Time-Series Data Mining in Transportation: A Case Study on Singapore Public Train Commuter Travel Patterns

Roy Ka-Wei Lee and Tin Seong Kam

***Abstract*—The adoption of smart cards technologies and automated data collection systems (ADCS) in transportation domain had provided public transport planners opportunities to amass a huge and continuously increasing amount of time-series data about the behaviors and travel patterns of commuters. However the explosive growth of temporal related databases has far outpaced the transport planners' ability to interpret these data using conventional statistical techniques, creating an urgent need for new techniques to support the analyst in transforming the data into actionable information and knowledge. This research study thus 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 the transportation temporal data. A case study on the Singapore public train transit will also be used to demonstrate the time-series data-mining framework and methodology.**

***Index Terms*—Time-series data mining, smart card, big data, transportation.**

I. Introduction

A. Motivation

Bagchi and White [1], in their paper, "*The potential of public transport smart card data*", introduced and explored the possibilities of using smart cards beyond fare collection. The smart cards data, beyond just the transacted fare prices, also contains rich information such as boarding and exiting time, as well as geospatial information such as the boarding and exiting bus stops and train stations. Since then, urban transport planners had tried to analyze the collected smart card data in attempt to discover useful information and knowledge on the commuters' travel patterns and behaviors.

However, the analysis of the smart card data was proven to be a challenge as the automatically collected smart card data were large and continue to increase in size. This explosive growth in such complex temporal data has far outpaced the urban transport planners' ability to interpret these data using conventional statistical techniques. As such, there is an urgent need for new techniques to allow urban transport planners to transform these massive complex data into actionable information ad knowledge.

B. Research Objectives and Contributions

In this research study, we therefore aim to explore and discuss 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 the smart card temporal data. A case study on the Singapore public train transit will also be used to demonstrate the time-series data mining framework and methodology.

In the following, we summarize the contribution of this paper:

- We discuss and explore the use of time-series data mining techniques in transportation domain. The application of time-series data mining would allow transport planners to effectively transform large amount of complex temporal data into actionable insights and knowledge.
- We demonstrate the use of time-series data mining in transportation domain with a case study on Singapore public train transit. The case study would apply time-series data mining techniques on over 60 million commuters' public train transit trips and generate travel patterns of commuters for 102 train stations.
- We derived several interesting insights on commuters' travel patterns and behavior from the case study. These generated insights provide an example to urban transport planners on how they could leverage time-series data miming to better understand the commuters' travel patterns and behaviors.

C. Paper Outline

The rest of the paper is organized as follows. Section II reviews the literatures related to our study. Section III discusses the time-series data mining technique. The case study on Singapore public train transit will be introduced in Section IV. Section V describes the application of time-series data mining methodology and framework used in the case study. We present the insights generated from this case study in Section VI before concluding the paper in Section VII.

II. Literature Review

In recent years, smart cards and automated data collection systems (ADCS) had also been widely adopted in public transport networks around the world. These smart cards, which are portable credit card size devices that store and process data, were credited with monetary value and used for public transit fare payment [2].

Although the preliminary purpose of smart cards is for collection of fares for public transport, Bagchi and White [1] discussed the potential use of the passively recorded

Manuscript received November 25, 2013; revised March 3, 2014.

The authors are with Singapore Management University, Singapore (e-mail: roylee.2013@phdis.smu.edu.sg).

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 [3] 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 [4], constructed a origin-destination (O-D) matrix using the smart card data collected from Japan's public train network and applied statistical analysis to study the change in passengers' travel patterns when the train operator changed its train timetable. Kim and Kang [5] also attempted a similar research on Seoul public transit network by using the transaction data collected from T-Money, South Korea's electronic fare card system.

Locally, there were also some researches done using the data collected from EZ-Link, Singapore's smart card used for public transit. Lee et al [6] did a case study to optimize serviceability and reliability of bus routes by running statistical analysis on the EZ-Link data collected from bus transit trips. Sun et al [7] also did a study using the EZ-Link data collected from public train transit to estimate the spatial-temporal density passenger onboard a train or waiting in the train station.

In the literature, there is a large body of work on applying statistical analysis on the smart card data collected from public transit trips. However, few researches were able to deal with the growing large amount of temporal data effectively or were there any extensive study done using the smart card data to reveal the commuters' travel patterns.

III. 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 [8], 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 widely studied time-series data mining technique is the dynamic time warping (DTW) algorithm proposed by Berndt and Clifford [9]. 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. 1. Comparing time-series using Euclidean distance.

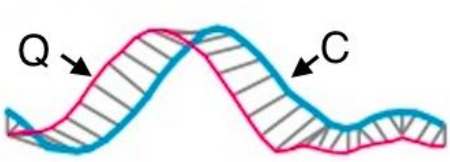

Fig. 2. Comparing time-series using dynamic time warping.

Take for example, Fig. 1, 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. 2, DTW algorithm is used to overcome this problem by accounting for the time factor when comparing the two time-series.

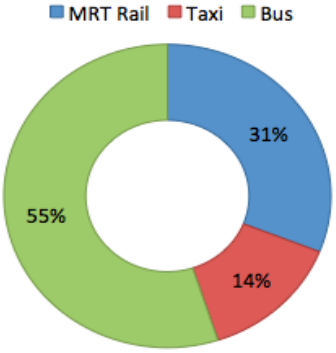

Fig. 3. Singapore public transport mode share.

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 [10] 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. Although there is an increased interest for organizations to explore time-series data mining, few research studies were done to apply time-series data mining in the transportation domain.

IV. Singapore Public Train Transit

A. 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 [11]. 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. Fig. 3 shows a pie chart on the breakdown in pubic transport mode market shares in 2008 [12].

Although the market shares for bus were higher, there was a steady increase in demand for MRT rail service. Accordingly to the household transport interview survey conducted from 1997 till 2008 [12], 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 Singapore Land Transport Authority (LTA) 2008 Transport Master Plan to "make public transport a choice mode"[11].

B. 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 MRT stations and 21 Light Rail Transit (LRT) stations, which are localized rail systems acting as feeder services to the MRT network. There are 4 main operating service lines and covering an estimated 149 km [13]. 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.

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.

C. 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 goals set 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 [12]. 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 [14].

As such, another objective for this research study is to generate analysis that could help Singapore transport planners to better understand the Singapore public transport commuters. With greater insights and knowledge on the public transport commuters' travel pattern, the transport planners would be able to structure policies that could serve the commuters better and ultimately increase the public transport mode share.

V. Application of Time-Series Data Mining

A. Datasets

The EZ-Link card is a contactless smart card used mainly for the payment of public transportation fares in Singapore. For this study, we were able to obtain one month (November 2011) worth of EZ-Link smart card transaction data from the LTA. An estimated total of 60 million train transit trip transactions were made in the month of November 2011. Each trip transaction consists of quite a number of data columns, which describe a train transit 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.

B. Data Transformation

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 (O-D) time interval format for time-series data mining. A Java application was written to perform this data transformation.

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	24	30	35	2345

Fig. 4. Data transformation from transaction data to O-D time interval

Fig. 4 shows the data transformation from transaction data into the O-D time interval format. In the O-D 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.

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

C. SAS Enterprise Miner

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.*, [15] in their paper, "*An Introduction to Similarity Analysis using SAS*", described in detail how DTW technique was implemented in SAS Enterprise Miner. In another similar work, Leonard and Wolfe [16] 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 generated O-D time interval data to investigate the travel pattern of Singapore public train commuters. Fig. 5 shows SAS Enterprise Miner work process.

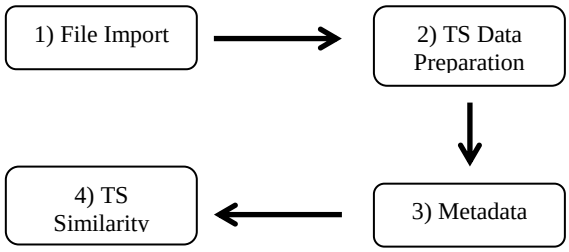

Fig. 5. SAS Enterprise Miner Time-Series Data Mining Work Process

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*.
- 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 and group time-series that exhibit similar characteristics over time. As the time series might have different lengths, DTW technique 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.

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.

VI. Time-Series Clustering and Analysis

A. 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.

TABLE I: Top 5 Stations in Various Distributions

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 I shows the general statistics of top 5 stations in 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. 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.

Secondly, The ranking of stations for SUM and MEAN is the same with the same 5 stations topping both measures. However the MAX top 5 stations are not the same as the MEAN and SUM top 5 stations. This could suggest that while certain 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.

B. Time-Series Data Plot (Overview)

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.

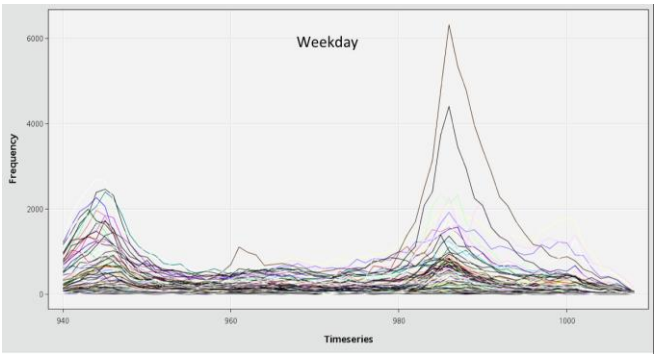

Fig. 6. Time-series plot for all train stations on weekday.

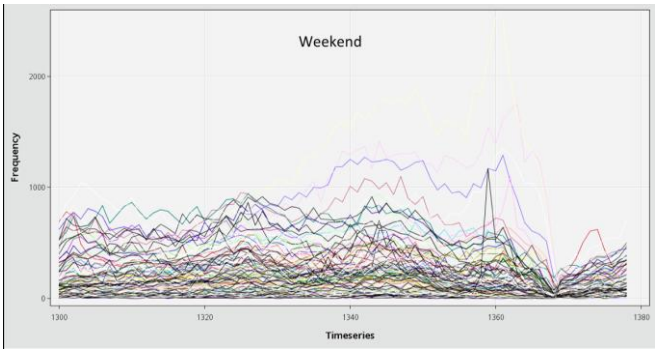

Fig. 7. Time-series plot for all train stations on weekend.

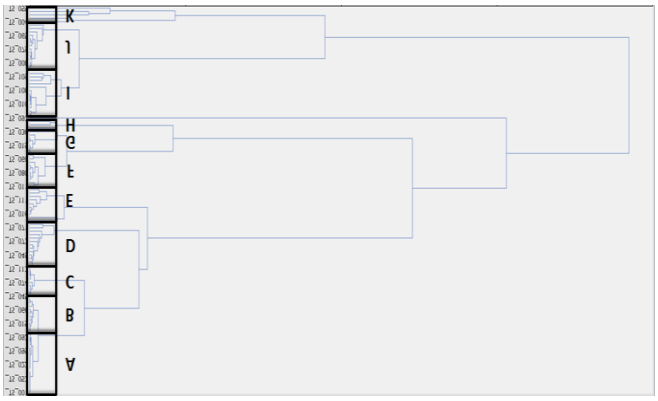

Fig. 8. Dendrogram of time-series data clustering.

Fig. 6 and Fig. 7 show 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 do 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.

C. Time-Series Data Cluster Analysis

The TSS node performs the clustering and similarity analysis on the train station time-series data.

Fig. 8 shows the dendrogram of the time-series data plots generated in the *Result panel* of the TSS node. 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. Fig. 9 shows the one-week time-series plot for the 11 clusters.

1) Cluster A – strong morning peak/ moderate evening peak

The time-series data plots in cluster A have displayed a strong morning peak and a relatively weaker evening peak on weekdays, suggesting that 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 on weekends, where the stations received relatively constant passenger volume throughout the day. Examining into the composition of cluster A, we found that it is made up of train 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.

2) Cluster B – strong morning peak

The time-series data plots in cluster B have displayed a strong morning peak on weekdays. However, the morning peak pattern was not observed on weekends. Examining into the composition of cluster B, we found that it is made up of LRT stations situated in residential areas. A possible preliminary explanation for the weekday morning peak could be the passengers living in residential areas were traveling to work on weekday morning. Another interesting observation is that 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 compared to MRT.

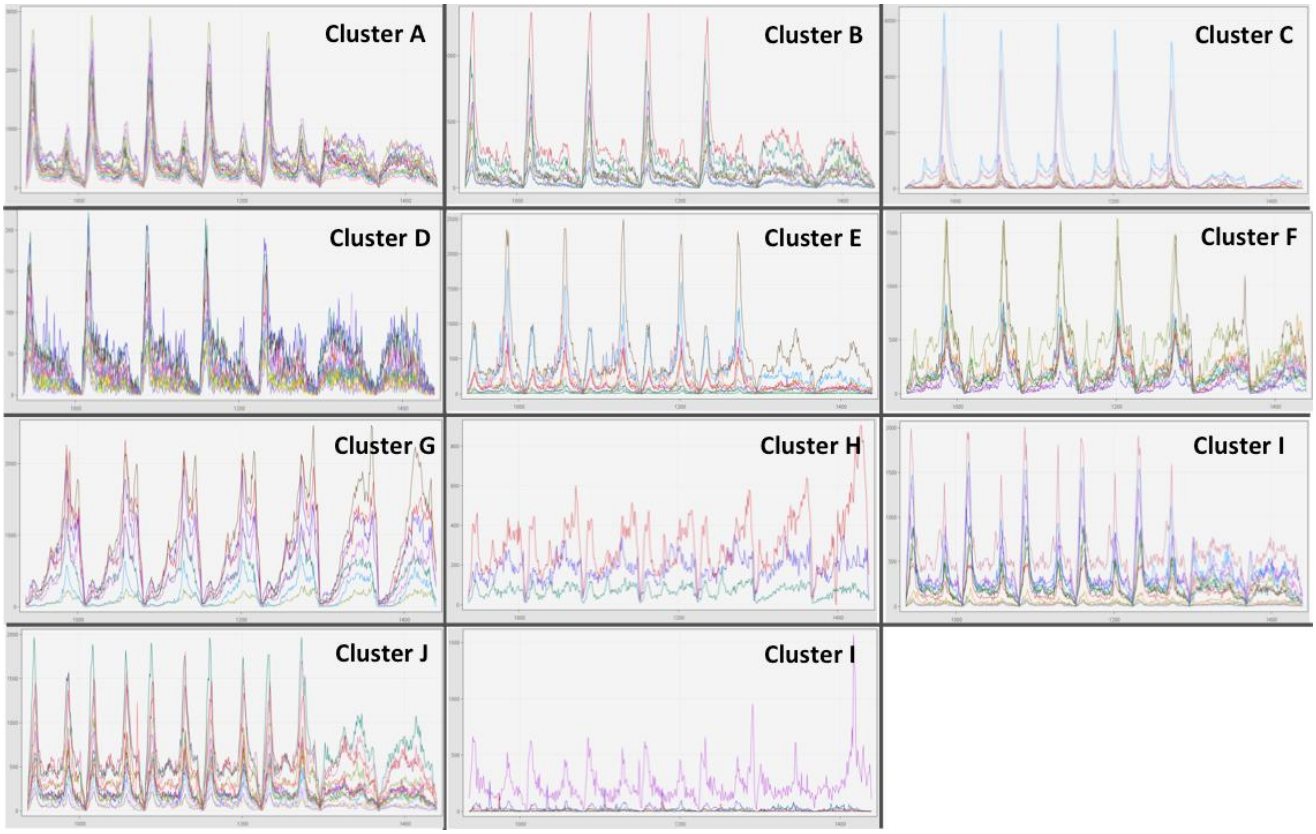

Fig. 9. One-week time-series plot for 11 clusters.

3) Cluster C – strong evening peak

The time-series data plots in cluster C have displayed a very strong evening peak on weekdays. However the evening peak pattern was not observed on weekends. Examining into the composition of cluster C, we found that it is made up of MRT stations situated in commercial and industrial areas. A preliminary explanation for the weekday evening peak could be that the passengers were leaving their workplace to return back home. Another interesting observation is that 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 due to more train stations serving residential areas than commercial and industrial areas.

4) Cluster D – moderate morning peak

Train stations in cluster D were experiencing moderately high passenger volume entering the stations in the morning. The morning peak pattern was again not observed on weekends. Examining into the composition of cluster D, we found that it is made 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 compared 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).

5) Cluster E – moderate morning peak/ strong evening peak

The time-series data plots in cluster E have displayed a moderate morning peak and strong evening peak on weekdays. However, these morning and evening peak patterns were not observed in weekends. Examining the composition of cluster E, we found that it is made up of MRT and LRT stations situated in residential areas that engage 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.

6) Cluster F – strong evening peak

Train stations in cluster F were experiencing high passenger volume entering the stations in the evening. However, the evening peak pattern was not observed on weekends although relatively higher passenger volume was seen in the evening. Examining the composition of cluster F, we found that it is made 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 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 time in the retail areas as compared to cluster C where the passengers were rushing to return home.

7) Cluster G – gentle evening peak

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 the composition of cluster G, we found that it is made 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 home.

8) Cluster H – weekend peak

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 that 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 that there were more passengers patronizing the retail facilities of the airport on weekends.

9) Cluster I – strong morning peak/ moderate evening peak

The time-series data plots in cluster I have displayed a strong morning peak and a relatively weaker evening peak on weekdays. However, the morning and evening peak patterns were not observed on weekends, where the stations received relatively constant passenger volume throughout the day. Examining the composition of cluster I, we found that it is made 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.

10) Cluster J – strong morning peak/ strong evening peak

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 on weekends. Examining the composition of cluster J, we found that it is made 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.

11) Cluster K – Seasonal peak

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.

VII. Insights and Discussion

The examination and analysis of the unique and distinctive passenger travel patterns in the 11 clusters have 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 heterogeneous 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 [17]. 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.

A. Research Limitations and Future Works

This research had demonstrated the usefulness of time-series data mining techniques for knowledge discovered on Singapore's public train passenger travel patterns. However, there are dimensions that this research did not cover and would be the 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.

VIII. 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 for other cities' transportation networks too.

Acknowledgment

We would like to thank the Land Transport Authority (LTA) of Singapore for sharing with us the MRT dataset. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

REFERENCES

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

[2] M. P. Pelletier, M. Trepanier, and C. Morency, "Smart card data use in public transit: a literature review," *Transportation Research Part C*, vol. 19, pp. 557-568, 2011.

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

[4] Y. Asakura, T. Iryo, Y. Nakajima, and T. Kusakabe, "Estimation of behavioural change of railway passengers using smart card data," *Public Transport*, vol. 4, no. 1, pp. 1-16, 2012.

[5] 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*, vol. 3683, pp. 95-100, 2005.

[6] 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.

[7] 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," in *Proc. the ACM SIGKDD International Workshop on Urban Computing*, pp. 142-148, 2012.

[8] S. Schubert and T. Y. Lee, "Time series data mining with SAS enterprise miner," in *Proc. SAS® Global Forum*, pp. 1-9, 2011.

[9] 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*, pp. 359–370, 1994.

[10] K. Nakkeeran, S. Garla, and G. Chakraborty, "Application of Time-series clustering using SAS® enterprise Miner™ for a retail chain," in *Proc. SAS® Global Forum*, pp. 1854-1874, 2012

[11] *Land Transport Authority (LTA)*, Land Transport Master Plan, Singapore, 2008

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

[13] Mass Rapid Transit Corporation, *The MRT Story*, Singapore, 1988

[14] I. Low, *Singapore's MRT Breakdown Chaos Leaves Thousands Stranded*.

[15] M. Leonard, J. Lee, T. Y. Lee, and B. Elsheimer, "An introduction to similarity analysis using SAS," in *Proc. International Symposium of Forecasting.*, pp. 302, 2008

[16] M. Leonard and B. Wolfe, "Mining transactional and time-series data," in *Proc. International Symposium of Forecasting*, pp. 1-26, 2002

[17] LTA Website. [Online]. Available: http://www.lta.gov.sg/content/ltaweb/en/public-transport/mrt-and-lrt-trains/travel-smart.html

Lee Ka Wei Roy is a PhD candidate under the supervision of professor Ee-Peng LIM from Singapore Management University, School of Information Systems. His main research interests include spatial-temporal related researches and social network mining and analysis. He had also completed his Master and Bachelor Degree on Information Systems from Singapore Management University.

Kam Tin Seong is a practice associate professor of information systems at the School of Information Systems, Singapore Management University. His current teaching and research interests are in visual analytics, business analysis and data mining, and geospatial science and technologies. He had received his PhD from University of London