ASSIGNMENT 2

ISSS602 – Data Analytics Lab

SINGAPORE URBAN TRAIN SYSTEM
Insights from Station Traffic Interactions
10-Oct-2021

Le Vu Anh Phuong (Joshua)
vaple.2021@mitb.smu.edu.sg

Contents

1. OVERVIEW

Origin-Destination (OD) data describes the total number of commuters on a transportation network at a particular time, such as Singapore's Mass Rapid Transit (MRT) and Light Rapid Transit (LRT) train systems. Due to the recent development of technologies, this data has been digitalized and made easily obtainable by the Land Transport Authority (LTA) for government as well as independent studies to provide insights on various questions such as congestion, system performance and stations linkages.

The data is commonly transformed and displayed in a matrix form – two-dimensional array of cells where row and column headers represent the origin and destination stations respectively, and each cell value is the number of trips that start and end on the respective OD pair. A model of a simplified OD matrix can be illustrated in *Table 1*.

		Destinations			
		Station 1	Station 2	…	Station n
Origins	Station 1	a_{11}	a_{12}	…	a_{1n}
	Station 2	a_{21}	a_{22}	…	a_{2n}
	…	…	…	…	
	Station n	a_{n1}	a_{n2}	…	a_{nn}

Table 1: Generalised OD Matrix

2. OBJECTIVES

The main objective of this study is to utilize appropriate clustering analysis methods to group MRT/LRT stations into homogeneous groups, based on the OD commuting interactions, followed by exploring the clusters' characteristics. This analysis will be caried out on three typical time windows: weekday morning journey-to-work peak (7-9am), weekday evening journey-to-home peak (5-7pm) and the weekend leisure time (11am-1pm). Through this study, we aim to discover insights on how the MRT/LRT stations display interactions in traffic flows and distributions.

3. DATA

3.1 Data Used

The July 2021 OD data named "*Passenger Volume by Origin Destination Train Stations"* from LTA data Mall is used for this study. This raw data is presented in the "long" format with each OD pair occupying a row below. Suitable data transformation steps will be performed to obtain the desired OD matrix.

YEAR_MONTH: Year and month of the data
DAY_TYPE: weekend of weekday
TIME_PER_HOUR: one hour time interval. Value 5 indicates 5-6 o'clock
PT_TYPE: mode of public transport train or bus.
ORIGIN_PT_CODE: origin MRT/LRT station code
DESTINATION_PT_CODE: destination MRT/LRT station code
TOTAL_TRIPS: number of trips by weekdays and weekends from origin to destination MRT/LRT stations.

3.2 Data Preparation

3.2.1 Creating OD Matrices from Long-Form Tables

To perform the analysis on the aforementioned three windows, the first step is to extract relevant rows belonging to these time periods. Row selection - Data view functionalities are used to select by Weekday/Weekend first (*Figure 1a*), followed by the 2 peak time periods for weekday, and the leisure period for weekend. Note that by data convention of the TIME_PER_HOUR column above, the 7-9am window will be obtained by setting the criteria to "equals – 7" and "equals – 8" (*Figure 1b*). Similar method is used to prepare the weekday evening peak, and weekend leisure time.

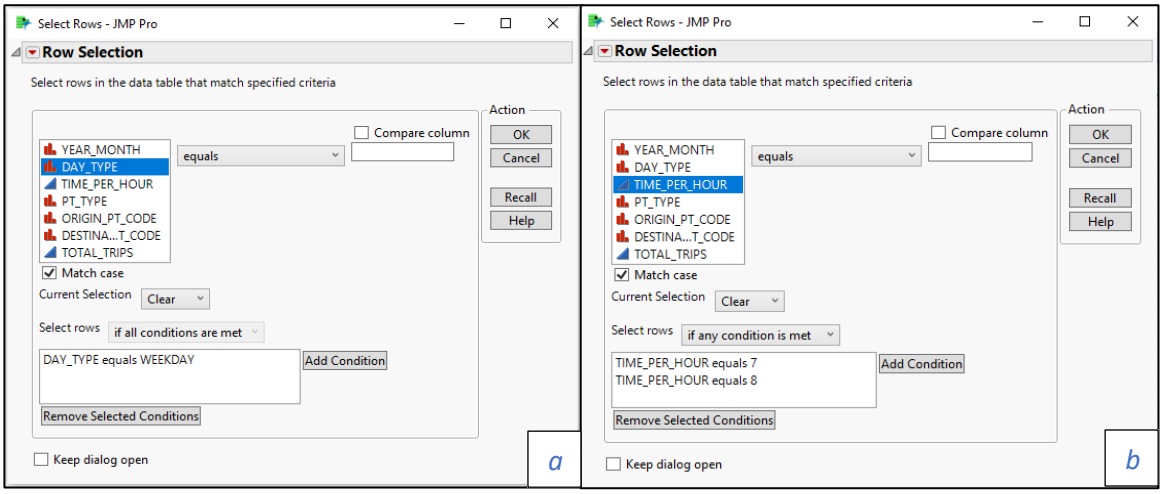


Figure 1: Data extraction by day and time

Next, Table Summary functionality is used to combine the 2-hr TOTAL_TRIPS between each OD pairs into a single row (*Figure 2a*). The resultant long-form table will only have one unique OD pair per row (*Figure 2b*), with each cell containing summation of all the trips starting from the respective origin to the respective destination over the 2-hr period.

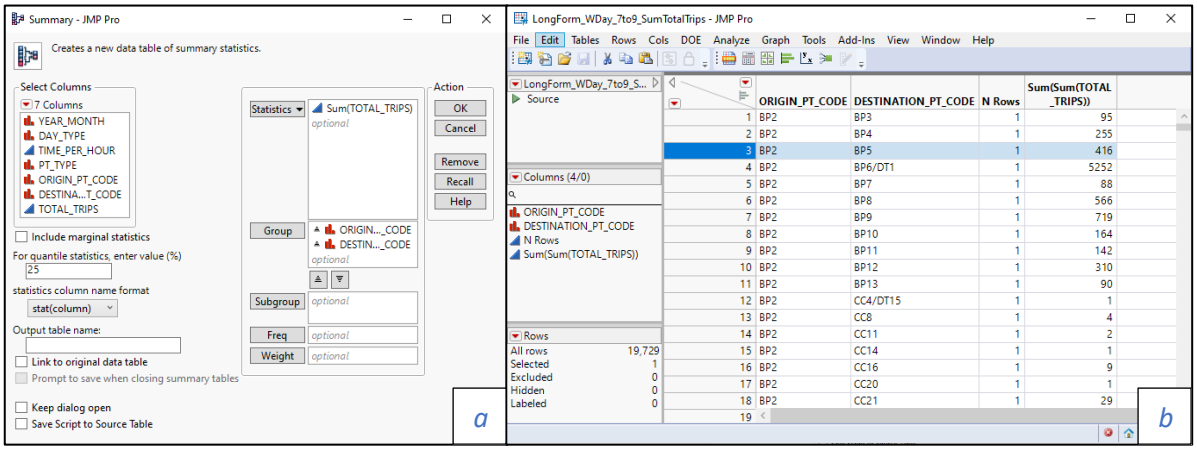


Figure 2: Combining 2-hr values

Finally, Split Table function is used to a create OD matrix from each long form table above (*Figure 3*), keeping only 3 relevant columns – DESTINATION_PT_CODE, ORIGIN_PT_CODE, and TOTAL_TRIPS. On the resultant OD matrix, replace empty OD pair values by zero as there are no traffic between them (*Figure 4*). Now the OD matrices for all 3 time periods are ready for further analysis.

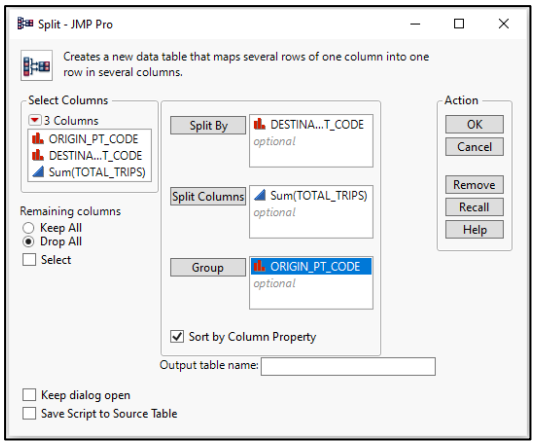


Figure 3: Creating OD Matrix from Long Form Table

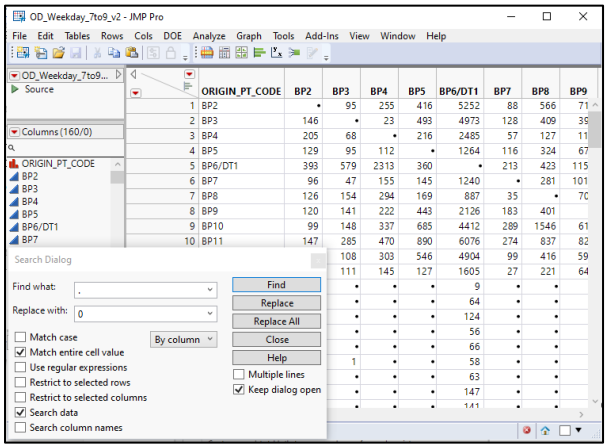


Figure 4: Replacing Empty Cells with Zeros

3.2.2 Data Normalisation

From the column overview of each OD matrix, we observe that the range of the number of trips is big, from a few dozens to several thousands. Hence, data normalisation is carried out to avoid bias towards high-magnitude columns.

Each column sum represents the total inflow into that MRT/LRT station, and it is contributed by the individual outflow from stations on each row. In this study, it is of interest to discover the interactions between stations in terms of traffic inflow and outflow. Hence each OD cell value is transformed into the percentage of total inflow to each destination column (column sum) by the following steps. The result of this normalisation is generalised and illustrated by Table 2.

		Destinations			
		Station 1	Station 2	…	Station n
Origins	Station 1	a_{11} / S1 *100	a_{12} / S2 *100	…	a_{1n} / S_n *100
	Station 2	a_{21} / S1 *100	a_{22} / S2 *100	…	a_{2n} / S_n *100
	…	…	…	…	…
	Station n	a_{n1} / S1 *100	a_{n2} / S2 *100	…	a_{nn} / S_n *100
		$S_1 = \sum_{i=1}^{i=n} a_{i1}$ (Col Sum 1)	$S_2 = \sum_{i=1}^{i=n} a_{i2}$ (Col Sum 2)	…	$S_n = \sum_{i=1}^{i=n} a_{in}$ (Col Sum n)

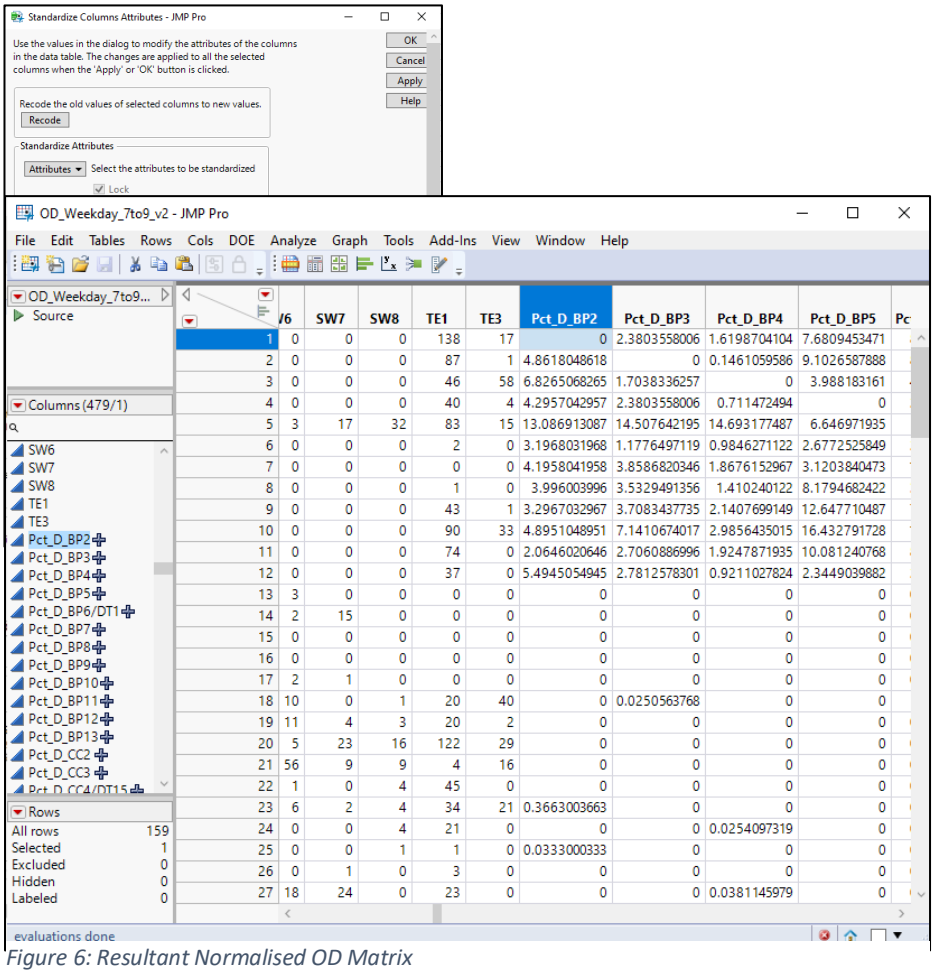
Table 2: Generalised Normalised OD Matrix

Step 1:

Firstly, we calculate the ratios of each cell value to the column sums to represent how each destination column total inflow is shared among the origin rows. This can be achieved by creating new formula columns based on the existing Destination columns, followed by

applying Standardized Column Attributes - Formula function on these new columns. In *Figure 5*, the example shows that each cell on the BP2 MRT station column will be divided by the column BP2 sum and then shown in percentage value.

Step 2:



	SW6	SW7	SW8	TE1	TE3	Pct_D_BP2	Pct_D_BP3	Pct_D_BP4	Pct_D_BP5	Pc
1	0	0	0	138	17	0	2.3803558006	1.6198704104	7.6809453471	
2	0	0	0	87	1	4.8618048618	0	0.1461059586	9.1026587888	
3	0	0	0	46	58	6.8265068265	1.7038336257	0	3.988183161	
4	0	0	0	40	4	4.2957042957	2.3803558006	0.711472494	0	
5	3	17	32	83	15	13.086913087	14.507642195	14.693177487	6.646971935	
6	0	0	0	2	0	3.1968031968	1.1776497119	0.9846271122	2.6772525849	
7	0	0	0	0	0	4.1958041958	3.8586820346	1.8676152967	3.1203840473	
8	0	0	0	1	0	3.996003996	3.5329491356	1.410240122	8.1794682422	
9	0	0	0	43	1	3.2967032967	3.7083437735	2.1407699149	12.647710487	
10	0	0	0	90	33	4.8951048951	7.1410674017	2.9856435015	16.432791728	
11	0	0	0	74	0	2.0646020646	2.7060886996	1.9247871935	10.081240768	
12	0	0	0	37	0	5.4945054945	2.7812578301	0.9211027824	2.3449039882	
13	3	0	0	0	0	0	0	0	0	
14	2	15	0	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	
17	2	1	0	0	0	0	0	0	0	
18	10	0	1	20	40	0	0.0250563768	0	0	
19	11	4	3	20	2	0	0	0	0	
20	5	23	16	122	29	0	0	0	0	
21	56	9	9	4	16	0	0	0	0	
22	1	0	4	45	0	0	0	0	0	
23	6	2	4	34	21	0.3663003663	0	0		
24	0	0	4	21	0	0	0	0.0254097319	0	
25	0	0	1	1	0	0.0333000333	0	0	0	
26	0	1	0	3	0	0	0	0	0	
27	18	24	0	23	0	0	0	0.0381145979	0	

Figure 6: Resultant Normalised OD Matrix

After carrying out appropriate column renaming (Pct_D = "Percentage Destination"), we will have additional 159 columns to the OD matrix with values indicating the percentage each station on the rows on an OD pair contribute to the total number of trips to each destination (column Sum). In other words, we have calculated how the total number of trips to each destination is distributed across origins (*Figure 6*).

4 DATA ANALYSIS

4.2 Selecting the Appropriate Clustering Method

In this study, the origin train stations will be clustered based on how similarly they contribute to the destination train stations across all columns. The number and size of each cluster should be left to vary depending on how similar the members in each cluster are to each other. Hence hierarchical clustering (HC) technique is preferred to K-mean or Normal Mixture as this technique does not require making initial guess of the number of clusters, and hence avoiding the issue of over-sensitivity on the initial selection of clusters (Jain, 1999). Furthermore, HC tends to avoid the limitations of same-size clusters of the other 2 techniques.

4.3 Performing the HC Clustering Technique

Using the OD matrix columns of Pct_D, we would like to explore how origin stations are grouped by their contribution to the destinations by the following steps:

Firstly, the HC technique is applied on the Pct_D columns (*Figure 7*). The Ward method is selected as after a trial-and-error process as it results in the most stable cluster criterion trend, and the resultant clusters are easily identified. This method is also frequently used due to its ability to consistently form homogeneous groups (Nowotny, 2003).

Secondly, the HC report is analyzed to obtain the most reasonable number of clusters.

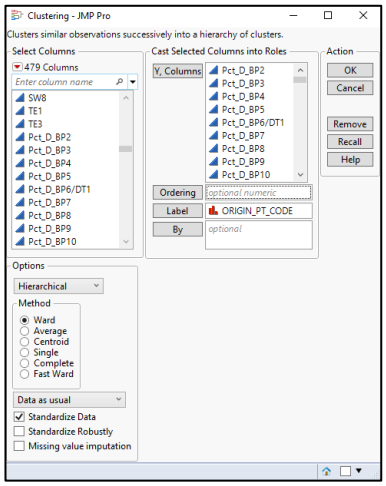


Figure 8: Performing HC by JMP Pro

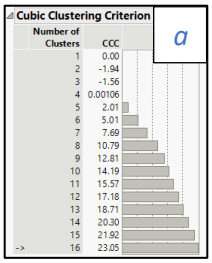
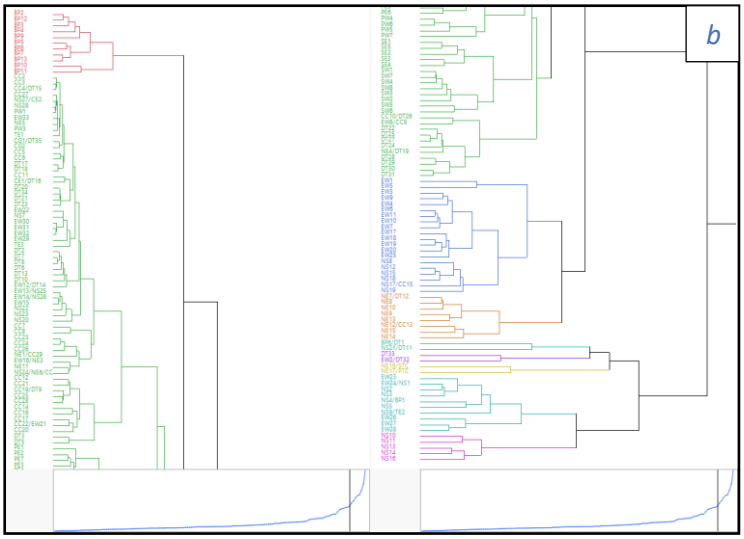


Figure 7: HC Report for Weekday Morning Peak



Taking the case of Weekday morning peak window as the example, JMP report suggests the number of clusters is 16 by the highest Cubic Clustering Criterion (CCC) value (*Figure 8a*). Adjustment is made by dragging the diamond shape on the dendrogram to 9 clusters, where the CCC is still stable, the distance diagram has a more drastic jump compared to the 16 clusters and the number of clusters with less than 5 members is reduced (*Figure 8b*). Similar approach is taken to determine the number of clusters formed for the other two time-windows. This trial-and-error procedure to adjust the number of clusters until certain characteristics are best observed is highly advisable for effective clustering analysis (Leonard Kaufman, 1990).

4.4 Analysing the Statistics and Characteristics of Clusters

4.4.1 The Weekday Morning Peak (7-9am)

Next, detailed studies on each cluster is performed. From *Figure 9* Cluster #1 contains only the Bukit Banjang LRT stations and the largest Cluster #2 contains two categories of stations: (1) work destinations (the CBD, commercial and industrial centers) and (2) small MRT/LRT stations connecting major population centres to work areas. In addition, the distance lengths connecting work destinations are much shorter compared to the rest of the distances. We can make the following characterization:

- Work-related stations and small LRT stations having relatively low number of commuters starting their trips from are grouped together. Cluster #2 with short distance lengths indicates that these stations are very similar in terms of traffic contribution to the destinations.
- Overall, these clusters are reflecting the real-life situation where there are consistently few people begin their journey from these stations to other stations during the journey to work window: from LRT serving the lightly populated areas, from small MRTs connecting large residential areas to workplaces, from new and relatively less populous residential areas, and especially from the workplaces themselves being the destinations for commuters, not sources of commuters.

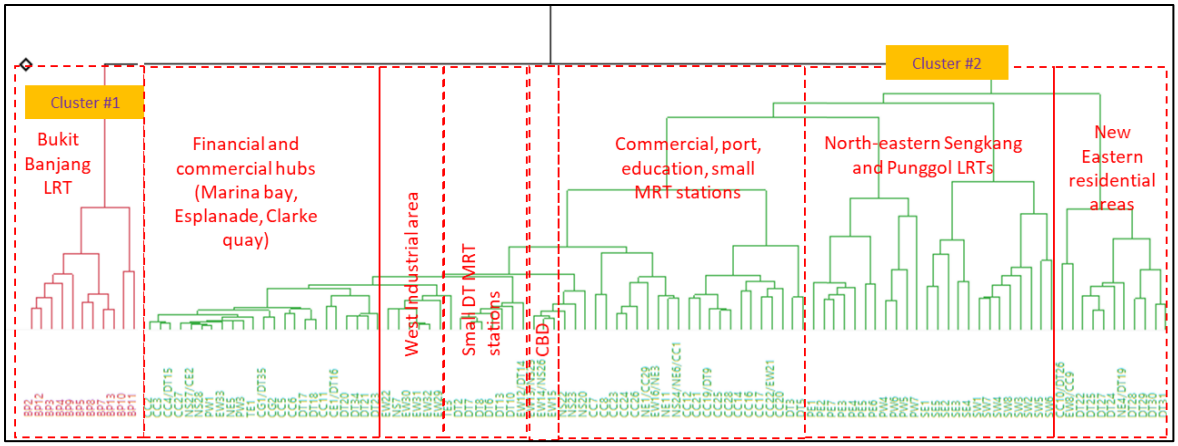


Figure 9: Cluster #1, 2 for Weekday Morning Peak

On the other end of the spectrum, as seen in *Figure 10* it is observed that residential areas are clustered together (Cluster #3, 4, 8, 9). Additionally, the cluster distances are much further. We can deduce the following characterizations:

- Residential areas have similarities in their contributions to other destinations, being the sources of commuters in the weekday morning peak.
- The longer distances before clusters are formed indicate the greater differences between these clusters as they are joined to form larger clusters, as compared to the work locations discussed previously. This illustrates the real-life situation because the number of commuters starting their journey to work in these stations vary more significantly being housing estates, compared to the work locations, which are generally uniformly empty in the morning peak, hence having relatively little differences.
- There are a few clusters with small size (Cluster #5-7) as the differences between them and the rest of the stations are too significant to merge (long inter-cluster distance) unless we reduce the number of clusters to a very small number and risk over-generalizing the rest of the stations. Specifically, most of them are interchange stations between a heavier traffic train line and a lighter traffic line (MRT-LRT interchanges, DT MRT line – EW/NS MRT line).

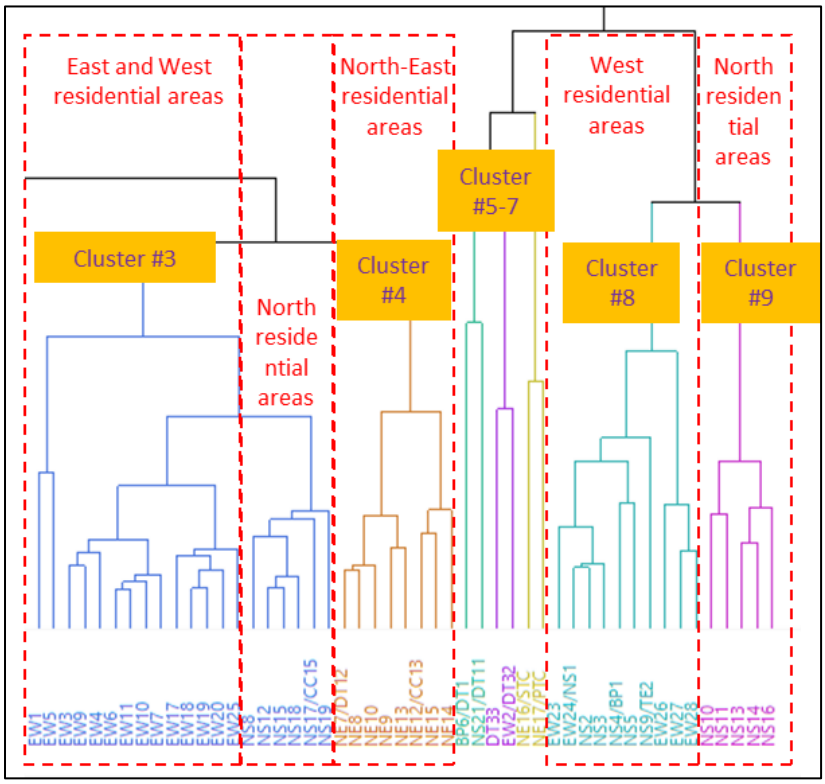


Figure 10: Cluster #3-9 for Weekday Morning Peak

4.4.2 The Weekday Evening Peak (5-7pm)

From the first cluster (*Figure 11*) a relatively reverse trend from the morning peak is observed. Apart from small LRT/MRT stations are again grouped together due to their low contribution to inflow traffic being both non-population and non-workplace locations, some residential areas in more remote locations such as the far East and North-West are now clustered with these low population areas. This indicates that they have similarly low contributions as the sources of commuters. This is reasonable because commuters at this time of the day do not generally start their journeys from these stations, but usually come to these locations to end the working day.

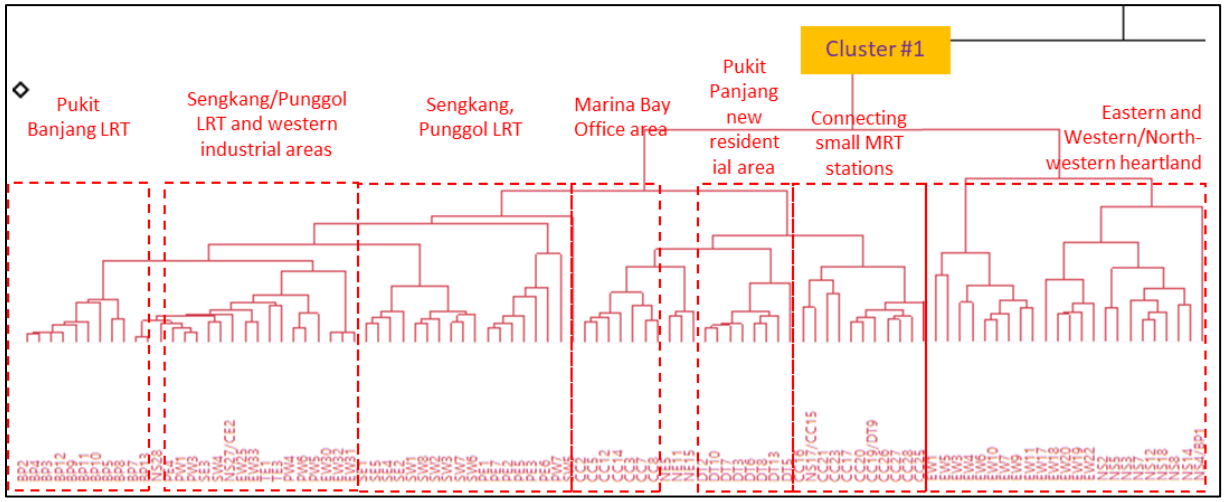


Figure 11: Cluster #1 for Weekday Evening Peak

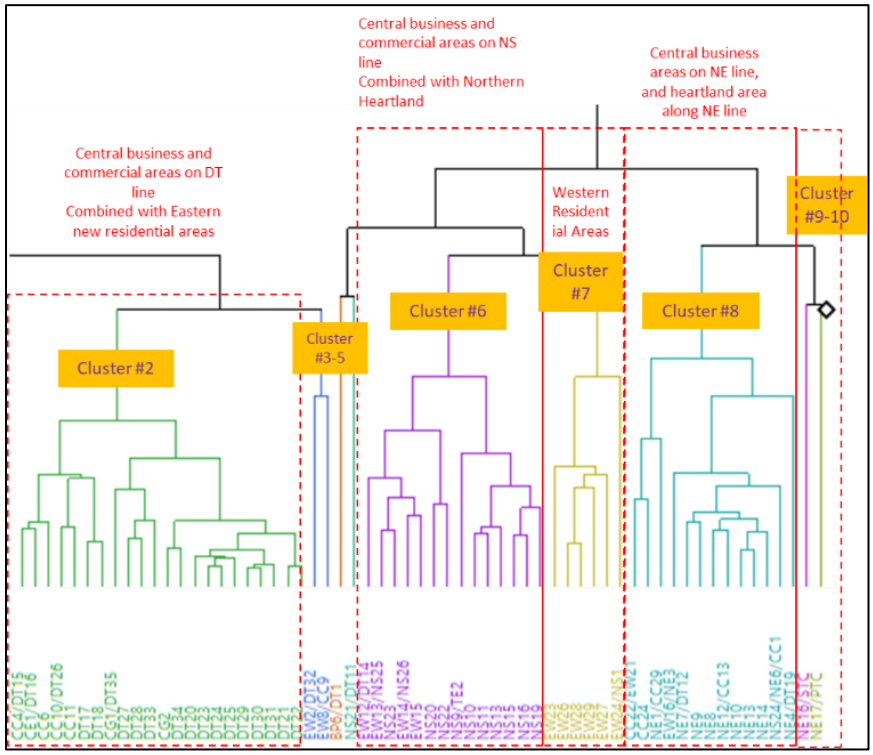


Figure 12: Cluster #2-10 for Weekday Evening Peak

However, more interesting interactions are discovered at the other few clusters (#2, #6, #7, #8 in *Figure 12*): Not only central business and commercial stations show high traffic inflow contribution as commuters leave the workplaces, the housing estates immediately connected to them along the North-South, North-East and East-West lines also demonstrate similar contribution and hence grouped. Commuters seem to display strong departing tendency from both 2 groups, one for travelling back home from work, one could be travelling to other commercial areas for evening activities.

This can be explained as followed: Since workplaces, after-work leisure locations and these residential areas are physically connected within close proximity, strong interactions are expected and strong similarity in departures are observed. This is different from the Eastern, Northern and Western heartland areas discussed previously. These residential areas are rather isolated from central commercial areas, and hence do not display strong contribution to the traffic to other commercial areas in the evening.
Similar to the morning peak window, there are a few clusters with small size (Cluster #3-5, 9, 10) with significant dissimilarities. Again, most of them are interchange stations between a heavier traffic train line and a lighter traffic line (MRT-LRT interchanges, DT MRT line – EW/NS MRT line).

4.4.3 The Weekend Leisure Peak (11am – 1pm)

From a high level, the clusters resemble largely to that of the weekday peak hours. From *Figure 13*, work locations (offices, ports, business parks, industrial areas), non/low-residential areas and small LRTs are clustered together in general (Cluster #1), as they are weak contributors to traffic to other destinations from the origin perspective, whereas main residential areas are clustered together (Cluster #2, 3, 4, 11, 12 in) due to their high contribution of departing passengers (*Figure 14*).
However, there is a noticeable difference when it comes to how commercial areas are clustered. They do not get cluttered with other work destinations, but with some residential areas as seen in cluster #1, 2, 3, 4. On this aspect, they are similar to that of the weekday evening peak.
The high interactions between the East, North-East and North-South residential areas, and central commercial areas can be understood as they are within close distance, enabling residents to commute easily from one station to another, generating high similarity levels.
For more secluded residential areas like the far West and North-West (Cluster #11, 12), they are neatly grouped separately without much interaction to the central commercial areas – the first small clusters are formed by stations on the same line with relatively proximity, before bigger clusters with another line are formed. This shows that stations of the same line and geographical locations have very similar traffic contribution.
In addition, the presence of small size clusters occurs (Cluster #8-10), similarly to the previous 2 time windows. Most of them are interchange stations between a heavier traffic train line and a lighter traffic line (MRT-LRT interchanges, DT MRT line – EW/NS MRT line).

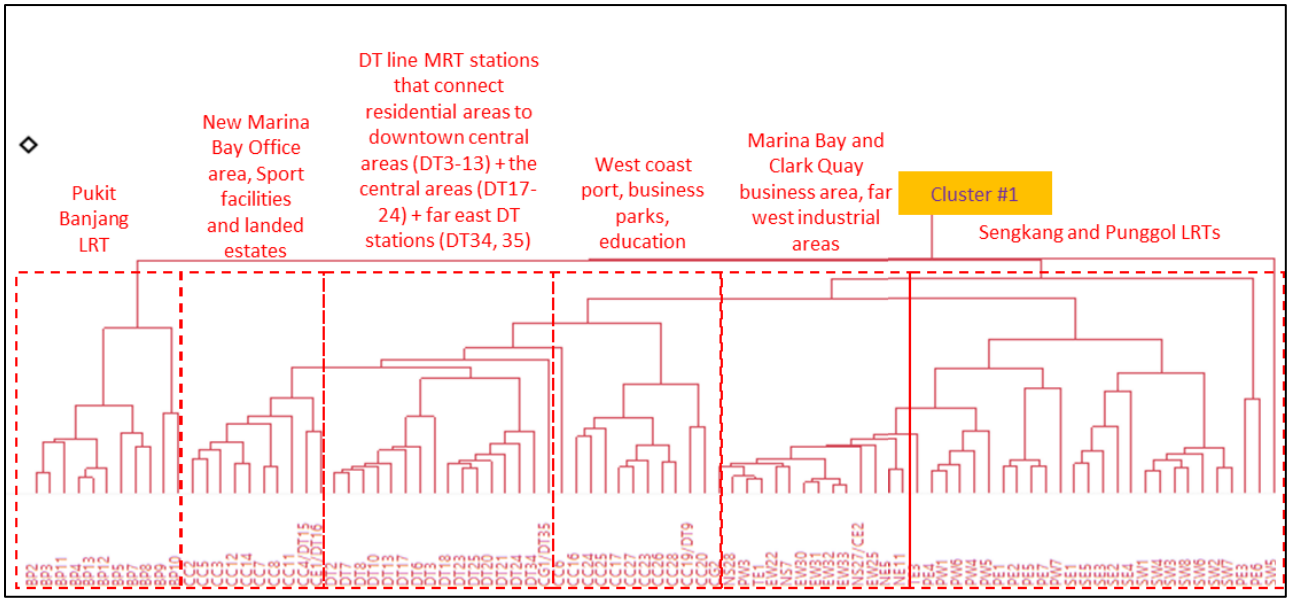


Figure 14: Cluster #1 for Weekend Leisure Peak

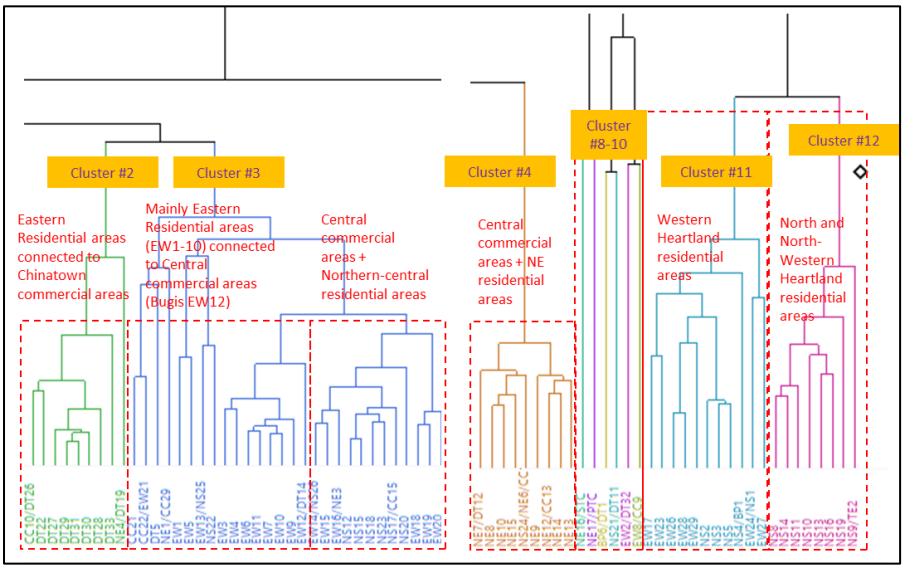


Figure 13: Cluster #2-12 for Weekend Leisure Peak

5 RECOMMENDATIONS

From the preceding analysis, the following summary can be drawn.

- For the weekdays:
  - West and North-West heartland areas display very clear distinct patterns: heavy inflow contributions in the morning, light in the evening peak.
  - Central business and commercial areas display light inflow contributions in the morning as they are not the origins for traffic but destination, and heavy in the evening as they become the origins of traffic to other stations.
  - North-East and North-South residential areas, due to closer proximity to the central areas, show heavy traffic inflow in both morning (journey to work) and evening (possibly for after work activities). It can be deduced that residents in these well-connected areas demonstrate more continuous engagement to the train service throughout the day for work and play. This is not observed in the West, and North-West residential areas, potentially due to their further distance from the central commercial regions.
- For the weekends:
  - Central commercial areas show strong interactions with the North-East, East and North residential areas, indicating similar traffic contributions.
  - Far West and North-West residential areas only demonstrate similarities within their own stations and hence form their own clusters.
- Interchanges of heavy traffic lines and lighter traffic lines form small clusters, implying very different traffic contribution patterns.

Based on the above summary, the following are recommended:

- From the congestion prevention point of view, it is reasonable to predict that the central commercial stations are prone to more overcrowding issues on the weekday evening peak and weekend leisure time. They display the strong interactions with other nearby residential areas. Hence operation data on these stations should be monitored to ensure the existing system is able to cope with the heavy traffic load.
- From the community interaction point of view, it was observed that the far-West and North-West regions display distinct work-leisure pattern without excessive interaction with other regions. Hence, there could be more operation capacity on these stations but at the same time, residents may encounter difficulties when travelling to more central areas either for work (labour mobility) or leisure. Hence more data on living satisfaction, career opportunity and labour flows can be collected to better develop suitable policies in these areas and compensate for their lack of accessibility to the central regions (Mitchell & Watts, 2010).
- Further studies on the characteristics of the interchanges between heavy traffic and light traffic lines should be conducted to obtain better understanding of the underlying reasons for these relatively distinct interactions behaviours from these stations.

Bibliography

Jain, A. K. (1999). Data Clustering: A Review. *ACM Computing Surveys*, 264–323.

Leonard Kaufman, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, Inc.

Mitchell, W., & Watts, M. (2010). Identifying Functional Regions in Australia Using Hierarchical Aggregation Techniques. *Geographical research, Vol.48 (1)*, 24-41.

Nowotny, B. A. (2003). lassificationof traffic data time series by cluster analysis, artificial neural networksand ANOVA. *10th world congress on Intelligent Transport Systems and Services*.