Master of Applied Information System

Final Report

Knowledge Discovery and Analytics in Time-Series Big Data: A Case Study on Singapore Public Train Commuter Travel Patterns

Version 2.4

<01 Dec 2013>

Version History

Version	Change Description	Author	Date
1.0	Initial Draft Final Report: • Draft of report structure • Draft of Literature Review for Automated Data Collection System (ADCS) • Draft of Data Preparation	RL	09.08.2013

2.0	Revision made with feedback from Prof Kam:
	• Confirmation of the project title
	• Add in the literature review for time-series mining
	• Revise and re-align literature review for smart card
	• Add methods and process of transforming spatial/temporal data into analytical format – transforming transaction data into OD matrix
	• Add challenges faced for data preparation into:
	  ○ Dealing with big data
	  ○ Timestamp to interval transofmation
		RL	29.08.2013

2.1	Revision made with additional inputs from Prof Kam: • Add in an overview of Singapore transport system	RL	07.10.2013
2.2	Revision made with inclusion of analysis	RL	27.10.2013
2.3	Revision made with review comments from Prof Kam.	RL	07.11.2013
2.4	Revision made with inputs from final presentation	RL	01.12.2013

ABSTRACT

The advances of automated data collection technologies and the rapid reduction of their costs provide businesses and organizations opportunities to amass a huge and continuously increasing amount of data about the consumption and lifestyle behaviors of their customers. The use of these data, however, tends to confine to simple tabular or dashboard reports. There is a general lack of analysis to optimize the return of investment on the collecting and managing of these data. This is particularly true when businesses and organizations seek to analyze the complex and voluminous time-series data that they have collected. The explosive growth of temporal related databases has far outpaced the analyst ability to interpret these data using conventional time-series statistical techniques, creating an urgent need for new techniques to support the analyst in transforming the data into actionable information and knowledge. In order to overcome this problem, this research study explores and discusses the potential use of time-series data mining, a relatively new framework by integrating conventional time-series analysis and data mining techniques, to discover actionable insights and knowledge from temporal transaction data. A case study on the Singapore public train transit will also be used to demonstrate the time-series data mining framework and methodology.

TABLE OF CONTENTS

1. INTRODUCTION 6
1.1. Overview 6
1.2. Motivation and Objectives 7
1.3. Content of the report 7
2. LITERATURE REVIEW 9
2.1. Conventional Data Mining 9
2.2. Time-Series Data Mining 10
3. DATA & METHODOLOGY 14
3.1. Case Study 14
3.1.1. Urban Public Transportation Study 14
3.1.2. Smart Cards and Automated Data Collection Systems (ADCS) 15
3.1.3. Singapore Public Transport System 16
3.1.4. Singapore MRT Rail Service 17
3.1.5. The Challenges for Public Transport 19
3.2. EZ-Link Smart Card 21
3.2.1. EZ-Link Smart Card: From Passive Data Collector to Smarting Informer 21
3.3. Data Preparation 22
3.3.1. EZ-Link Smart Card Transaction Data 22
3.3.2. From Transaction Data to Temporal O-D Matrix 23
3.4. Analytical Process and Methods 24
4. ANALYSIS & DISCUSSION 30
4.1. General Statistics on Station Passenger Volume 30
4.2. Time-Series Data Plot 33
4.3. Time-Series Data Cluster Analysis 35
4.3.1. Cluster Dendrogram 35
4.3.2. Cluster A: Strong Morning Peak/ Moderate Evening Peak – Residential Area 37
4.3.3. Cluster B: Strong Morning Peak – Residential Area 39
4.3.4. Cluster C: Strong Evening Peak – Industrial/ Commercial Area 41
4.3.5. Cluster D: Moderate Morning Peak – Residential Area 43
4.3.6. Cluster E: Moderate Morning/ Peak Strong Evening Peak – Industrial/ Commercial/ Residential Area 45
4.3.7. Cluster F: Strong Evening Peak – Industrial/ Commercial/ Retail Area 47
4.3.8. Cluster G: Gentle Evening Peak – Retail Area 49
4.3.9. Cluster H: Weekend Peak – Special Weekend Activities Area 51

4.3.10. Cluster I: Strong Morning Peak/ Moderate Evening Peak – Industrial/ Commercial/ Residential Area 53
4.3.11. Cluster J: Strong Morning Peak/ Strong Evening Peak – Industrial/ Commercial/ Residential Area 55
4.3.12. Cluster K: Seasonal Peak –Special Activies Area 57
4.4. Summary of Analysis and Insights for Urban Transport Planners 58
4.5. Research Limitations and Future Works 58
5. CONCLUSION 60
6. REFERENCE 61

1. INTRODUCTION

1.1. Overview

The advances of automated data collection technologies and the rapid reduction of their costs provide businesses and organizations opportunities to amass a huge and continuously increasing amount of data about the consumption and lifestyle behaviors of their customers. The use of these data, however, tends to confine to simple tabular or dashboard reports. There is a general lack of analysis to optimize the return of investment on the collecting and managing of these data. This is particularly true when businesses and organizations seek to analyze the complex and voluminous time-series data that they have collected. The explosive growth of temporal related databases has far outpaced the analyst ability to interpret these data using conventional time-series statistical techniques, creating an urgent need for new techniques to support the analyst in transforming the data into actionable information and knowledge. In order to overcome this problem, this practical research study explores and discusses the potential use of time-series data mining, a relatively new framework by integrating conventional time-series analysis and data mining techniques, to discover actionable insights and knowledge from temporal transaction data. A case study on the Singapore public train transit will be used to demonstrate the time-series data mining framework and methodology.

Urban public transport planners today continue to face challenges in understanding the public transport commuters' travel behaviors as the conventional transport survey, which is still the main tool used for transport studies, remains resource intensive and expensive. Furthermore most of the insights gained from these transport surveys were obsoleted by the time the planners leveraged on it for urban public transport planning. To overcome these challenges, the case study

in this paper explored and demonstrated the use of sensing data such as smart card transport transaction data and time-series data mining to help urban public transport planners to better understand the Singapore public transport commuters' travel behaviors.

1.2. Motivation and Objectives

This practical research study aims to explore and demonstrate the effective use of time-series data mining in analysing complex temporal big data. This research study has specifically used the Singapore public train transportation as a case study for the industrial application of time-series data mining to discover insights on Singapore commuters' travel behaviours. The high level goals of this research include:

I. To introduce to time-series data mining as a new framework to overcome the limitations of conventional time-series statistical techniques when analysing temporal big data

II. To discuss in details the existing time-series data mining techniques and methodologies employed by researchers and the industry.

III. To demonstrate time-series data mining techniques and methodologies in urban public transport planning context.

1.3. Content of the report

This paper starts with an overview of the motivation and challenges faced by modern data analyst especially with relation to large and temporal database. This is followed by a review of relevant literatures of conventional time-series analysis, data mining, time-series data mining, smart card and automated data collection system (ADCS). Next, the Singapore public train transit case study will be introduced. This includes the EZ Link system, the Singapore version of

smart card system. A detail discussion on the implementation of the time-series data mining framework using the EZ Link data will be covered. This is followed by a comprehensive discussion on the analysis results obtained and interesting insights gained from the analysis results. Last but not least, the extensibility of the framework to other businesses and organization and future research direction will also be discussed.

2. LITERATURE REVIEW

The literature review segment of this report is divided into two parts. The first part of this segment introduces the conventional data mining techniques and their limitations when applied on time-series data. The second part introduces the time-series data mining techniques, in particularly, the Dynamic Time Warping technique is discussed in detail. The challenges of time-series data mining will also be discussed in this segment.

2.1. Conventional Data Mining

Knowledge Discovery in Database (KDD) refers to the process of discovering insights and knowledge from a collection of data and data mining is the heart and analytical step of K process [24]. In the data mining step, a variety of data analysis tools were used to discover useful, non-obvious and previously unknown patterns or data trends [23]. It is important to note that the employment of different data mining methods and techniques could generate different type of insights and knowledge for users. Thus, analysts will need to be conversant with the different types of data mining techniques in order to adopt them well for their research and industry context.

Conventionally, there were mainly three types of data mining techniques: association rules, classification and statistical. Association rules were typically applied in a transactional database, which it takes the form $X \rightarrow Y$, where X and Y are sets of items appearing in transactions. For example, the rule $\{tomato, lettuce\} \rightarrow \{salad\ dressing\}$, found in sales data of a supermarket indicated that customers who bought tomato and lettuce had a certain likelihood of buying salad dressing. Data mining in a transactional database involves discovering of all such association rules. The main limitation with this data mining technique is it does not consider the sequence

and order of items either within transaction or across transaction. As such, this data mining technique is not suitable to be used on time-series data.

Another data mining technique is classification, which is a method that involves generating a set of rules for classifying instances into predefined classes from a set of trial data and then predicts the classes of new instances according to the generated rules. It is important to note that in clustering, the classes are not predefined or known at the point of learning, but is part of the learning task in the clustering to define and identify the classes. It is possible to apply classification technique on time-series data. However it is not without challenge; special techniques will have to employed on the time-series data to make sure that each instance of the time-series are of equal length for the application of classification techniques.

The last and perhaps most common conventional data mining technique used by analyst is statistical data mining. Statistic is a useful tool to analyze and make inferences about the data to discover useful patterns from a dataset. However it required sophisticatedly high level of statistical skills to construct statistical data mining models to analyze time-series data.

Although the conventional data mining techniques are useful in many industrial settings, they are not suitable for performing data mining on time-series data. As such, new techniques will have to be developed to cater for such types of data.

2.2. Time-Series Data Mining

Time-series data refers to data collected in a routine, continuous and sequential manner. This type of data, which typically accompanied with a timestamp, has always been collected by

businesses and organizations in their daily operations. Examples of such data include, sales transaction, delivery orders, stock quote prices etc. Increasingly, businesses and organizations seek to analyze these time-series data to uncover more business insights. However, the analysis of time-series data posted new challenges, as traditional data mining and statistical techniques are inappropriate when analyzing data that contain a time factor. These challenges eventually became the motivation for the development of time-series data mining techniques.

According to Schubert and Lee [22], there are two main challenges in analyzing time-stamped data. Firstly it is a tedious process to transform time-stamped data into table formats, which is suitable for the application of traditional statistical techniques. Secondly it is challenging to apply pattern detection on time-stamped data using traditional data mining as the time-stamped data might be irregularly recorded. Time-series data mining however, tried to overcome these challenges by presenting a framework methodology that converts time-stamped data into time-series format suitable for analysis and pattern detection.

One of the most wisely studied time-series data mining technique is the dynamic time warping (DTW) technique proposed by Berndt and Clifford [18]. As mentioned previously, one of the common problems in analyzing time-series data is that time-stamped data might be irregularly recorded, which might cause two different time-series that have common trends to occur at different time. If the two time-series did not occur simultaneously, the application of traditional data mining techniques will not discern such a relationship, as it does not consider time as a factor in comparison.

Fig 2a Euclidean Distance

Fig 2b Dynamic Time Warping

Take for example, Fig 2a, a traditional data mining similarity measure such as Euclidean distance is used to compare the similarity between two time series Q and C, and a relationship is not discern because it is the two time-series are out of phase. In Fig 2b, dynamic time warping technique is used to overcome this problem by accounting for the time factor when comparing the two time-series.

The DTW technique first constructs an *n-by-m* matrix *M* for two time-series *P* and *Q* (a.k.a input and target sequence). The ($i^{th,}$ j^{th}) element of the matrix *M* contains the Euclidean distance $d(q_i, p_j) = (q_i - p_j)^2$ between the two points q_i and p_j. A warping path, $W = \{W_1, W_2, \ldots, W_k\}$, which is a contiguous set of element in Matrix *M*, is then determined. The warping path is also bounded by the following constraints:

(1) **Boundary conditions**: $W_1 = (1,1)$ *and* $W_k = (n, m)$, This will require the warping path to start and finish diagonally.

(2) **Continuity**: Given $W_k = (a, b)$ and $W_{k-1} = (a', b')$, $a - a' \leq 1$ and $b - b' \leq 1$, This will require the next allowable step to be adjacent cell.

(3) **Monotonicity**: $a - a' \geq 0$ and $b - b' \geq 0$. This will force the points in W to be monotonically spaced in time.

A large number of warping paths would have fulfilled the above constraints but only the warping path with the minimized cost will be of interest. A common implementation would be the minimization of distance between adjacent elements; it can be calculated with the simple method:

$$DTW(Q,P) = min_w[\sum_{k=1}^{K} d(w_k)]$$

Using the DTW technique, similarity measures between several time sequences can be generated to form a similarity matrix. For example, given K time sequences, a $(K \times K)$ symmetric matrix can be constructed whose ij^{th} element contains the similarity measure between i^{th} and j^{th} sequences. This type of similarity matrix is commonly used in time-series clustering.

There were already time-series data mining case studies done on business and organizations. However most case studies were done in the retail business domain. In a recent case study, Nakkeeran et al [16] demonstrated how time-series data mining techniques can be used to perform clustering of retail store-level revenue over time and how profiling of such clusters generates greater business insights. Hebert [17], in his master capstone project, performed a similar case study using a different set of retail data. Although there is an increase interest for organization to explore time-series data mining, little studies were done to apply time-series data mining in transportation domain. This research thus hope to explore and apply time-series data mining techniques to investigate the passenger travel behaviors within a transportation network, and in this research case study, the Singapore public train transportation network.

3. DATA & METHODOLOGY

3.1. Case Study

3.1.1. Urban Public Transportation Study

The public transport service plays a vital role in any country; it facilitates the shifting of people between different points in space, which ultimately helps to support the economy and livelihood of the people [1]. Furthermore, an effective and efficient public transport system would help to reduce the high usage of private transport such as cars and ease off congestion on the roads. It is therefore imperative for urban transport planners to conduct careful studies and research to understand the urban public transport network and public commuters' travel behaviors so as support their decision making processes.

Traditional transportation studies typically involve the conduct of household transportation survey as a mean to collect data. After the data are collected, analyst will apply conventional statistical analysis to draw insights on the commuter usage of the transportation system. The traditional transportation studies, however, has its own challenges and limitations. Firstly the conduct of household transportation survey is a long and labor-intensive process where the data collected might have obsoleted before the conclusion of the transportation study. Secondly the data collected were aggregated; while the aggregated data are good for the application of statistical analysis, they are limited in fully describing the individual commuter's travel patterns. Lastly conventional statistical analysis also has its limitation in analyzing complex data with a temporal element.

These limitations and challenges has motivated this case study where this research explored and demonstrated the use of sensing data, such as smart card transit transaction data, and time-series data mining techniques to generate insights on the Singapore public train commuters' travel behaviors.

3.1.2. Smart Cards and Automated Data Collection Systems (ADCS)

Smart cards are portable credit card size devices that store and process data [21]. As the smart cards are very portable and durable, businesses had explored many innovative ways to leverage on this technology for their business operations. Currently, the smart cards are mainly used in applications involving identification, authorization and payment.

In a literature review on smart card data in public transit, Pelletier et al [21] highlighted that in recent years, smart cards and automated data collection systems (ADCS) had also been widely adopted in public transport networks around the world. The main concept of this smart card application involved enabling the public transit commuters to make their fare payment using smart cards that were credited with monetary value.

Although the preliminary purpose of smart cards is for collection of fares for public transport, Bagchi and White [9] discussed the potential use of the passively recorded transaction data for travel behavior analysis and public transport planning. Since then, various researches and case studies were done on the data collected from public transport smart card systems around the world. Morency et al [10] explored the data mining techniques used to analyze the spatial and temporal variability of Canadian pubic transit network passengers using different card types. Asakura et al [11], through the use of smart card data collected from Japan's public train network,

studied the change in passengers' travel patterns when the train operator changed its train timetable. Kim and Kang [12], also attempted to use the transaction data collected from T-Money, South Korea's electronic fare card system, to develop an origin-destination (O-D) matrix for the Seoul intermodal public transit network. From the O-D matrix, transport planners will gain a better understanding of passenger flow between stations at different temporal segments.

In contrast to the traditional transportation study surveys, the passive collection of data through the use of smart cards and ADCS is a more cost-effective alternative. As the smart card transit transaction data are sequential and temporal, time-series data mining techniques can be applied on the data to reveal insights and discover knowledge on the passenger travel behaviors.

3.1.3. Singapore Public Transport System

There are three main modes of public transportation in Singapore: Taxi, bus and mass rapid transit (MRT) rail network. Singapore also adopted a hub-and-spoke integrated public transport system as its public transportation strategy [5]. In this system, the bus services will serve the transport within a town to the hub, and the MRT rail services will be used for longer distance transport between hubs. The below Figure 3a shows a pie chart on the breakdown in pubic transport mode market shares in 2008 [7].

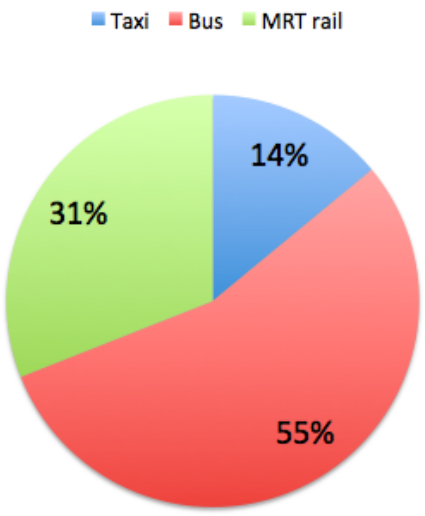


Figure 3a, Singapore Public Transport Mode Market Share in 2008

Although the market shares for bus higher, there was a steady increase in demand for MRT rail service. Accordingly to the household transport interview survey conducted from 1997 till 2008 [7], the demand for MRT rail service had increased from only 19% in 1997 to 31% in 2008. Furthermore, between the two modes of public transport, the survey found that the MRT rail service was a better alternative mode that could compete with the car on speed for long urban trips. As such, the MRT rail network is an important mode of public transport where the Singapore government would continue to invest in it to achieve the goals set in the 2008 transport master plan to "make public transport a choice mode"[5].

3.1.4. Singapore MRT Rail Service

The Mass Rapid Transit (MRT) rail system was built in mid-1980s with its first segment opened in 1987. Since then, a number of expansion works were done on the MRT network. Currently the MRT network has 102 stations covering an estimated 149 km [6]. The below Figure 3b shows a map of the Singapore MRT Rail network.

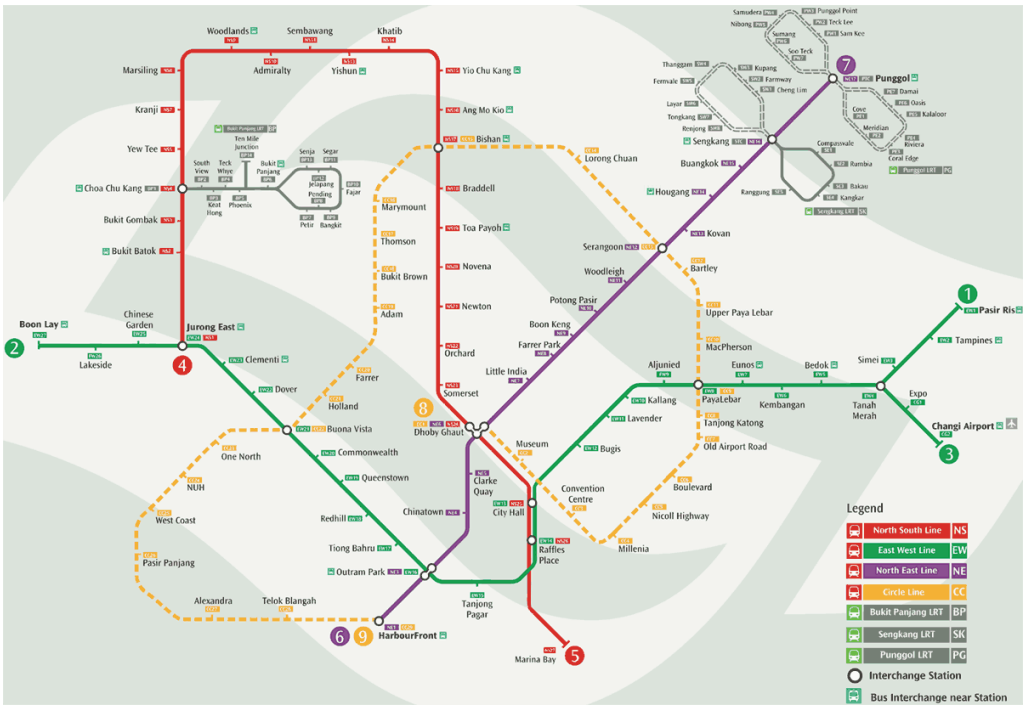


Figure 3b, MRT Rail Service Network Map (Current)

There are currently four main operating service lines in the MRT rail service network. A number of expansion works were also in progress where new train service lines and train stations will be built to increase its coverage to 278 km by 2020. The below Table 3a shows the current operating service lines and the expansion estimated to be completed b 2020.

Service Line	No. of Stations	Length (km)	Extension Works
Operational			
North South Line (NSL)	26	45	Marina South Pier Station, estimated to complete in 2015
East West Line (EWL)	35	57.2	Tuas West extension, estimated to complete in 2016
North East Line (NEL)	16	20	N.A.
Circle Line	30	35.7	N.A.
Under Construction			
Downtown Line	34	42	Estimated to complete in 2017
Thomson Line	22	30	Estimated to complete in 2021

Table 3a, Service Lines of MRT Rail Network

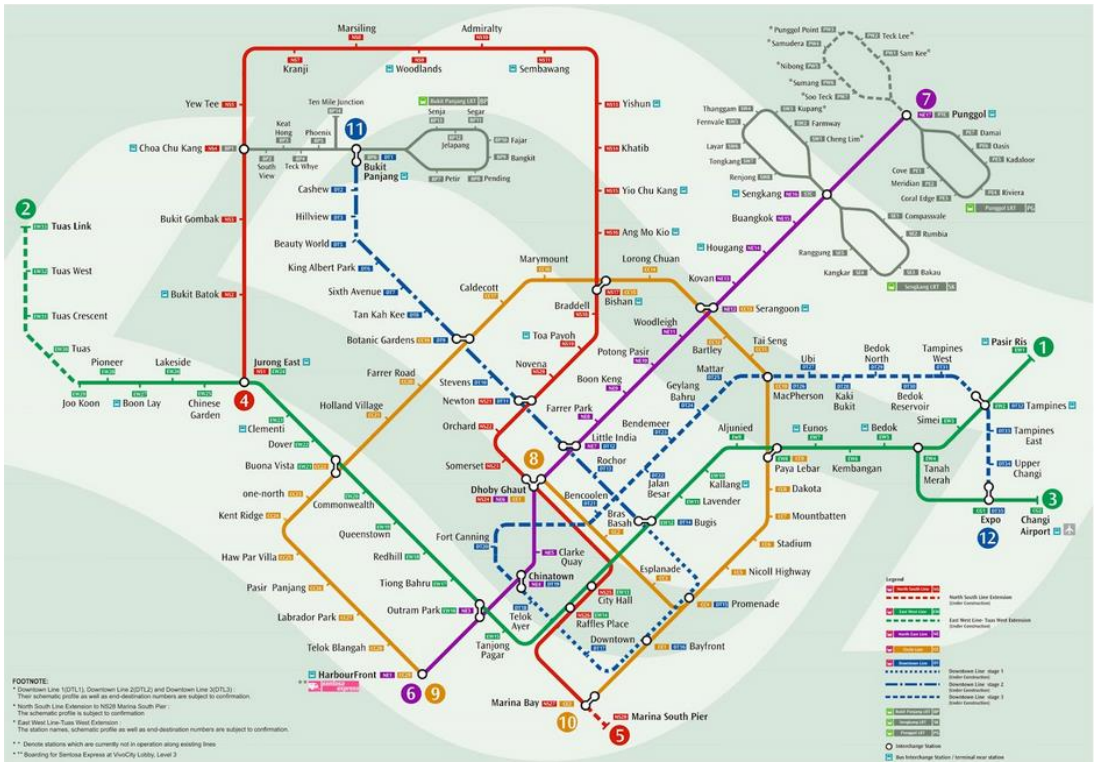


Figure 3c, MRT Rail Service Network Map (in 2020)

With the number of expansion works and increasing of service lines, it is important for Singapore urban transport planners to have a better understanding of the local public train commuters travel behaviors in order structure better policies and construct expansions that serve the commuters better.

3.1.5. The Challenges for Public Transport

Although the Singapore government had placed great emphasis and efforts on the public transport, there are still challenges in realizing its goal in 2008 transport master plan. Accordingly to the Singapore's Household Interview Travel Survey from 1997 to 2008, the public transport's share of total daily trips had dropped from 63% in 1997 to 58% in 2004, and falling even further to 56% in 2008 [7]. Furthermore, the series of MRT service breakdowns in

2011 and 2012 has decreased the public confidence on the promise of MRT being a good alternative to cars [8]. Below Table 3b shows the series of MRT service breakdowns between 2011 and 2012.

S/N	Date	Service Line Affected	Remarks
1	17 Oct 2011	North South Line (NSL)	Train disrupted between 10.30AM to 11.10AM between Ang Mo Kio and Bishan station
2	15 Dec 2011	North South Line (NSL)	Train disrupted between 6.50PM to 12.15AM between Bishan and Marina Bay station
3	17 Dec 2011	North South Line (NSL	Train disrupted between 7.55AM to .18PM between Toa Payoh and Marina Bay station
4	18 Dec 2011	North South Line (NSL) and East West Line (EWL)	Postponed opening hours for NSL and EWL
5	15 Mar 2012	North East Line (NEL)	Train disrupted between 6.30AM to 4.35PM between Harbourfront and Dhoby Ghaut station
6	17 Aug 2012	North East Line (NEL)	Train disrupted whole day between Harbourfront and Dhoby Ghaut station
7	10 Jan 2013	North East Line (NEL)	Train disrupted between 9.50AM to 4.35PM between Harbourfront and Dhoby Ghaut station

Table 3b, MRT Service Disruptions

In order to better understand the problem of decreasing public transport mode share, this research aims to collect the train transport trip data and analyze the travel pattern behaviors of its passengers. This result of this analysis would help policy makers and transport service operators to better understand the Singapore public transport commuters and thereafter structure policies and plans that could better serve the commuters better and ultimately increase the public transport mode share.

3.2.EZ-Link Smart Card

Implemented in 2001, the EZ-Link card is a contactless smart card used mainly for the payment of public transportation fares in Singapore. In 2009, the new CEPAS EZ-Link, which allows the smart card to be used in a wide variety of payment applications, replaced the original EZ-Link smart card. The replacement had also enabled users to make limited small payments in certain retail stores, for example in Singapore branches of McDonald's fast food outlets. There were also exploratory projects where EZ-Link planned to work with NETS to create a new hybrid card that integrates the function of EZ-Link and CashCard. In 2007, Starhub, one of Singapore telecommunication companies, had also embarked on a six-month trial project to explore the possibility of integrating EZ-Link functions into phones. The EZ-Link cards are sold, distributed and managed by EZ-Link Pte. Ltd, a subsidiary of Singapore's Land Transport Authority.

3.2.1. EZ-Link Smart Card: From Passive Data Collector to Smarting Informer

In spite of the great upbeat and proliferation of smart cards and ADCS, there were few studies done on the smart cad data collected from public transport network in Singapore. Lee et al [13] did a case study to optimize serviceability and reliability of bus routes by analyzing the data collected from EZ-Link, Singapore's smart card system. Sun et al [14] also did a study using the EZ-Link data to estimate the spatial-temporal density passenger onboard a train or waiting in the train station. Another group of researchers, Soh et al [15], used EZ-link data to perform a weighted complex network analysis of Singapore's public transport network. Although these studies had demonstrated some applications of the smart card data collected from the Singapore public transport network, no extensive study had been done thus far to reveal the travel patterns of the Singapore public transport passengers.

3.3. Data Preparation

Although the EZ-Link smart card transaction data contain rich information on the passenger travel pattern, it is not formatted to perform any meaningful analysis. Thus there is a need to prepare and transform the raw smart card transaction data into proper data formats for further analysis. The below Figure 3c shows the data preparation process where raw EZ-Link smart card transaction data are transformed into time-series data plots for our analysis.

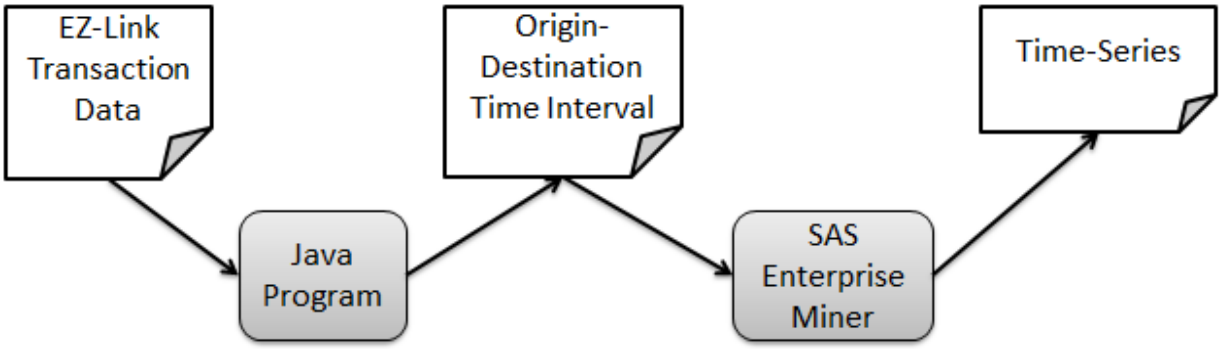


Figure 3c, Data Transformation Process Overview

3.3.1. EZ-Link Smart Card Transaction Data

For this study, we were able to obtain one month (November 2011) worth of EZ-Link smart card transaction data from the Singapore Land Transport Authority (LTA). An estimated total of 60 million train trip transactions were made in the month of November 2011 within the Singapore public train network, which consist of 102 MRT train stations and 21 LRT train station. Each trip transaction consists of quite a number of data columns, which describe the trip. However for the purpose of this study, we are only interested in the following data columns: the origin station, destination station and passenger entry timestamp into the origin station. While the time factor of the entry timestamp for each trip transaction remains critical for our analysis, the absolute time value was not "analytical friendly" for performing time-series data mining. As such, the

transaction data with entry timestamp will need to be transformed into origin-destination (OD) time interval format for time-series data mining.

3.3.2. From Transaction Data to Temporal O-D Matrix

A Java application is written to perform the data transformation. The EZ-Link smart card transaction data was first loaded and stored in the PostgreSQL relational database, taking up as much as 130GB worth of storage space. The Java application will then query the database and output the result in the OD time interval format. As the EZ-Link smart card transaction data is large, a B^+ tree index was built on the origin and destination station ID columns to reduce the query time.

In the OD time interval format, transaction data are aggregated to 15 minutes time interval per day. The "Origin ID" and "Destination ID" columns refer the station IDs of the origin and destination station respectively. The "Day of Month" column captures the day of the month for that given record (for example, if the date is 13 Nov 2011, the day of the month would be 13). The "Time Interval" column captures the time interval of the record; it will be a 15 minutes time interval starting from 0600H to 2345H. The "Passenger" column captures the number of passengers that is traveling from the origin to the destination in that particular day of the month and time interval. Figure 3d shows the data transformation from transaction data into the origin-destination (OD) time interval format.

Transaction Data

Origin ID	Destination ID	Entry Time	…
14	25	2011-11-13 15:12:44	…
67	12	2011-11-13 15:15:44	…

OD Time Interval

Origin ID	Destination ID	Day of Month	Passenger	Time Interval
14	25	13	56	0600
14	25	13	21	0615
14	25	13	12	0630
		. . .		
14	25	30	35	2345

Figure 3d, Data Transformation from Transaction Data to O-D Time Interval

The output of the transformation is saved as CSV file for the performance of time-series data mining using SAS Enterprise Miner.

3.4. Analytical Process and Methods

The SAS Enterprise Miner is an analytical software application that streamlines data mining processes and allows users to perform predictive and descriptive analytics on large volumes of data. The application has interactive visualization functions, which allows users to perform data exploration and discovery.

Leonard et al [19] in their paper, *An Introduction to Similarity Analysis using SAS*, described in detail how DTW technique is implemented in SAS Enterprise Miner. In another similar work, Leonard and Wolfe [20] had also explained how the DTW technique could be used in SAS Enterprise Miner for mining transactional and time-series data. For the purpose of this research,

SAS Enterprise Miner will also be used as the tool to perform time-series mining and clustering on the smart card data to investigate the travel pattern of Singapore public train commuters.

The SAS Enterprise Miner is used as data transformation and analytical tool for this research. The OD time interval data generated by the Java application is imported into the SAS Enterprise Miner where to convert it into time-series data plots. Figure 3e shows SAS Enterprise Miner work process.

Figure 3e, SAS Enterprise Miner Time-Series Data Mining Work Process

There are four nodes in the SAS Enterprise Miner time-series data mining work process:

- File Import – The File Import node allows user to upload and convert external flat files, spreadsheets, and database table into format that SAS Enterprise Miner can recognize as a data source and use it in the subsequent data mining processes.

- TS Data Preparation – The TS Data Preparation (TSDP) node converts the input data into time-series data for analysis. A few settings were set for this research analysis. Firstly, the *timeseries* column, which contains the time interval in the input OD Time Interval format, was set to the role of *Time ID* in this analysis. The *Time ID* would form up the x-axis in the generated time-series data plots. The *passenger* column, which contains the frequency of number of passengers, was set to the role of *Target*. The *Target* would form up the y-axis in the generated time-series data plots. As we are interested to examine the passenger volume of each MRT train station, we will set the *Origin* column, which

contain the origin train station ID, as the cross-sectional variable, *Cross ID*. Refer to Figure 3f for the setting in TSDP node.

Name	Use	Role	Level
DOM	**Default**	Rejected	Interval
DOW	**Default**	Rejected	Interval
Destination	**Default**	Rejected	Nominal
Origin	**Default**	Cross ID	Nominal
Passenger	**Default**	Target	Interval
Timeseries	**Default**	Time ID	Interval

Figure 3f, Setting in TSDP node

The *Result* function of the TSDP node also allows enables the analyst to perform certain basic time-series data analysis. Figure 3g shows the display panel of the *Result* function of the TSDP node.

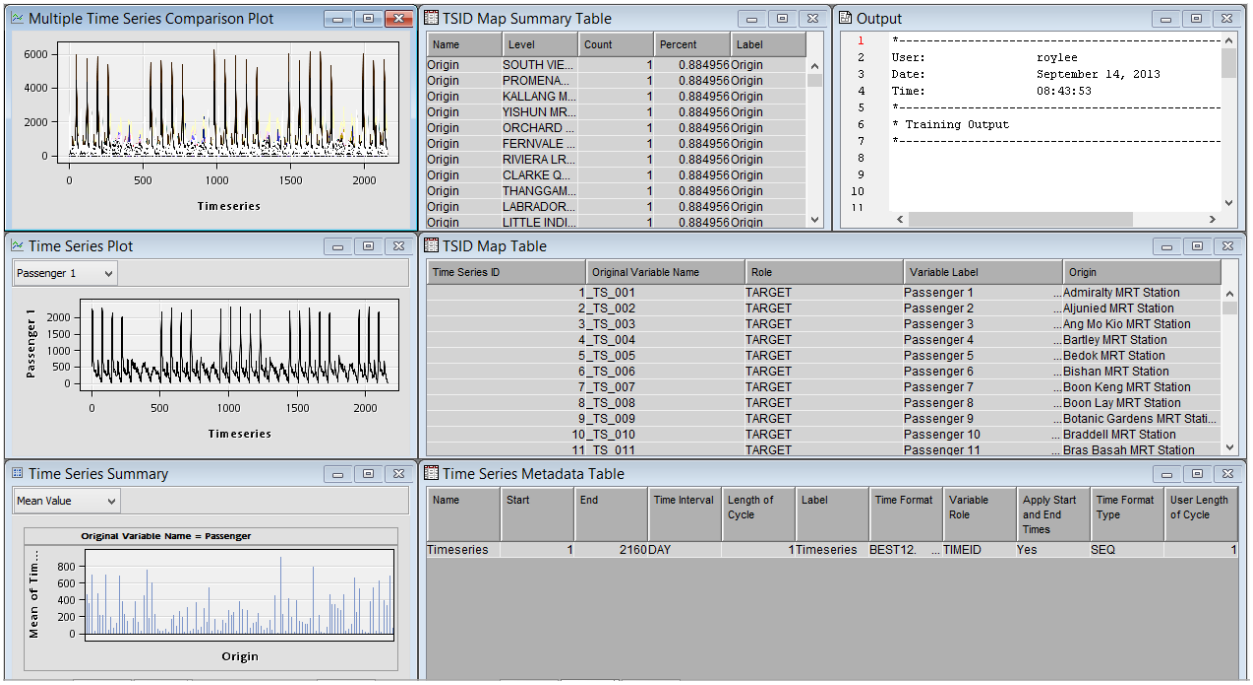


Figure 3g, Result panel of TSDP node

Some of the sub-functions displayed in the Result panel of TSDP include:

- *Time Series Summary* – Allow user to select and visualize the different statistical distribution of passenger flow of each train station. Examples of statistical distribution include MEAN, MAX, MIN and SUM of passenger flows in each train stations.
- *Time Series Plot* – Allow user to select and visualize time-series graph plot of a particular train station.
- *Multiple Time Series Comparison Plot* – Allow user to select and visualize multiple train stations' time-series graph plots.
- *TSID Map Summary Table* – Table showing the data of cross-section variable used in the analysis
- *TSID Map Table* – Table showing the mapping of time-series plot to the origin train station.

- Metadata – The Metadata node allow users to modify certain data attributes so that the data is suitably formatted for the next process node.

- TS Similarity – The TS Similarity (TSS) node performs the clustering and similarity analysis by comparing the time-series, classified and group time-series that exhibit similar characteristics over time. As the time series might have different lengths, dynamic time warping (DTW) method will be applied to compare two time-series; the input and target sequences. The TSS node also calculates the similarity measures between the compared input and target sequences. The *Result* function of the TSS node visualizes the results of the similarity and clustering analysis. Figure 3h shows the display panel of the *Result* function of the TSS node.

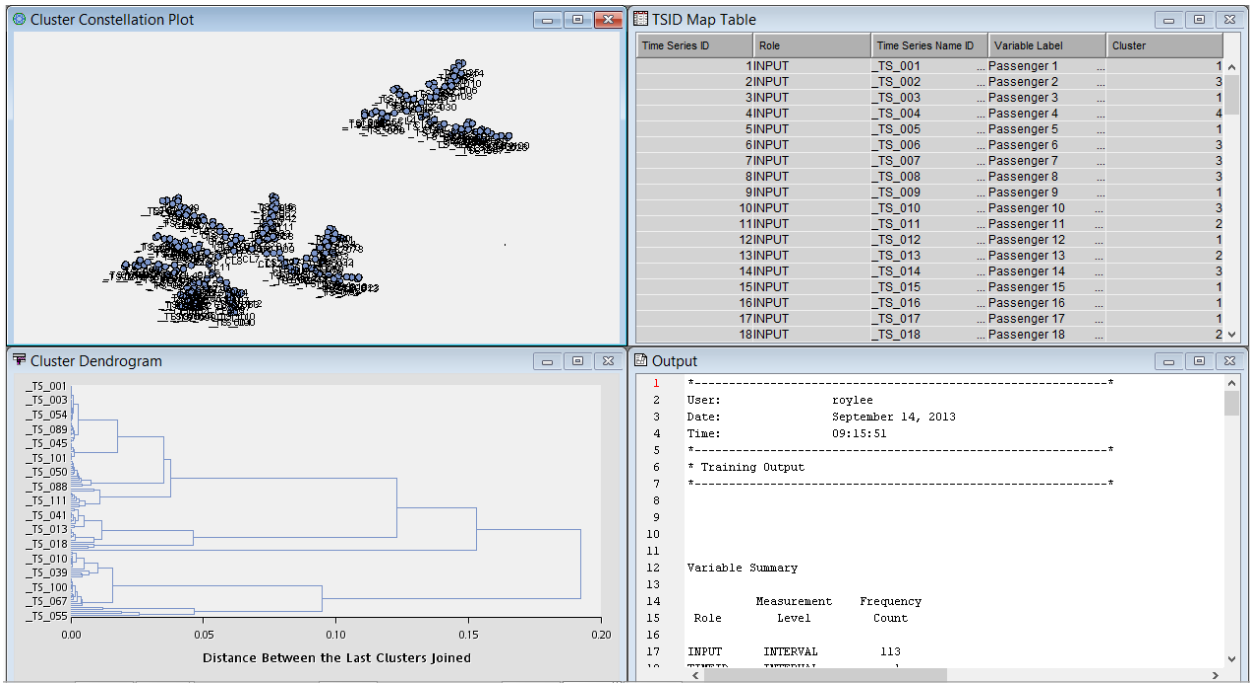


Figure 3h, Result panel of TSS node

Some of the visualizations displayed in the Result panel of TSS include:

- *Constellation Plot* – The constellation plot provides a visualization understand on similarity relationship between train stations time-series. The circles represent the input variables, which are the time-series of train station. The circles that are linked and grouped together show higher similarity relationship.
- *Dendrogram* – The dendrogram visualized the arrangement of clusters produced by hierarchical clustering. The train station time-series will join up with other time-series that are similar to them, and similar cluster will join up with other similar clusters.

Through the SAS time-series data mining work process, travel patterns of the passenger in each MRT stations are generated. For the purpose of this research, the modeling the passenger volume

in a station will be based on the passenger entry time into the station. The results and insights generated from TSDP and TSS node will be analyzed and discussed in the next section.

4. ANALYSIS & DISCUSSION

4.1. General Statistics on Station Passenger Volume

Some basic and general statistics and distributions on the time-series data can be generated using the *Time-Series Summary* module from the *Result panel* of TSDP node. Figure 4a shows the MAX, MIN, MEAN and SUM distributions of the train stations' time-series data sorted in decreasing order.

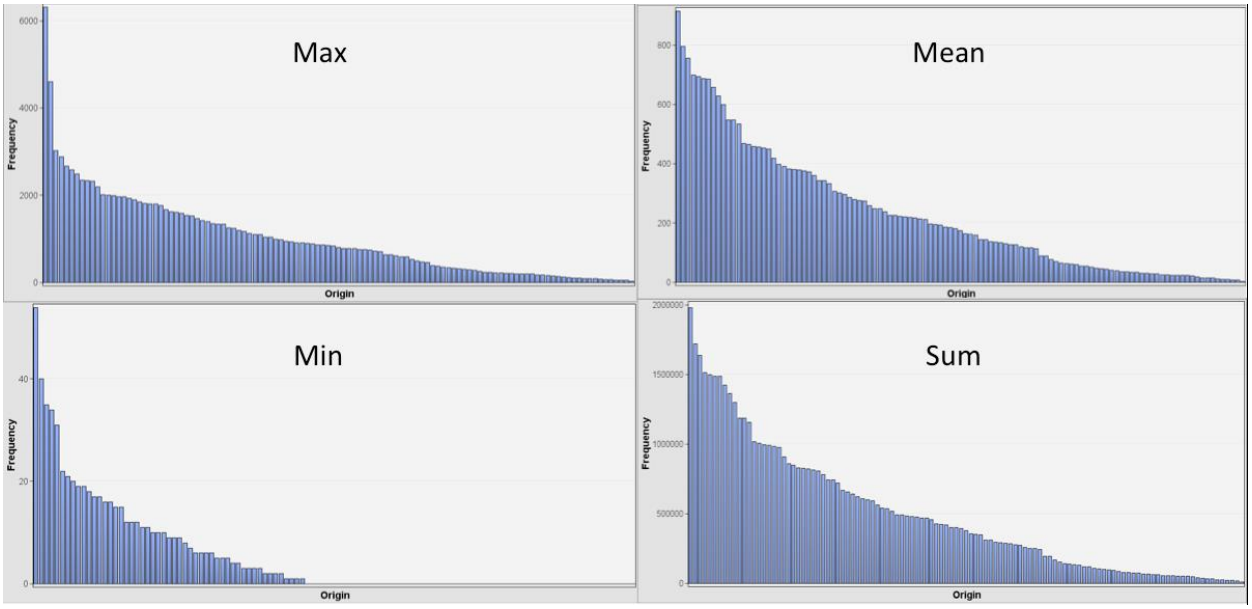


Figure 4a, MAX, MIN, MEAN and SUM distribution of time-series data sorted in decreasing order

The below Table 4a further shows the general statistics of top 5 stations in various distribution:

Rank	Station	Value (# of passengers
MEAN		
1	Orchard MRT Station	916
2	Raffles Place MRT Station	796
3	City Hall MRT Station	757
4	Ang Mo Kio MRT Station	700
5	Boon Lay MRT Station	694
MAX		
1	Raffles Place MRT Station	6322
2	Tajong Pagar MRT Station	4604
3	Yishun MRT Station	3019

4	Orchard MRT Station	2879
5	Tampines MRT Station	2664
MIN		
1	Ang Mo Kio MRT Station	54
2	Yishun MRT Station	40
3	Woodlands MRT Station	35
4	Bugis MRT Station	34
5	Clementi MRT Station	31
SUM		
1	Orchard MRT Station	1,979,034
2	Raffles Place MRT Station	1,720,209
3	City Hall MRT Station	1,636,709
4	Ang Mo Kio MRT Station	1,513,342
5	Boon Lay MRT Station	1,499,068

Table 4a, Top 5 train stations by various distributions

Note that MAX, MIN and MEAN distribution were measured by number of passengers entered a train stations in a time interval (15 minutes) while SUM distribution were measured by the total number of passengers in a given train stations for November 2011.

There are some interesting observations made from the various distributions. Firstly, noticed from the MAX distribution that there were two stations that have much higher passenger volume in a time interval, especially the top passenger volume train station, which has almost double the passenger volume of the third train station. Cross referencing this from the top 5 ranking of statins for the MAX distribution, the top passenger volume train station, Raffles Place MRT Station, has a passenger volume of 6322, which was indeed more than twice of the amount of passenger volume of the third train station, Yishun MRT Station. This suggests that the traffic peak period of certain stations would experience much higher passenger volume as compared to the peak period of other stations.

Another interesting observation is the comparison between MEAN and MAX distribution. As observed from the distributions, the highest mean passenger volume is significantly lower than the highest passenger volume in a time interval. This is also observed from top 5 ranking for both distributions; top passenger volume train nation for MAX distribution, Raffles Place MRT Station, has nearly 7 times higher passenger volume than highest mean passenger volume train station, Orchard MRT Station. This again suggest train stations may experience exponential increase in passenger volume during certain period of the day.

The ranking of stations for SUM and MEAN is the same with the same 5 stations topping both measures. However we are able to observe an interesting phenomenon where the MAX top 5 stations are not the same as the MEAN and SUM top 5 stations. The top stations for MEAN and SUM, Orchard MRT Station, is ranked 4th for MAX. Raffles Place MRT Station, which ranked 2nd for MEAN and SUM, ranked top for MAX. While there are 3 stations, namely Tajong Pagar MRT Station, Yishun MRT Station and Tampines MRT Station, which only appears in the top 5 in MAX but not in MEAN and SUM. This could suggest that while these 3 stations might not have a high average volume of passengers per interval, there are certain periods in the day where it experience large surge in passengers.

The review on the general statistics on the passenger volume of the MRT stations had highlighted the importance to analyze the travel patterns of passengers from a time-series perspective instead of simply looking at the average or total volume of passengers for each station. There could be more insights revealed as we analysis the passenger travel patterns in a time-series with shorter time interval.

4.2. Time-Series Data Plot

The *Multiple Time-Series Comparison Plot* module in the *Result panel* of TSDP node was used to visualize the time-series data for all train stations.

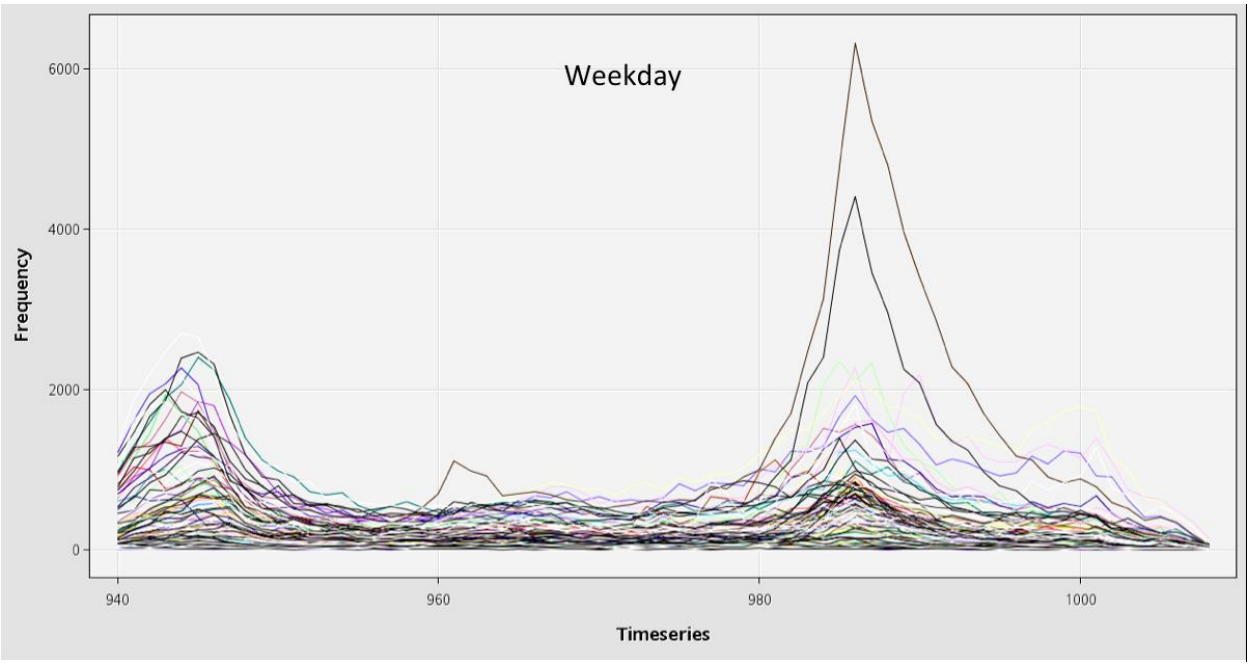


Figure 4b, Time-series plot for all train stations on weekday

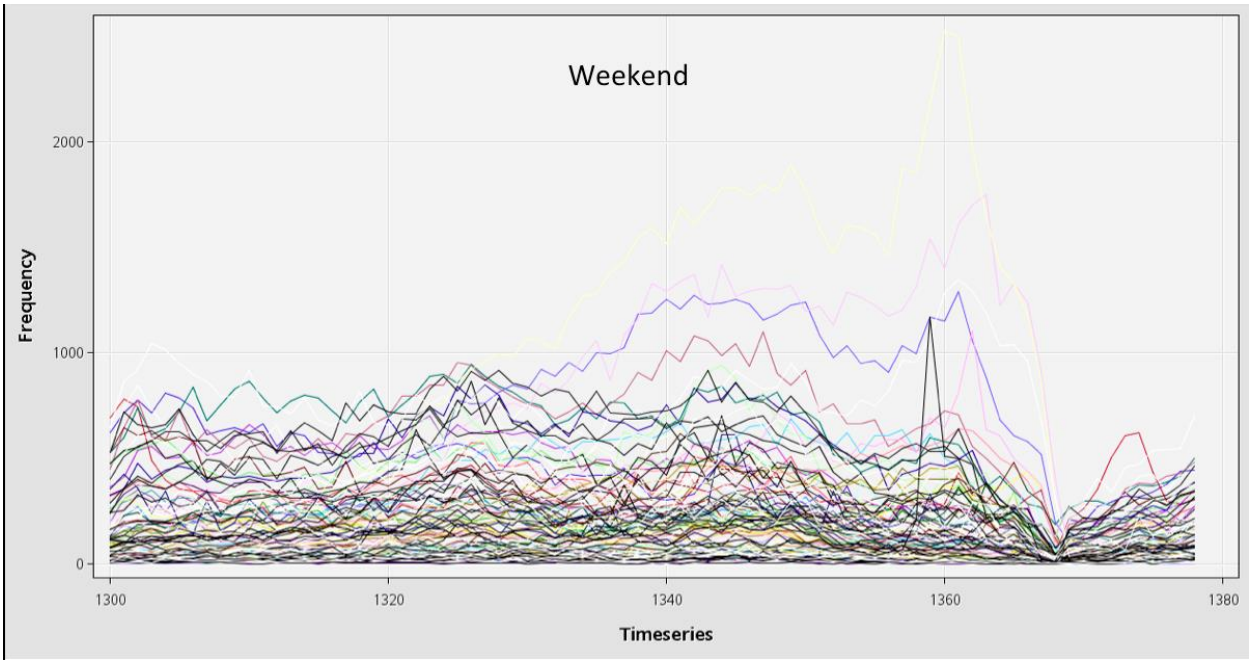


Figure 4c, Time-series plot for all train stations on weekend

The above Figure 4b and Figure 4c shows the time-series plot of all train stations on a weekday and weekend respectively. There are two interesting observations made from the above time-series data plot. Firstly, there are significant differences between the weekday and weekend's train stations time-series data plot; the weekday time-series data plot shows a clear morning and evening peaks in passenger volume while the weekend time-series data plot seems to show a more uniformly distributed passenger volume throughout the day.

Secondly, it is observed in both the time-series data plots that the different train stations does exhibit different time-series patterns. For example, while there are some train stations showing a morning and evening peak in its weekday time-series data plot, the same travel pattern cannot be observed in other stations. Thus, clustering and further analysis will need to be done to identify and classify these different travel patterns among the train stations.

4.3. Time-Series Data Cluster Analysis

4.3.1. Cluster Dendrogram

The TSS node performs the clustering and similarity analysis on the train station time-series data. Figure 4d shows the dendrogram of the time-series data plots generated in the *Result panel* of the TSS node.

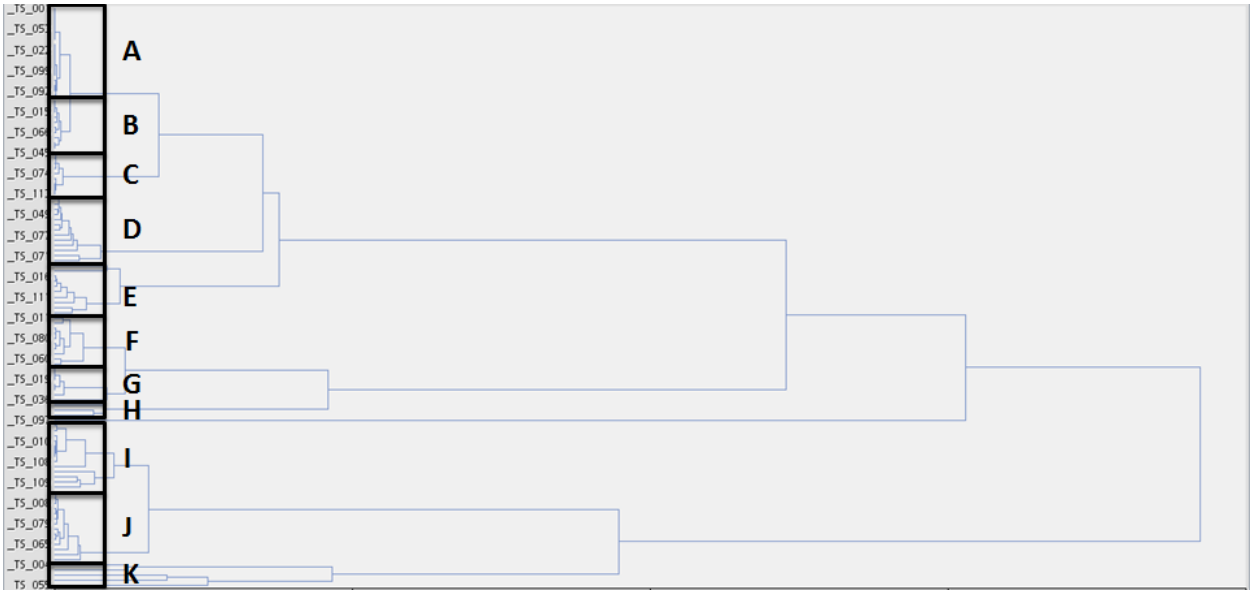


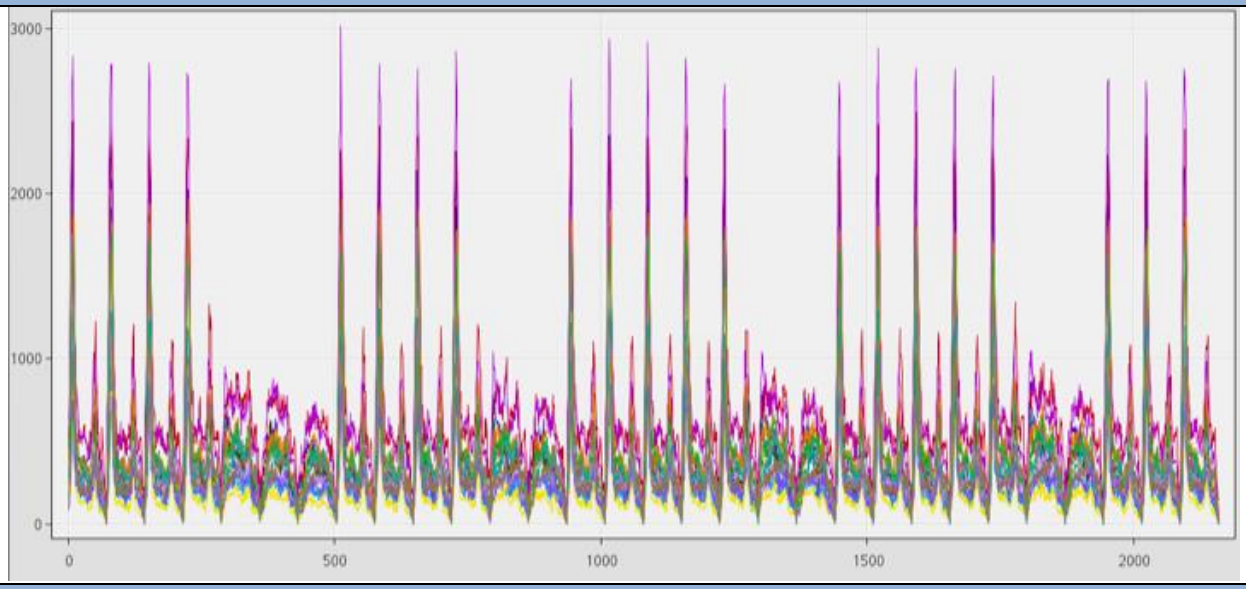
Figure 4d, dendrogram of time-series data clustering

The default number of clusters for TSS node similarity analysis is 5. However, upon examining the 5 clusters, the results were not satisfactory as there were still different distinct travel patterns within each cluster that could be further refined and classified. As such, a trial and error process was initiated to explore the optimal number of clusters need to be generated such as each cluster exhibit unique and interesting passenger travel patterns. After much trial and error on the hierarchical clustering, 11 different clusters (labeled A – k) were identified to be the optimal number of clusters for our analysis. Each of the clusters has exhibit unique and interesting

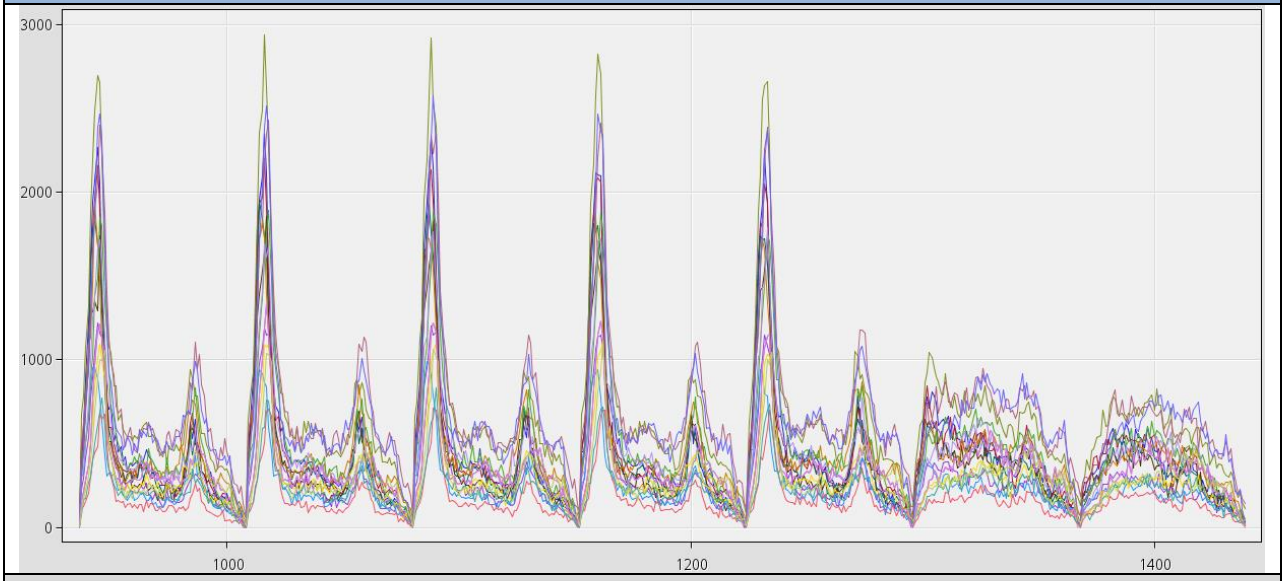
passenger travel patterns based on their time-series data plots. The 11 clusters will be analyzed and discussed in greater detail in the subsequence sections.

4.3.2. Cluster A: Strong Morning Peak/ Moderate Evening Peak – Residential Area

Month View



Week View



Cluster A – Admiralty MRT Station, Sembawang MRT Station, Yew Tew MRT Station, Lakeside MRT Station, Khatib MRT Station, Yishun MRT Station, Kovan MRT Station, Ang Mo Kio MRT Station, Choa Chu Kang MRT Station, Bedok MRT Station, Pasir Ris MRT Station, Simei MRT Station, Timpines MRT Station, Kembangan MRT Station, Sengkang LRT Station, Sengkang MRT Station, Serangoon MRT Station

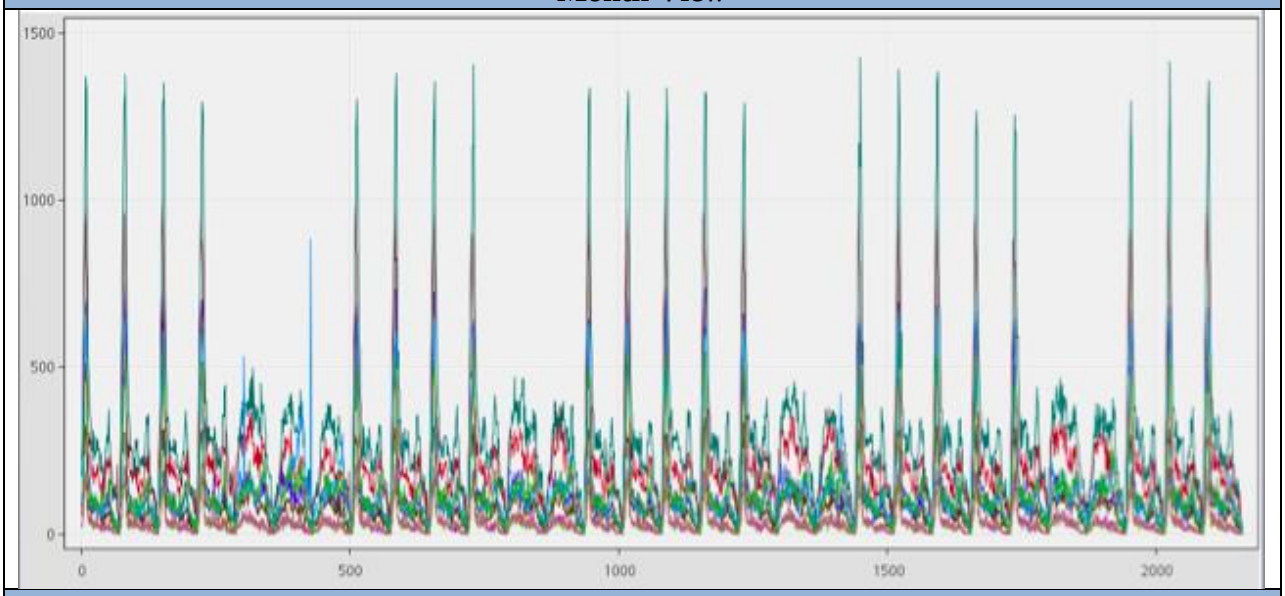
The time-series data plots in cluster A have displayed a strong morning peak and a relatively weaker evening peak on weekdays. This suggests the train stations in cluster A were experiencing high passenger volume entering the stations in the morning and relatively lesser

passenger volume in the evening. However the morning and evening peak patterns were not observed in weekend, where the stations received relatively constant passenger volume throughout the day.

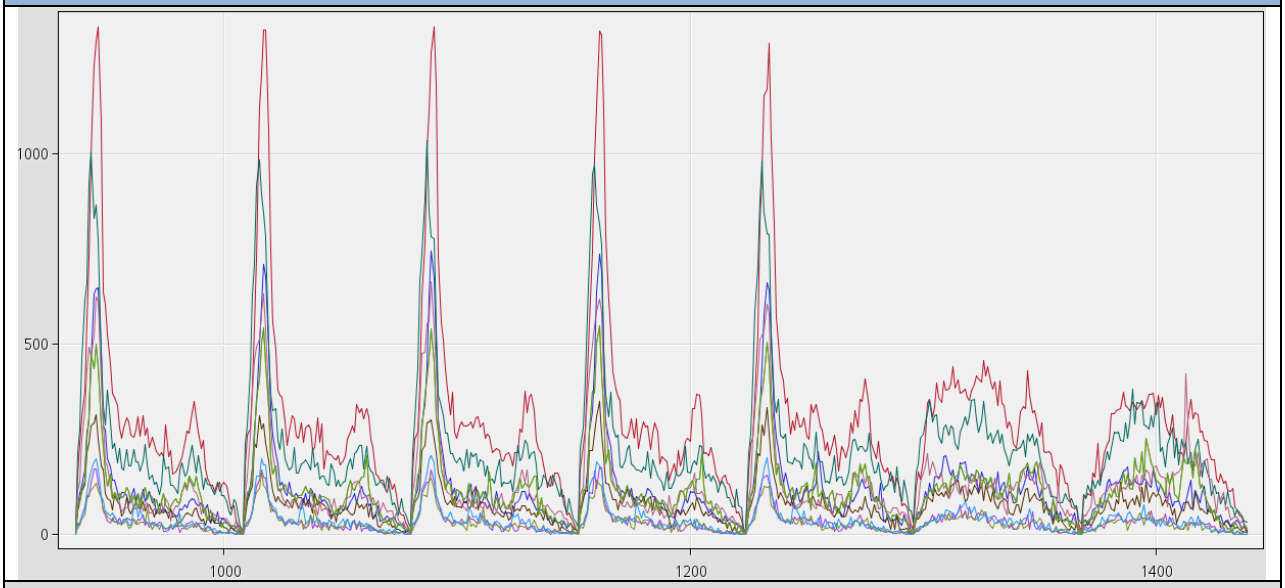
Examining into the composition of cluster A, we found that it is make up of MRT stations situated in residential areas. This could give us a preliminary explanation for the weekday morning peak where the passengers living in residential areas were traveling to work on weekday morning. As for the relatively lower weekday evening peak, a possible explanation could be that the passengers, whom had travelled to the schools or small offices located in the residential areas, were returning home from work.

4.3.3. Cluster B: Strong Morning Peak – Residential Area

Month View



Week View



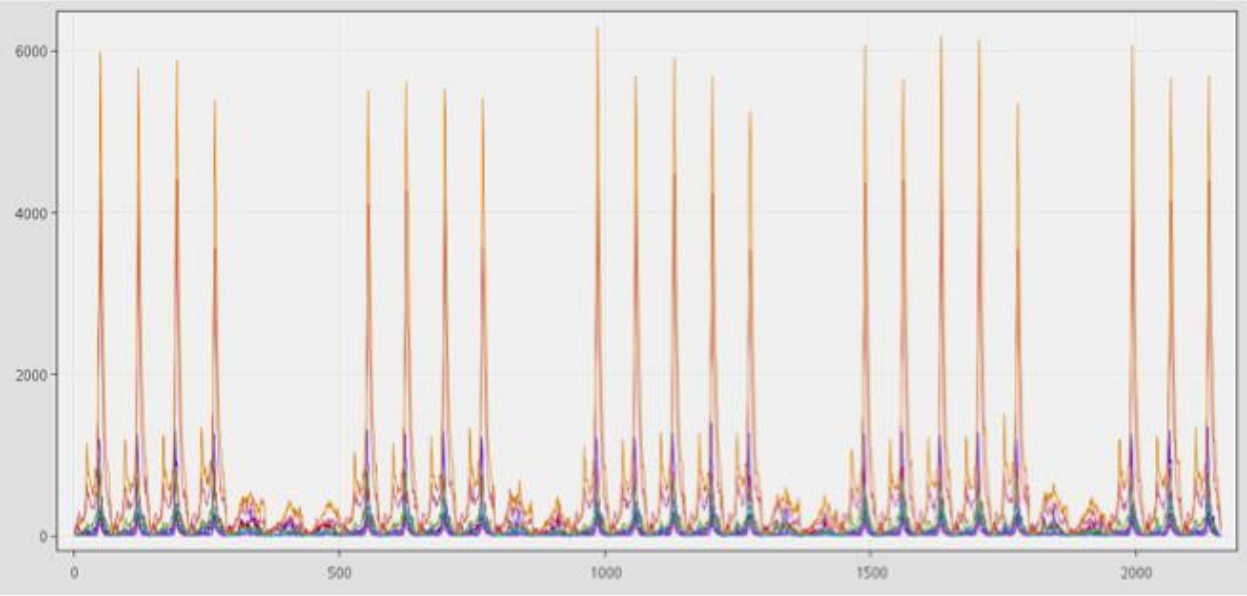
Cluster B – Buangkok MRT Station, Hougang MRT Station, Bukit Gombak MRT Station Rumbia LRT Station, Tongkang LRT Station, Cove LRT Station, Meridian LRT Station, Punggol LRT Station, Punggol MRT Station

The time-series data plots in cluster B have displayed a strong morning peak on weekdays. This suggests the train stations in cluster B were experiencing high passenger volume entering the stations in the morning. However the morning peak pattern was not observed in weekend, where the stations received relatively constant passenger volume throughout the day.

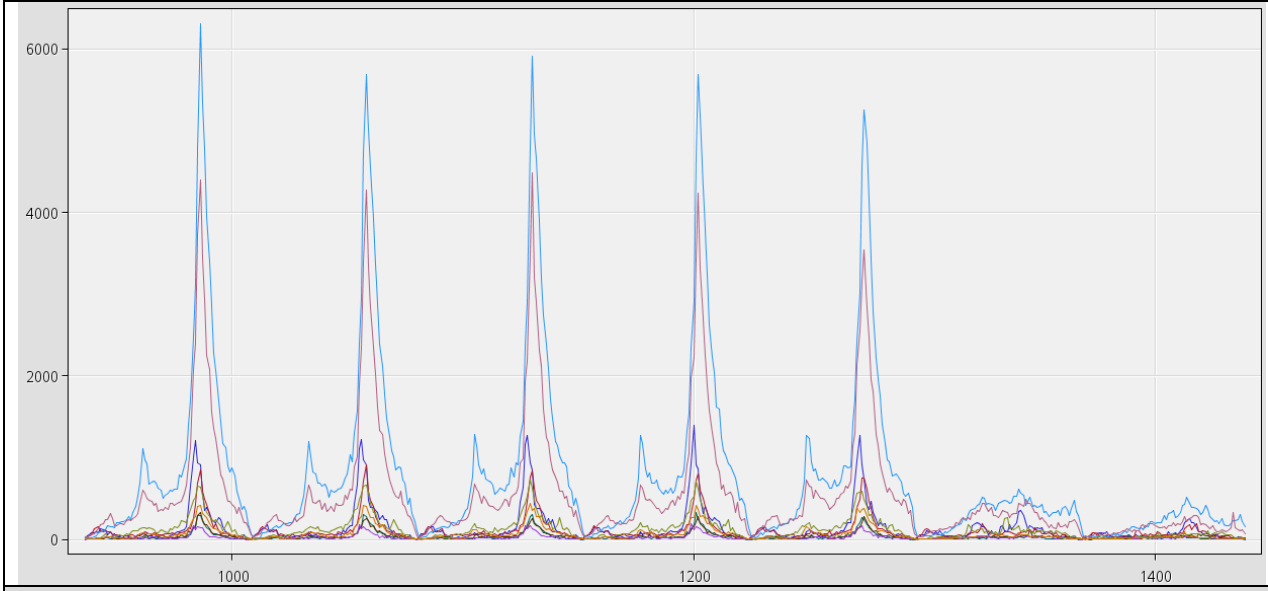
Examining into the composition of cluster B, we found that it is make up of LRT stations situated in residential areas. This could give us a preliminary explanation for the weekday morning peak where the passengers living in residential areas were traveling to work on weekday morning. Another interesting observation is the morning passenger volume of cluster B was lower than the morning passenger volume of cluster A. This might be due to the limited capacity of LRT as it has smaller carriages compare to MRT.

4.3.4. Cluster C: Strong Evening Peak – Industrial/ Commercial Area

Month View



Week View



Cluster C – Joo Koon MRT Station, Tai Seng MRT Station, Labrador Park MRT Station, Nicoll Highway MRT Station, Pasi Panjang MRT Station, Kent Ridge MRT Station, Raffles Place MRT Station, Tanjong Pagar MRT Station, One-North MRT Station

The time-series data plots in cluster C have displayed a very strong evening peak on weekdays.

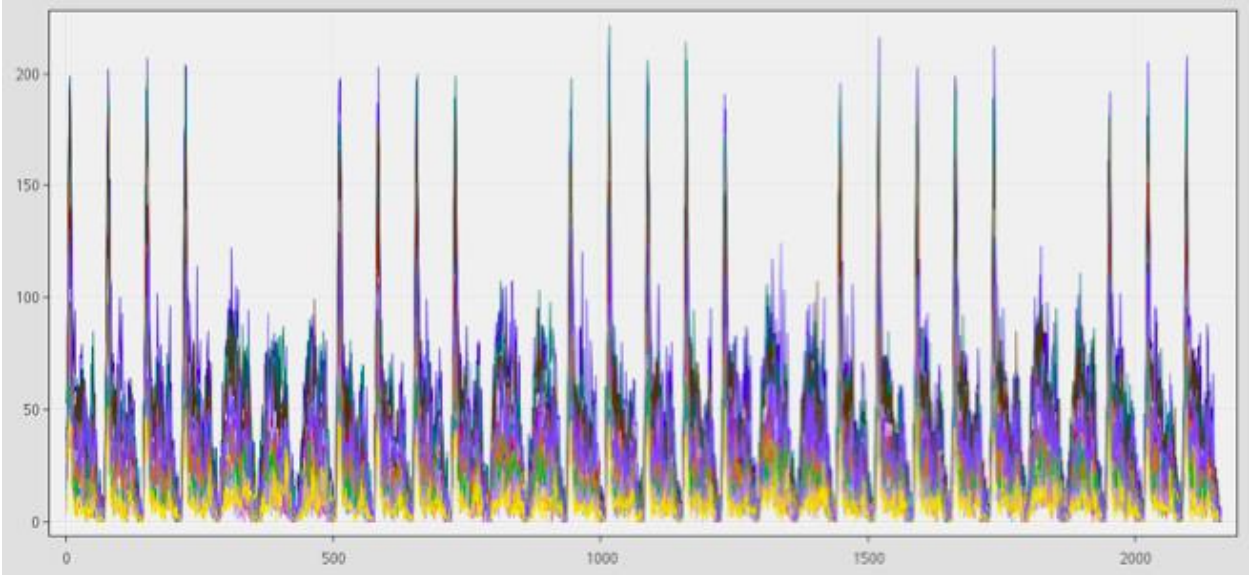
This suggests the train stations in cluster C were experiencing high passenger volume entering

the stations in the evening. However the evening peak pattern was not observed in weekend, where the stations received relatively constant passenger volume throughout the day.

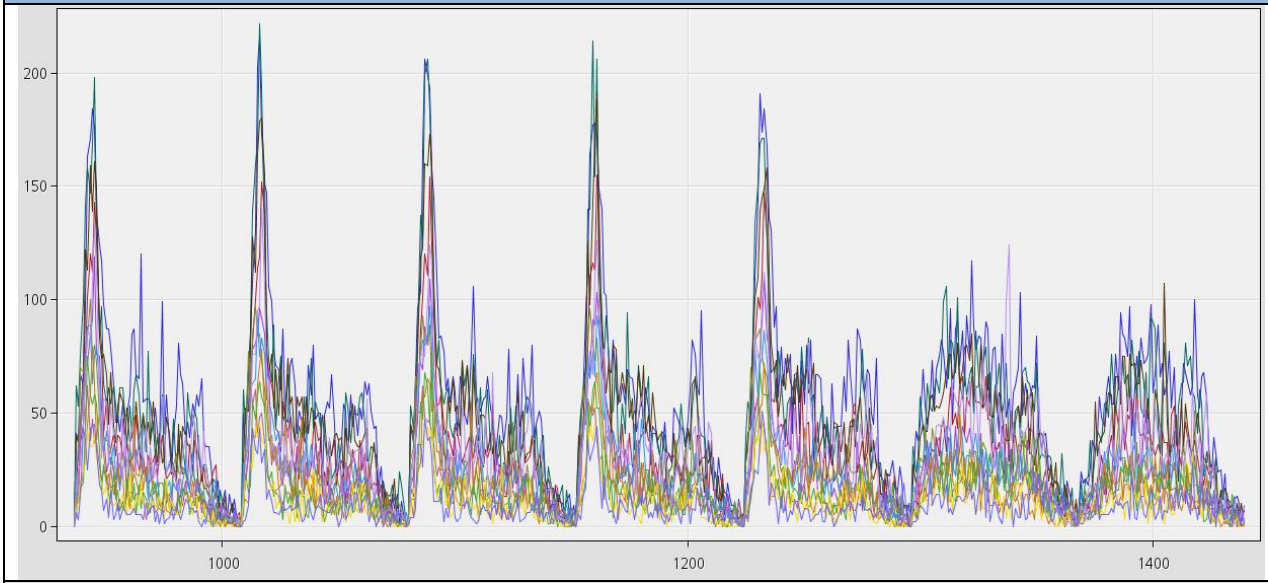
Examining into the composition of cluster C, we found that it was make up of MRT stations situated in commercial and industrial areas. This could give us a preliminary explanation for the weekday evening peak where the passengers were leaving their workplace to return back home. Another interesting observation is the weekday evening passenger volume of commercial and industrial area (cluster C) was higher than weekday evening passenger volume of residential area (cluster A and B). This might be because there were more train stations serving residential areas than commercial and industrial areas.

4.3.5. Cluster D: Moderate Morning Peak – Residential Area

Month View



Week View



Cluster D – Compassvale LRT Station, Coral Edge LRT Station, Fernvale LRT Station, Kangkar LRT Station, Kadaloor LRT Station, Keat Hong LRT Station, South View LRT Station, Phoenix LRT Station, Renjong LRT Station, Damai LRT Station, Ranggung LRT Station, Oasis LRT Station, Riviera LRT Station

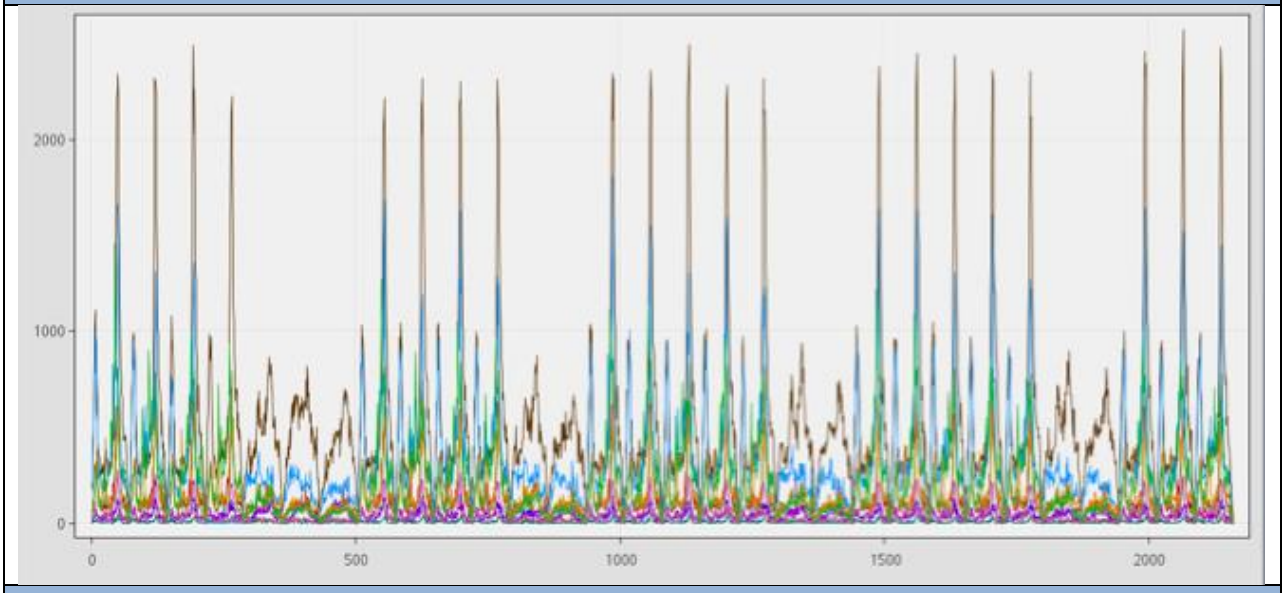
The time-series data plots in cluster D have displayed a moderate morning peak on weekdays. This suggests that the train stations in cluster D were experiencing moderately high passenger volume entering the stations in the morning. However the morning peak pattern was not

observed in weekend, where the stations received relatively constant passenger volume throughout the day.

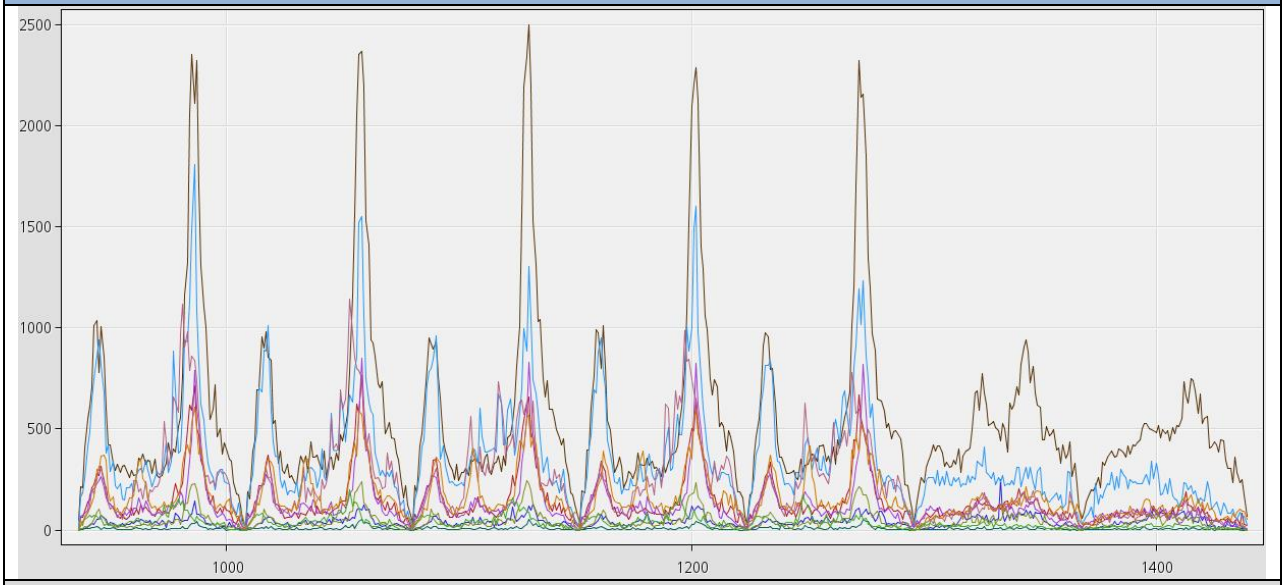
Examining into the composition of cluster D, we found that it was make up of LRT stations situated in residential areas. This could give us a preliminary explanation for the weekday morning peak where the passengers living in residential areas were traveling to work on weekday morning. The morning peak in cluster D also generally had a relatively lower passenger volume compare to the passenger volume in cluster B. This could be due to lesser residents in the residential areas served by stations in cluster D, or the residents could have a better alternative mode of transport (i.e. public bus or private cars).

4.3.6. Cluster E: Moderate Morning/ Peak Strong Evening Peak – Industrial/ Commercial/ Residential Area

Month View



Week View



Cluster E – Botanic Gardens MRT Station, Caldecott MRT Station, Buona Vista MRT Station, Jurong East MRT Station, MacPherson MRT Station, Haw Par Villa MRT Station, Dover MRT Station, Yio Chu Kang MRT Station, Newton MRT Station, Teck Whye LRT Station

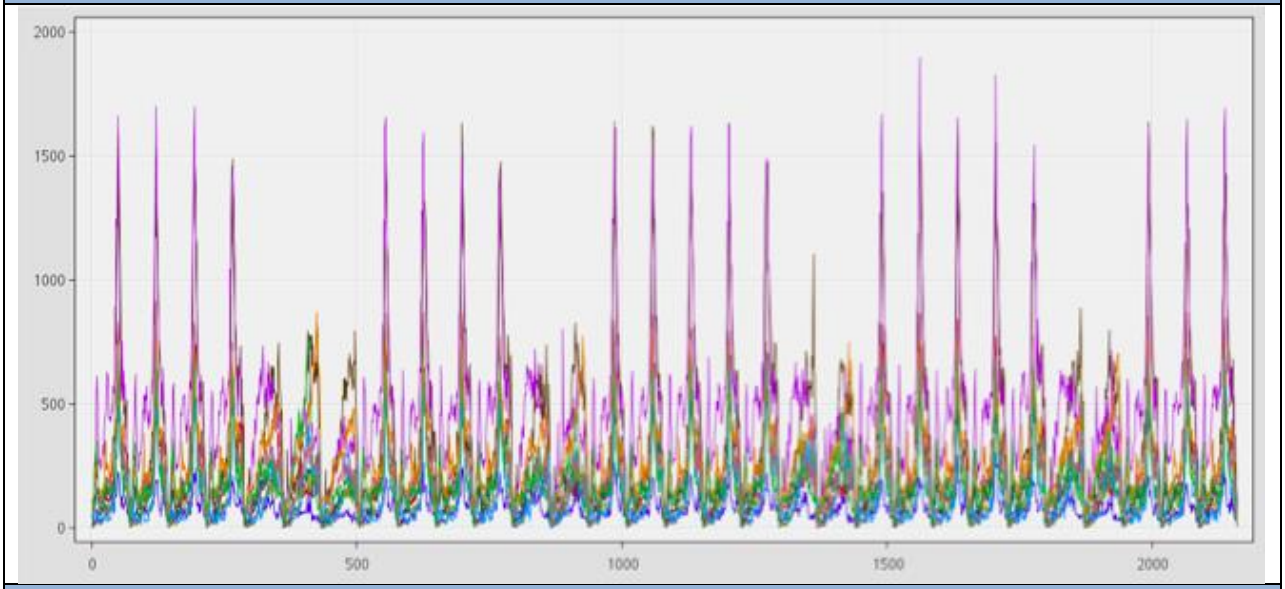
The time-series data plots in cluster E have displayed a moderate morning peak and strong evening peak on weekdays. This suggest that the train stations in cluster E were experiencing moderately high passenger volume entering the stations in the morning and relatively higher

passenger volume in the evening. However this morning and evening peak patterns were not observed in weekends, where the stations received relatively constant passenger volume throughout the day.

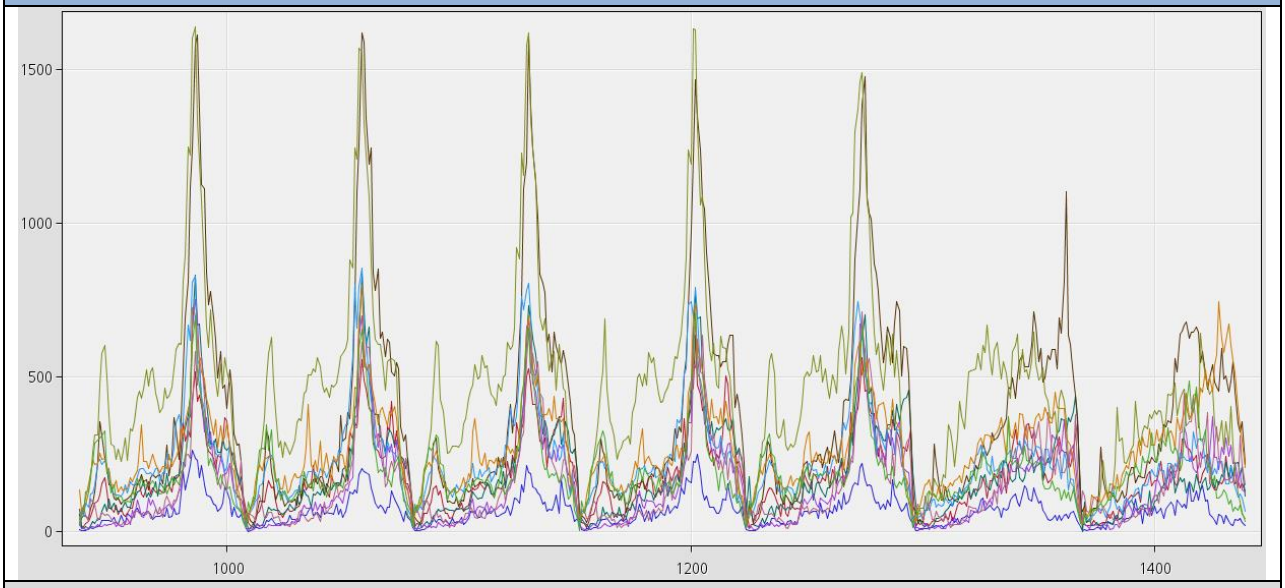
Examining into the composition of cluster E, we found that it was make up of MRT and LRT stations situated in residential areas that engages in significantly high amount of commercial and industrial activities. Contrasting this to cluster A, where the train stations were experiencing higher passenger volume in the morning than evening, the stations in cluster E might be serving an area where there were lesser residents and more commercial and industrial activities. Thus, we observed higher passenger volume entering the stations in the evening to travel back home from work than the morning passenger volume where the residents depart for their workplace.

4.3.7. Cluster F: Strong Evening Peak – Industrial/ Commercial/ Retail Area

Month View



Week View



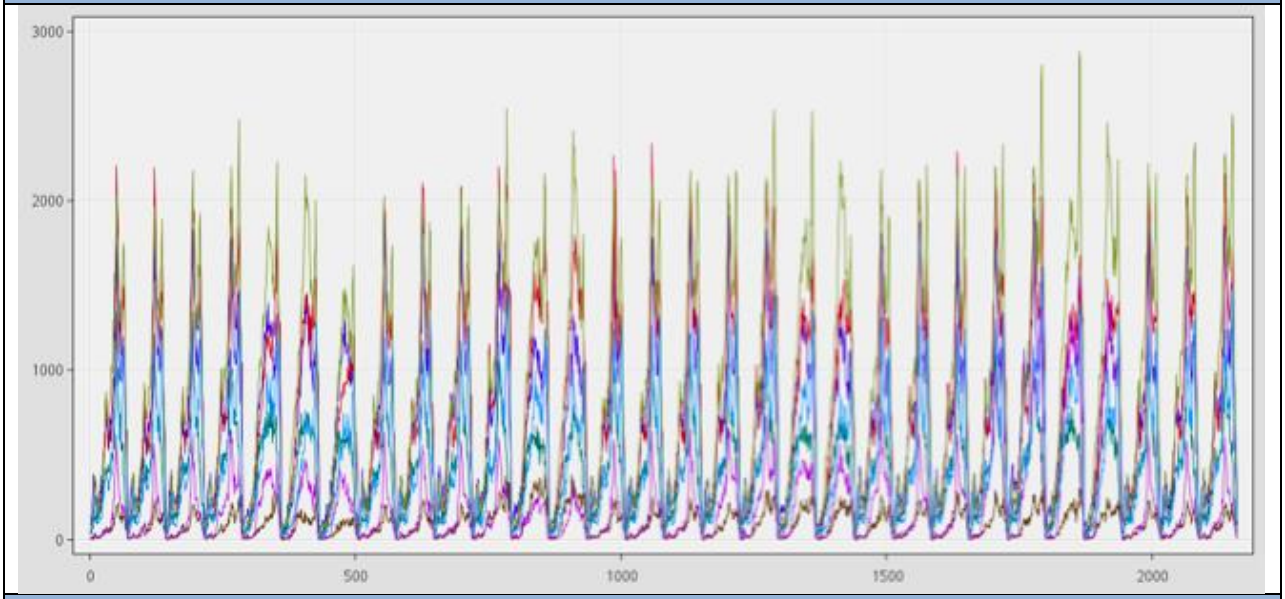
Cluster F – Bras Basah MRT Station, Dhoby Ghaut MRT Station, Clarke Quay MRT Station, Harbour Front MRT Station, Promenade MRT Station, Novena MRT Station, Outram Park MRT Station, Marina Bay MRT Station, Little India MRT Station, Paya Lebar MRT Station

The time-series data plots in cluster F have displayed strong evening peak on weekdays. This suggests the train stations in cluster F were experiencing high passenger volume entering the stations in the evening. However the evening peak pattern was not observed in weekend although relatively higher passenger volume was seen in the evening.

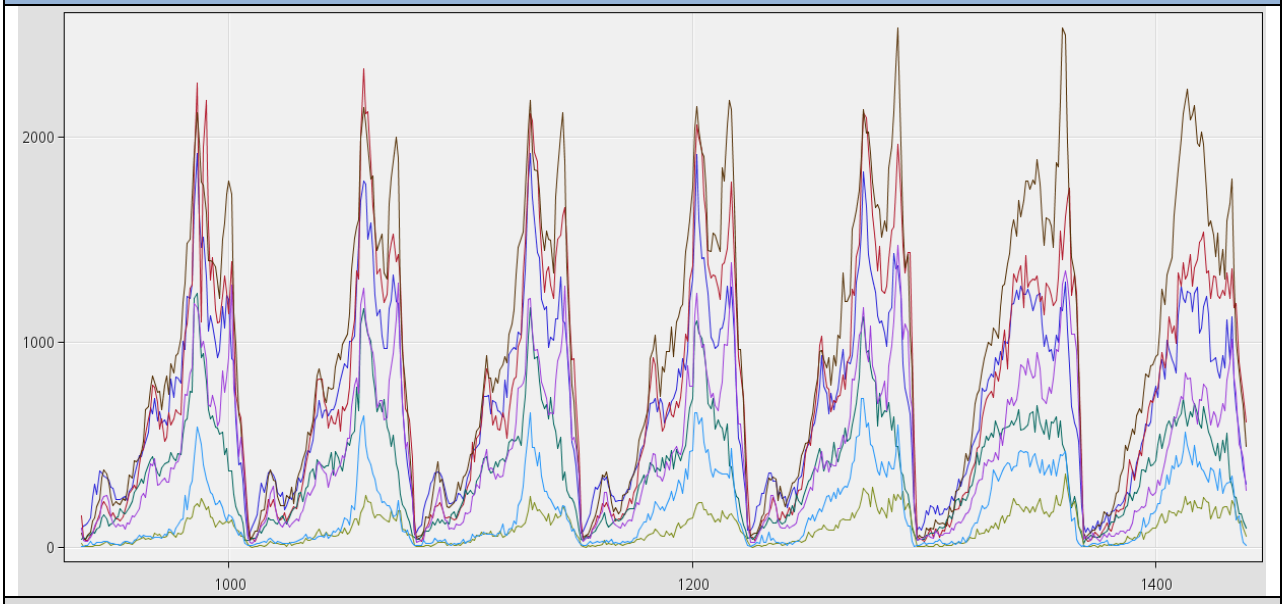
Examining into the composition of cluster F, we found that it was make up of MRT stations situated in commercial and retail areas. This could give us a preliminary explanation for the weekday evening peak where the passengers were leaving their workplace or retail areas to return back home. Comparing cluster F and other evening peak time-series plots such as cluster C, cluster F displayed relatively higher passenger volume in late evening. This might be due to the passengers spending more times in the retail areas as compare to cluster C where the passengers are rushing to return home.

4.3.8. Cluster G: Gentle Evening Peak – Retail Area

Month View



Week View



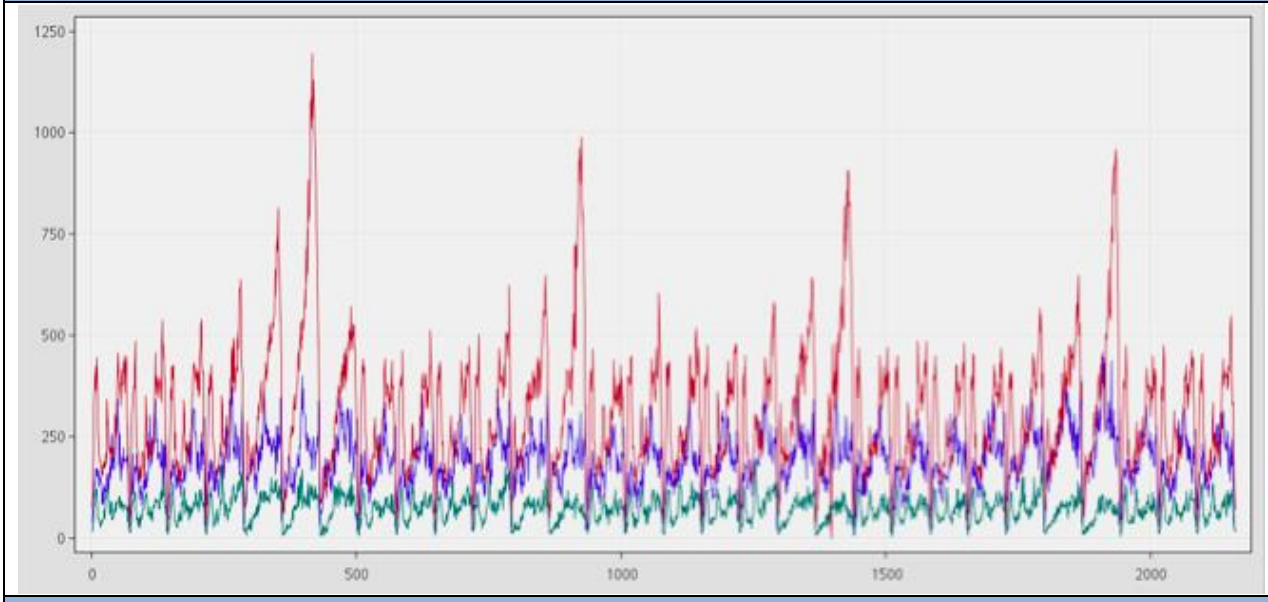
Cluster G – Bugis MRT Station, City Hall MRT Station, Chinatown MRT Station, Orchard MRT Station, Somerset MRT Station, Esplanade MRT Station, Expo MRT Station

The time-series data plots in cluster G have displayed a gentle to evening peak pattern for both weekend and weekday. This suggests the train stations in cluster G were experiencing gentle building up of passenger volume which peak at every evening.

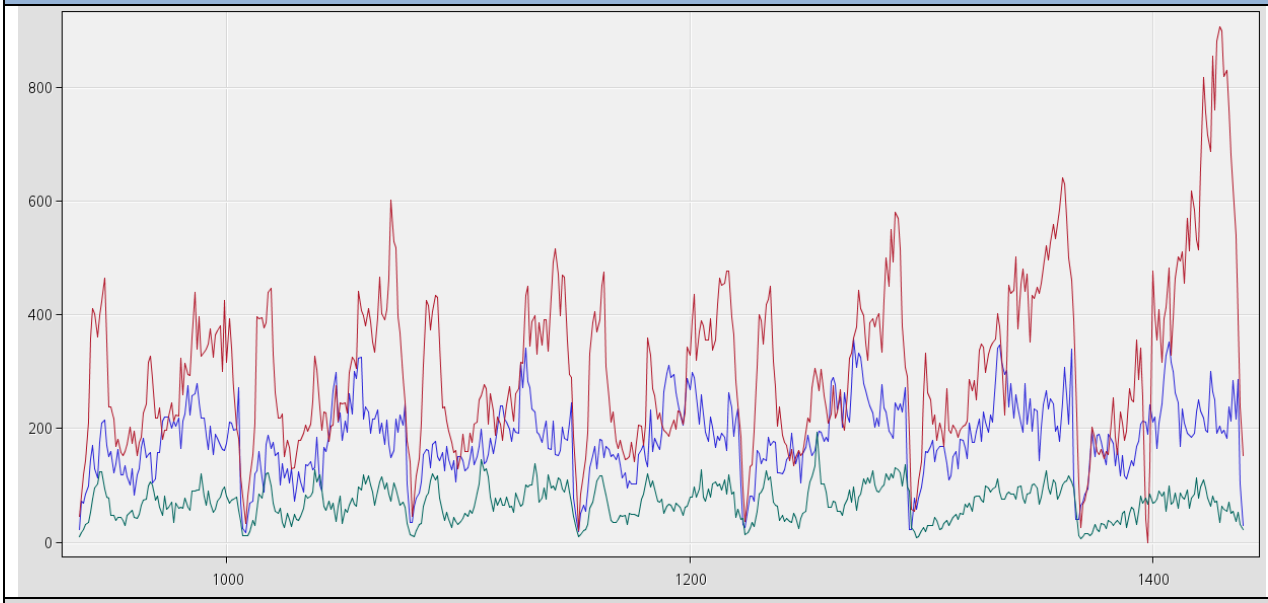
Examining into the composition of cluster G, we found that it was make up of MRT stations situated in retail areas. This could give us a preliminary explanation for the consistent gentle evening peaks where the passengers visiting the retail areas were leaving to return back home.

4.3.9. Cluster H: Weekend Peak – Special Weekend Activities Area

Month View



Week View



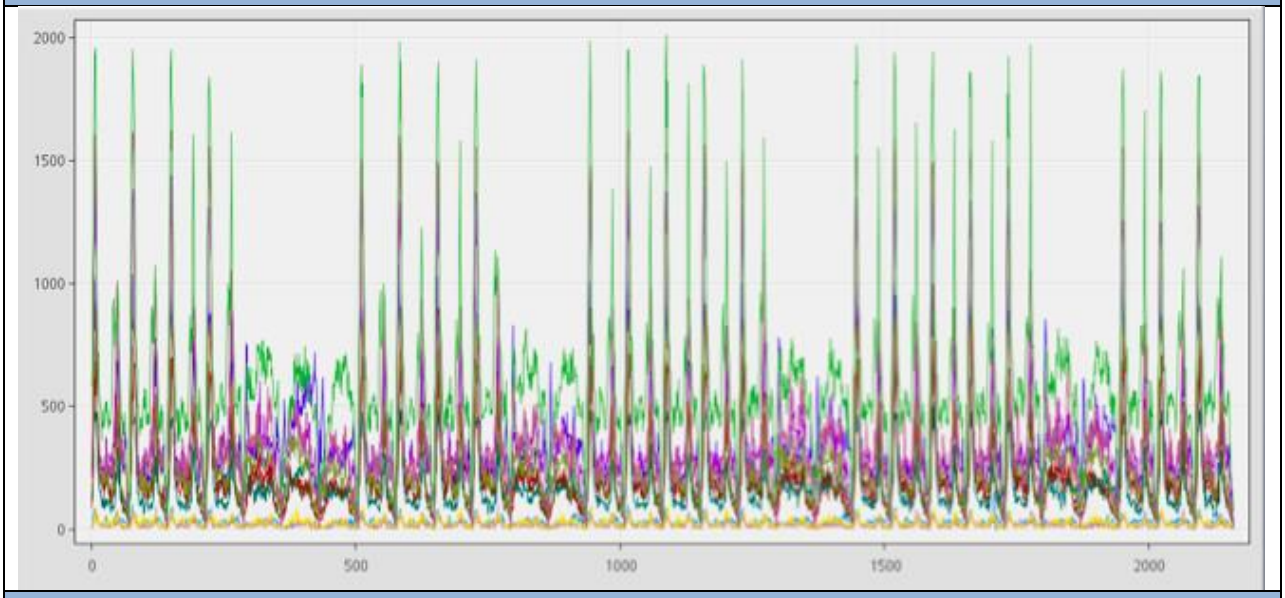
Cluster H – Changi Airport MRT Station, Farrer Park MRT Station, Holland Village MRT Station

The time-series data plots in cluster H have displayed a fairly constant pattern for both weekday and weekend with the exception of Changi Airport (red line), which is observed to have a weekend evening peak. This suggests the train stations in cluster H were experiencing fairly evenly distributed passenger volume throughout the day. However, the Changi Airport station

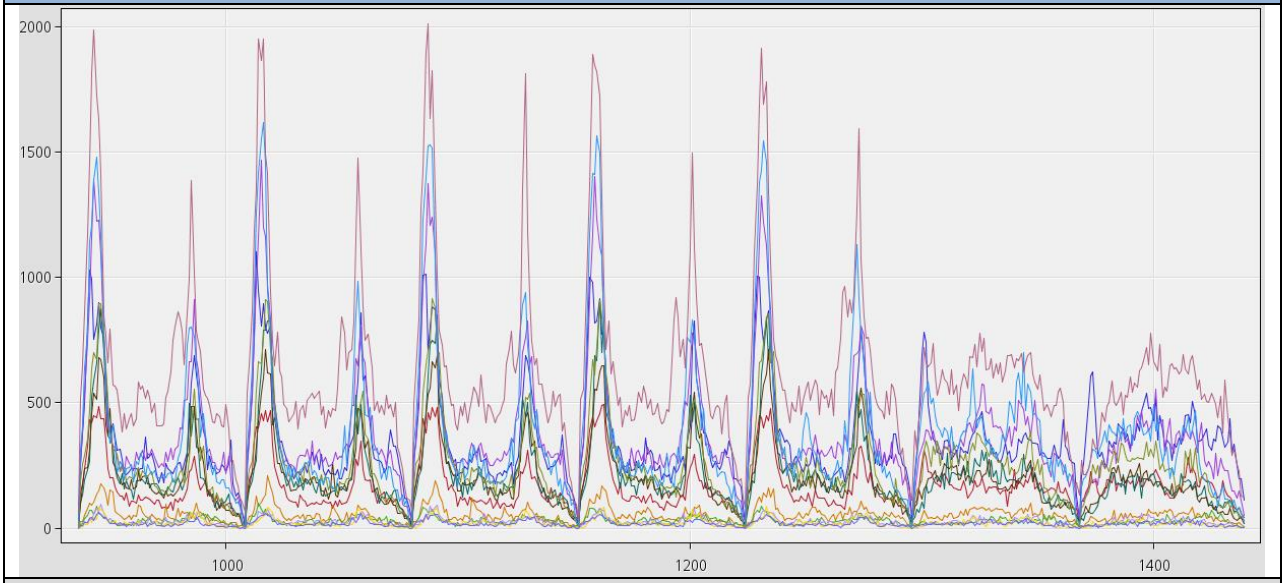
seems to experience higher passenger volume on weekend evenings. One possible explanation could be there were more passengers patronizing the retail facilities of the airport on weekends.

4.3.10. Cluster I: Strong Morning Peak/ Moderate Evening Peak – Industrial/ Commercial/ Residential Area

Month View



Week View



Cluster I – Aljunied MRT Station, Kallang MRT Station, Bishan MRT Station, Braddell MRT Station, Bukit Batok MRT Station, Eunos MRT Station, Pioneer MRT Station, Woodlands MRT Station, Dakota MRT Station, Farrer Road MRT Station, Woodleigh MRT Station, Telok Blangah MRT Station

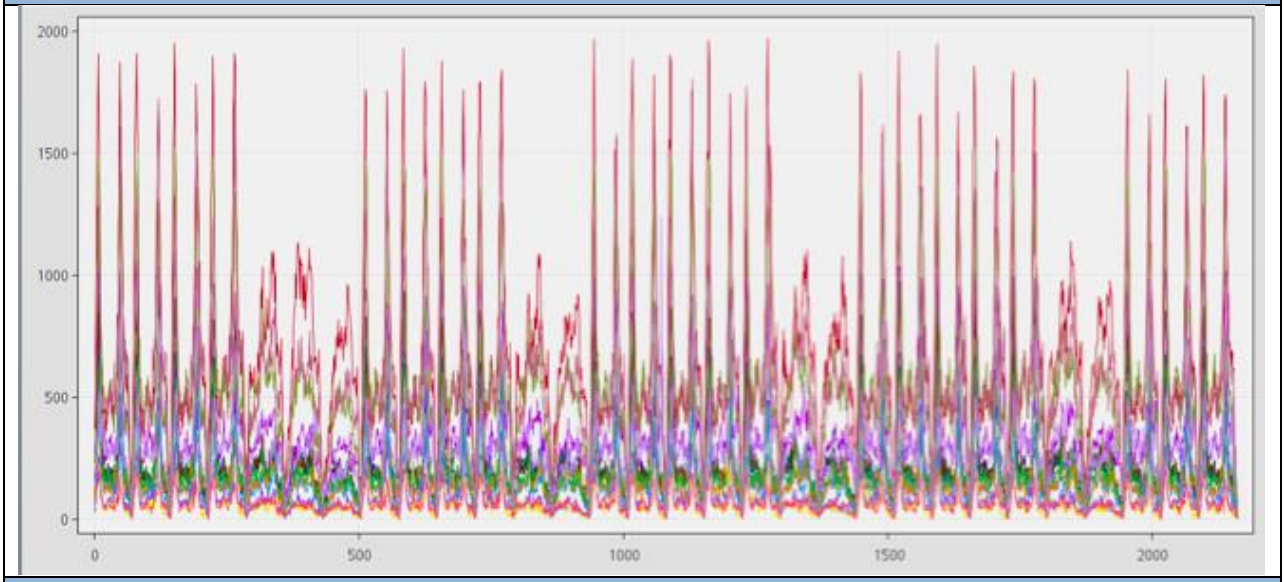
The time-series data plots in cluster I have displayed a strong morning peak and a relatively weaker evening peak on weekdays. This suggests the train stations in cluster I were experiencing

high passenger volume entering the stations in the morning and lesser passenger volume in the evening. However the morning and evening peak patterns were not observed in weekend, where the stations received relatively constant passenger volume throughout the day.

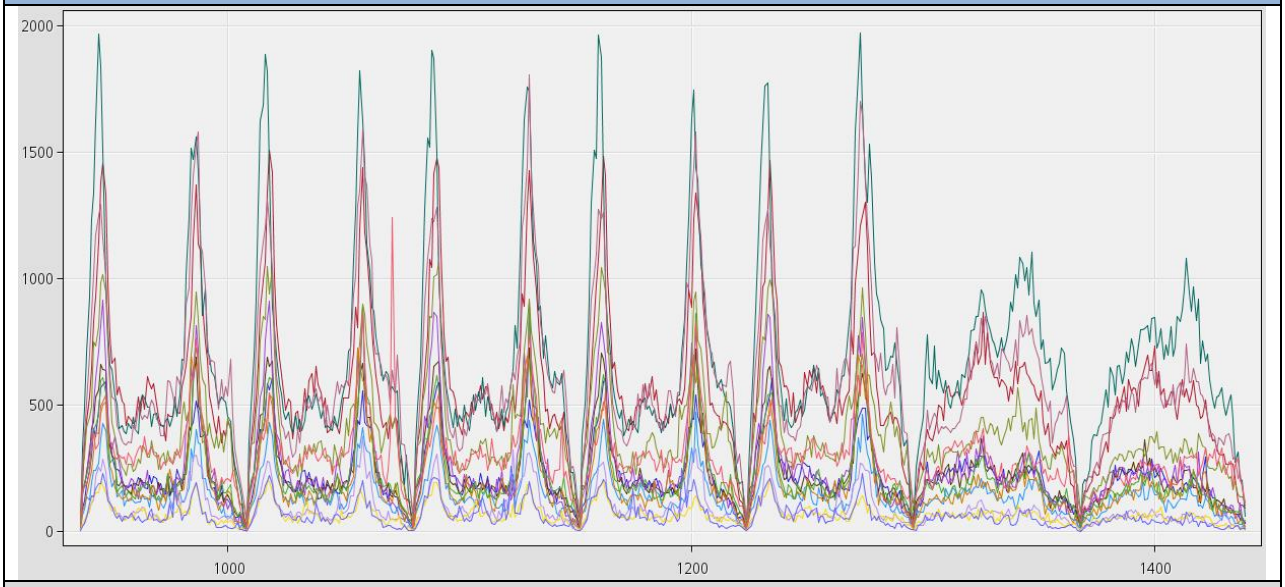
Examining into the composition of cluster I, we found that it was make up of MRT stations situated in residential areas that engage in some commercial and industrial activities. This could give us a preliminary explanation for the weekday morning peak where the passengers living in residential areas were traveling to work on weekday morning while the passengers working in the areas are returning home in the evening.

4.3.11. Cluster J: Strong Morning Peak/ Strong Evening Peak – Industrial/ Commercial/ Residential Area

Month View



Week View



Cluster J – Book keng MRT Station, Toa Payoh MRT Station, Boon Lay MRT Station, Commonwealth MRT Station, Tanah Merah MRT Station, Tiong Bahru MRT Station, Potong Pasir MRT Station, Clementi MRT Station, Queenstown MRT Station, Rehill MRT Station, Marymount MRT Station, Mountbatten MRT Station, Lorong Chuan MRT Station, Lavender MRT Station

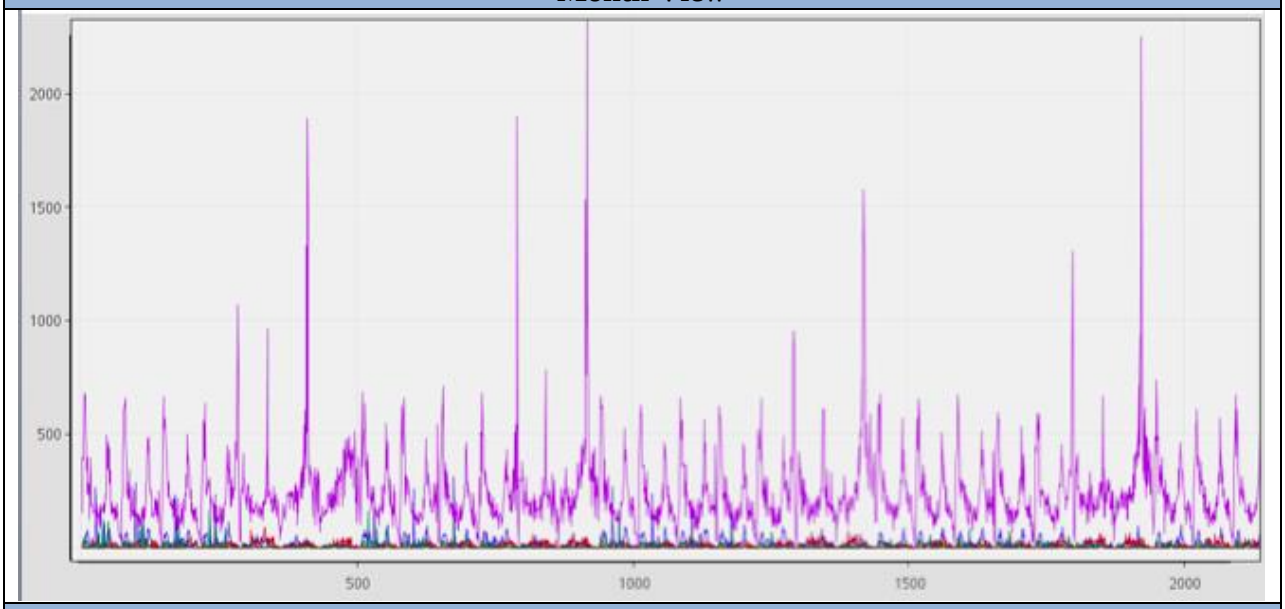
The time-series data plots in cluster J have displayed a strong morning and evening peak on weekdays. This suggests the train stations in cluster J were experiencing high passenger volume

entering the stations in both morning and evening. However the morning and evening peak patterns were not observed in weekend, where the stations received relatively constant passenger volume throughout the day.

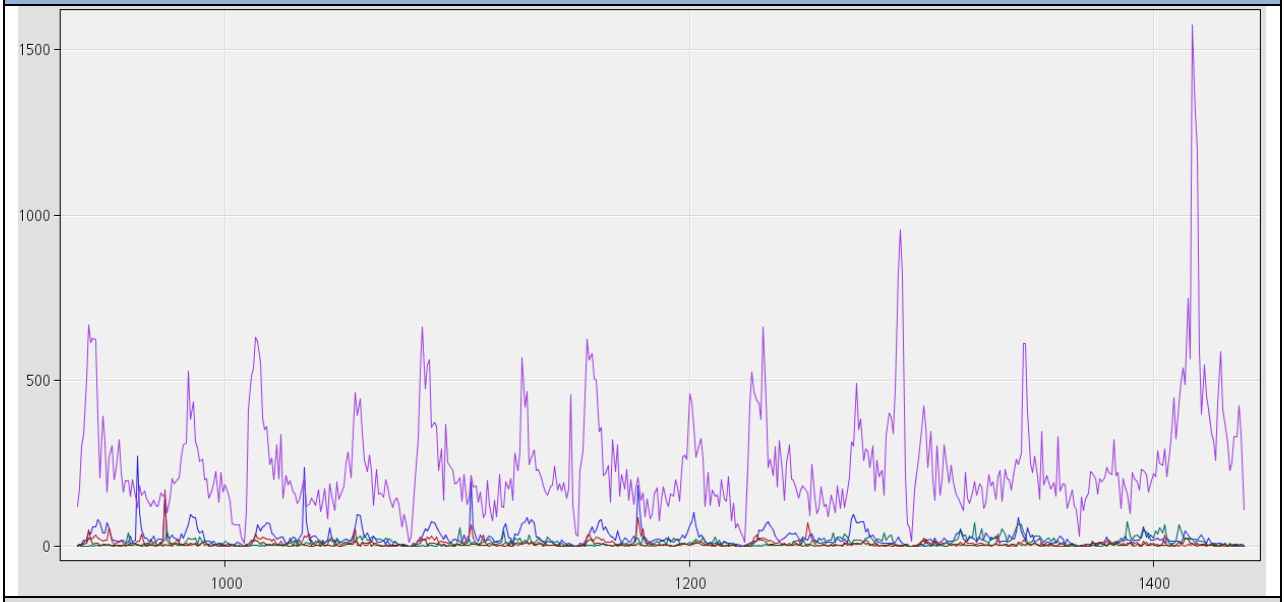
Examining into the composition of cluster J, we found that it was make up of MRT stations situated in residential areas that engage in commercial and industrial activities. This could give us a preliminary explanation for the weekday morning and evening peak where the passengers living in residential areas were traveling to work in the morning while the passengers working in the areas are returning home in the evening.

4.3.12. Cluster K: Seasonal Peak –Special Activies Area

Month View



Week View



Cluster K – Bartley MRT Station, Farmway LRT Station, Layar LRT Station, Kranji MRT Station, Thanggam LRT Station

The time-series data plots in cluster K to be haphazard and does not display any patterns. This could be because the train stations are situated in less developed areas where there were not much residential, industrial and commercial activities.

4.4. Summary of Analysis and Insights for Urban Transport Planners

The examination and analysis of the unique and distinctive passenger travel patterns in the 11 clusters has revealed that passengers travel patterns from different train stations are not homogenous. As such, urban transport planners would have to structure more dynamic policies that take into considerations the heterogenous passenger travel patterns.

For example, the Land Transport Authority of Singapore had decided to allow public train commuters to travel free on weekdays if the commuters exit 16 stations in central city area before 7.45AM [25]. The objective of this initiative is to get public train commuters to travel earlier so as to ease off the huge surge in passenger volume in late evening. Based on the findings in this research, an improvement to this initiative could be allowing commuters to travel free if they could enter residential area stations before their peak hour. This improvement will allow more certainty in easing off passenger volumes of the origin stations and prevent building huge crowd at the destination at 7.45AM.

4.5. Research Limitations and Future Works

This research had demonstrated the usefulness of time-series data mining techniques for knowledge discover on Singapore's public train passenger travel patterns. However, there are dimensions that this research did not covered and would be potential areas for future works. Some of these areas include:

- Exit Timing: This research was done based on the entry timestamp. It would complete the analysis if another time-series data mining were done based on the exit timestamp of passengers exiting the train stations.

- Gravity Model of Migration: As seen in each of the cluster analysis and insights interpretation, it would be helpful if we could ascertain if indeed the train stations are situation in residential, commercial office or retail areas. This will help us to explain the public train passengers' travel patterns in greater details.
- Predictive Analytics: Predictive analytics can be done using the results of this research to predict how the passengers would behave when the public train extension works for 2020 are completed.

5. CONCLUSION

With the application of time-series data mining techniques and sensing data in transportation studies, urban transport planners and analysts will be able to analyze the passenger travel patterns faster and gain greater insights beyond what could be provided by conventional statistical analysis or traditional data mining techniques. There are also a number of future works that could be done to generate greater insights and knowledge discovery. The time-series data mining framework proposed in this research is also extensible to study other transport modes such as buses and taxis, and beyond the urban transportation industry too.

6. REFERENCE

[1] A. M. Yam. (2008, Nov.). Shaping Urban Journeys, Journey. [Online]. 1, Available: http://ltaacademy.lta.gov.sg/doc/Yam%20Ah%20Mee.pdf

[2] C.V Ferber,T. Holovatch Y. Holovatch and V. Palchykov, "Public transport networks: empirical analysis and modeling", The European Physical Journal B, Vol 68, pp 261-275, 2009

[3] H. S, S. Lim, T. Zhang, X. Fu, G. K. K. Lee , T. G. G.Hung, P. Di, S. Prakasam and L. Wong, "Weighted complex network analysis of travel routes on the Singapore public transportation system", Physica A, Vol 389, No. 24, pp. 5852-5863, December 2010

[4] J.B. Gordon, "Intermodel Passenger Flows on London's Public Transport Network", M.S. thesis, Dept. Urban Studies and Planning, Univ. California, Berkeley, 2012

[5] Land Transport Authority (LTA). Land Transport Master Plan, Singapore, 2008

[6] Mass Rapid Transit Corporation. The MRT Story, Singapore, 1988

[7] Choi.C and Toh.R. Household Interview Surveys from 1997 to 2008—a Decade of Changing Travel Behaviours. *Journeys*, 2, page 52–61, 2010

[8] Low. I (December 16, 2011). Singapore's MRT Breakdown Chaos Leaves Thousands Stranded. *The Straits Times*.

[9] M.Bagchi and P.R.White. The potential of public transport smart card data. *Transport Policy*, 12, page 464-474, 2005

[10] C.Morency, M.Trepanier and B.Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14, page 193-203, 2007

[11] Y.Asakura, T.Iryo, Y.Nakajima and T. Kusakabe Estimation of behavioural change of railway passengers using smart card data. *Public Transport*, 4(1), page 1-16, 2012

[12] J.Kim, and S.Kang. Development of Integrated Transit-Fare Card System in the Seoul Metropolitan Area. *Knowledge-Based Intelligent Information and Engineering Systems Lecture Notes in Computer Science*, 3683, page 95-100, 2005

[13] D-H.Lee, L.Sun and A.Erath. Study of Bus Service Reliability in Singapore Using Fare Card Data. Paper submitted for The 12th Asia-Pacific ITS Forum & Exhibition 2012, Kuala Lumpur, April 2012.

[14] L.Sun, D-H.Lee, A.Erath and X.F.Huang. Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System.

[15] H.Soh, S.Lim, T.Zhang, X.Fu, G.K.K.Lee, T.G.G.Hung, P.Di, S.Prakasam and L.Wong. Weighted complex network analysis of travel routes on the Singapore public transportation system. *Physica A*, 389, page 5852-5863, 2010

[16] K.Nakkeeran, S.Garla and G.Chakraborty, Application of Time-series Clustering using SAS® Enterprise Miner™ for a Retail Chain, Proc of SAS® Global Forum, 2012

[17] D.Hebert, Time-series Data Mining: A Retail Application Using SAS Enterprise Miner,

[18] D.J.Berndt and J.Clifford, Using dynamic time warping to find patterns in time-series, *AAAI Working Notes of the Knowledge Discovery in Databases Workshop*, page 359–370, 1994

[19] M.Leonard, J.Lee, T.Y.Lee and B.Elsheimer, An Introduction to Similarity Analysis Using SAS, Proc of International Symposium of Forecasting. SAS Institute Inc., 2008

[20] M.Leonard and B.Wolfe, Mining Transactional and Time-series Data,

[21] M.P.Pelletier, M.Trepanier and C.Morency, Smart Card Data Use in Public Transit: A Literature Review, *Transportation Research Part C*, 19, page 557-568, 2011

[22] S.Schubert and T.Y. Lee, Time Series Data Mining with SAS Enterprise Miner, Proc of SAS® Global Forum, 2011
[23] W.Frawley, C.Piatesky-Shapiro and C.Matheus, Knowledge Discovery in Database: An Overview, *AI Magazine*, Fall, page 213-228, 1992
[24] U.Fayyad, G.Piatesky-Shapiro and P.Smyth, From Data Mining to Knowledge Discovery in Database, *AI Magazine*, Fall, page 32-54, 1996
[25] LTA Website, *Travel Early*, *Travel Free*, *Travel on the MRT*, article retrieved from http://www.lta.gov.sg/content/ltaweb/en/public-transport/mrt-and-lrt-trains/travel-smart.html