transportation networks. *Journal of Transportation and Statistics*, *4*(2/3), 1-14.

[6] Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems*, *5*(3), 287-301.

[7] Kobayashi, T., & Miller, H. (2014). Exploratory visualization of collective mobile objects data using temporal granularity and spatial similarity. In *Data Mining for Geoinformatics* (pp. 127-154). Springer New York.

[8] Dodge, S., Weibel, R., & Lautenschütz, A. K. (2008). Towards a taxonomy of movement patterns. *Information Visualization*, *7*(3-4), 240-252.

[9] Dykes, J. A., & Mountain, D. M. (2003). Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications. *Computational Statistics & Data Analysis*, *43*(4), 581-603.

[10] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (*KDD*-96), 226-231.

[11] Kieu, L. M., Bhaskar, A., & Chung, E. (2014) Transit passenger segmentation using travel regularity mined from Smart Card transactions data. Paper presented at the Transportation Research Board 93rd Annual Meeting, January 2014, Washington, DC.

[12] Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H., & Attanucci, J. P. (2013). Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board*, *2343*, 17-24.