A Machine Learning Approach for Detecting Long-Term Changes in the Weekly Trip Patterns of Public Transport Passengers

Chun Yong Moon · Jiwon Kim · Mark Hickman

Abstract
This study proposes a machine learning framework for detecting public transit passengers with long-term behaviour changes from smart card data. In this study, we focus on detecting changes in temporal trip patterns, such as the time-of-day and the day-of-week of transport usage. Following the detection, we develop an automatic detection algorithm in order to identify the passengers who have changed their temporal transport usage patterns. The algorithm also detects passengers who have stopped using public transit altogether or have become new frequent transit users. Automatically detecting such passengers from massive smart card data helps transit operators identify more easily the causes of major behavioural changes of passengers, allowing them to develop better transit services and customer retention strategies.

Keywords: Public Transport · Long Term Travel Behaviour · Smart Card Data · Machine Learning
1 Introduction

Researchers are interested in using large amounts of smart card data to gain greater insight into passengers’ mobility patterns and travel behaviours within an urban transport system. Huge amounts of data on passenger mobility are collected regularly from public transport (PT) systems through the use of smart cards, as automated transit fare collection devices. Smart card data contain detailed information on the boarding, alighting and transferring actions of passengers. Large-scale smart card data that are collected over a long-term period allow in-depth investigation into individual passengers’ repeated trip patterns, such as regular and variable daily and weekly trip patterns.

The variability of a PT passenger’s trip patterns is categorized into two types of temporal changes: short-term, transient fluctuations and long-term, lasting changes. The latter is particularly interesting from a PT operator’s perspective because understanding what drives such systematic changes gives PT operators valuable insight into passengers’ behaviour. For instance, we may observe passengers changing their departure times in response to a new PT schedule or fare policy. A casual user may become a frequent user (or vice versa) following a change in a PT service or a new availability of other transport modes. Inferring the causes of these changes can help PT operators improve their services and design better customer retention strategies. To enable such analyses, the required first step is to identify instances of long-term, lasting changes in passengers’ repeated trip patterns from historical data. It is not practical to detect manually such changes from large amounts of data. Therefore, the goal of this study is to develop a method for automatically detecting long-term changes in repeated trip patterns of individual passengers using smart card data.

The development of automated fare collection systems has inspired many studies to investigate topics that are related to passenger behaviour. In the study by Bruno et al.[23], passenger groups are classified according to their traffic behaviours using smart card data and data mining techniques. Through the classification process, Bruno et al. outline key characteristics of traffic patterns within each passenger group. In the study by Bingfeng et al.[24], a subway route selection model is developed using smart card data from 2008 so as to solve the problem of high transit costs in Beijing. Mahrsi et al.[25] study the location of regular passengers’ domicile by extracting information from smart card data and using clustered socio-economic information. Mahrsi et al.’s study aimed to discover a new group of passengers that have similar characteristics of behaviours. Devillaine et al.[26] investigate the times of day when different passenger groups use smart cards within different cities. In addition, their study aims to uncover the underlying purposes the actions of passengers. Consequently, Devillaine et al. analyse changes in behavioural patterns in response to sociological, cultural and geopolitical differences by comparing their results from different cities.

In the study by Haitao et al.[27], passenger flows are analysed by using deep-learning models and smart card data (Chinese Shaoxing BRT IC card records) in order to analyse short-term bus rapid transit (BRT) performance. They use a Stacked Autoencoder model to extract the features of passenger flows. Following the extraction, they develop a model that is used to predict BRT passenger flows. During a 525-day period between the 25th of June 2013 to the 2nd of December 2014, smart card data was recorded. During the first 500 days, the Stacked Autoencoder model was trained and the final 25 days were used to test the model.
Even though the study’s objective was to predict the BRT performance and to analyse passenger flows, the study was not concerned with the long-term pattern changes of individual passengers. Similarly to the previous studies, Haitao et al. use also clustering techniques to collect and analyse transit passengers’ patterns from smart cards.

Most of the studies discussed use statistical data on all passengers that are taken from smart cards and apply clustering techniques to the data. However, the analysis of the long-term pattern changes of individual passengers is equally important in comparison to the statistical analysis of all passengers. In order to overcome the lack of research on the detection of long-term behaviour changes and the inefficiency of detecting these changes manually, this study proposes a model that uses an Autoencoder, which is a type of unsupervised deep learning model, and various similarity measurement methods to detect long-term changes in a passenger's movements.

2 Data Set
This study uses one year’s worth of Go-card data (the public transit smart card in Queensland Australia) that are provided by Translink (Queensland’s sole transit authority). This study uses the data to detect and classify the long-term temporal and spatial patterns and pattern changes of individual passengers. The data are collected from July 2015 through to June 2016 and include 2,528,460 cards and 151,682,564 transaction records. The data include information on operator (bus, rail, and ferry), route, smart card ID, passenger type, boarding location and time and alighting location and time [1, 2].

3 Methodology
This study adopts a machine learning approach to detect long-term changes in repeated trip patterns in an unsupervised manner. We extract features from the weekly trip patterns of each PT passenger and apply a clustering algorithm to identify major patterns in a passenger’s weekly PT usage over a long period of time. After applying a clustering algorithm, we apply a classification algorithm to each passenger in order to determine whether this passenger exhibits a long-term behavioural pattern change.

3.1 Framework
This study analyses the main patterns of individual passengers and the long-term changes in passengers’ movement. The term “main pattern” is used to refer to a representative weekly boarding pattern of a passenger that repeats over several weeks. For instance, a passenger who goes to work at 7:00 AM and leaves at 5:00 PM will continue this behaviour unless the company’s work hours change. In our study, we define a set of main patterns for each passenger by identifying a set of distinct weekly boarding patterns. A long-term behaviour change is then detected by identifying the condition where a passenger’s weekly boarding pattern has been changed from one main pattern to another main pattern. Figure 3-1 shows the overall structure of the proposed approach.
This section describes a framework for detecting passengers with long-term trip pattern changes. The framework consists of four modules:

- **Module 1** converts the Go-card data into a data format that is suitable for extracting temporal trip patterns of individual passengers.
- **Module 2** extracts the Compressed Weekly Pattern (CWP) of an individual passenger in order to characterize the trip pattern of the passenger for any given week. Module 2 determines whether there is a change in a passenger’s weekly trip pattern by comparing the CWPs between two consecutive weeks. Module 2 exacts a new main pattern (if found) to store in the passenger’s main pattern database. Main pattern refers to a representative pattern of similar CWPs and each main pattern has a class number.
- **Module 3** assigns each week a class label of the most similar main pattern by comparing the given week’s CWP and the main patterns in the main pattern database.
- **Module 4** determines if a given passenger shows a long-term behaviour change by looking at the assigned main patterns across all weeks.

The smart card data are input into our proposed model (Figure 3-1). We run all modules while we change the Similarity Ratio (SR) from one hundred percent to one percent at one percent intervals. The SR is a pre-defined similarity threshold that represents the minimum required similarity level (in percentage) for two CWPs to be considered “similar”. The SR is used to cluster CWPs into a few distinct main patterns. The details of each module are presented below.
Module 1. Changing Data Format

Module 1 extracts the boarding and alighting times, dates and smart-card IDs of individual passengers and creates a 168-by-\(n\) binary matrix of hourly transit usage for a week. The week is denoted by \(W\), for each passenger, where rows represent weeks (\(n\) is the total number of weeks in the data) and columns represent 1-hour time intervals over one week (168 is the number of hours per week). The value 1 is assigned to row \(i\) and column \(j\) if the passenger uses public transit on week \(i\) at a time interval \(j\); otherwise, the value is assigned 0. Figure 3-2 shows an example of the data transformation conducted in Module 1.

![Module 1 Data Format](image)

Figure 3-2: The data formatted by Module 1

Module 2. Extracting Weekly Main Patterns

Module 2 extracts weekly main patterns from passengers’ weekly movements. The main patterns are used as a basis for Module 3. This module consists of Module 2-1, which extracts the Compressed Weekly Pattern (CWP) and Module 2-2, which compares the similarity of CWPs with existing main patterns.

Module 2-1. Extracting Compressed Weekly Pattern by Autoencoder

The key function of Module 2-1 in our model is to extract key features and patterns from various passenger movements. To implement this function, we apply an unsupervised learning method called “Autoencoder.” An autoencoder is a deep neural network model that performs well in extracting features from data in an unsupervised manner. Through its hidden layer called the “bottleneck”, which has fewer neurons than the input or output layers [11], the autoencoder can compress the input data representation, reducing dimensions and extracting patterns. Using the autoencoder, we extract a Compressed Weekly Pattern (CWP) for each week of a given passenger by compressing weekly trip patterns over 5-consecutive weeks (two previous weeks + target week + two following weeks) into one representative weekly pattern, as shown in Figure 3-3. The CWP is a 168-by-1 feature vector characterizing a passenger’s weekly trip pattern by summarizing one 5-week group.

Let \(W_i\) be a 168-by-1 binary vector representing the \(i\)th column of \(W\) and \(S_i\) be a 168-by-5 matrix concatenating \(W_i\) over the five consecutive weeks as follows:

\[
5\text{-week group. } S_i = (W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2})
\]
The process of extracting the CWP entails expressing $S_i$ in terms of a 168-by-1 CWP vector that is denoted by $CWP_i$. The CWP is extracted in order to remove week-to-week temporary fluctuations (noise) and reveal the underlying systematic weekly trip patterns.

**Module 2-2: Identifying Main Patterns by Merging CWPs**

This module compares each pair of two consecutive CWPs across all $n$ weeks to identify a set of distinct CWP clusters over the observation period. In comparing two consecutive CWPs from the two consecutive 5-week groups, we call the first 5-week group “standard group” denoted by $S$ and call the second 5-week group “target group” denoted by $T$. For instance, for week $i$, we consider standard group $S_i = (W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2})$ and target group $T_i = (W_{i-1}, W_i, W_{i+1}, W_{i+2}, W_{i+3})$, where the alignment of the target group’s weeks to the standard group is delayed by one week.

Then we obtain the CWPs from the standard group and target group by compressing 168-by-5 matrices $S_i$ and $T_i$ into 168-by-1 binary feature vectors, denoted by $CWP_i^S$ and $CWP_i^T$, respectively.

Next, we measure the similarity between $CWP_i^S$ and $CWP_i^T$ using a pre-defined similarity metric. To measure the similarity between two CWPs, we test various similarity metrics, including Jaccard Similarity, symmetrical Dynamic Time Warping (DTW), asymmetrical DTW and Euclidean Distance [13, 14, 15, 16, 17, 18, 19, 20, 21, 22].

One of the most popular methods for calculating similarity is the Jaccard Similarity Coefficient and Euclidean Distance. Jaccard Similarity (Equation 3) is the size of the intersection between two data sets which is divided by the size of the union. The Jaccard Similarity has a value between 0 and 1.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$  \hspace{1cm} (Equation 3)
The Euclidean distance calculates the distance between two points \( p \) and \( q \) (Equation 4). The point \( p = (p_1, p_2, ..., p_n) \) and the point \( q = (q_1, q_2, ..., q_n) \) in the Cartesian coordinate system are denoted in Equation 4.

\[
\|p - q\| = \sqrt{(p - q) \cdot (p - q)} = \sqrt{\|p\|^2 + \|q\|^2 - 2p \cdot q} \quad \text{(Equation 4)}
\]

Another method used for measuring the similarity between two time-series arrays is Dynamic Time Warping (DTW). The time-series array data varies in time or in speed. Two sets of data can be compared because there is a non-linear scaling deviation between the two arrays (patterns) with the passage of time. In general, DTW distance performs a similarity search with greater precision than Euclidean distance when finding various distortions (warping), such as position shifts.

Given the measured similarity between two CWPs, the decision whether to cluster these two CWPs into one main pattern is based on a pre-defined similarity threshold, denoted by the Similarity Ratio (SR). The SR is defined as the minimum required similarity level in percentage for two CWPs to be considered “similar”. The lower the threshold value, the easier it is to merge two CWPs and we can obtain fewer and more aggregated final clusters (or main patterns).

Given the similarity metric chosen and a pre-defined SR value, Module 2-2 compares \( \text{CWP}_1^S \) from the standard group and \( \text{CWP}_1^T \) from the target group at each week \( i \). If the similarity between \( \text{CWP}_1^S \) and \( \text{CWP}_1^T \) is less than the required SR value, we consider that a passenger’s weekly boarding pattern in week \( i+1 \) (\( \text{CWP}_1^T \)) is significantly different from that in week \( i \) (\( \text{CWP}_1^S \)) and thus consider \( \text{CWP}_1^T \) as a candidate for a new main pattern. This candidate pattern is then matched against the existing main patterns in the main pattern database that have been identified from the previous weeks. If the pattern represented by \( \text{CWP}_1^T \) does not exist in the current main pattern database, we add this candidate to the dataset as a new main pattern and assign a class number to this new main pattern. If the similarity between \( \text{CWP}_1^S \) and \( \text{CWP}_1^T \) is greater than the SR, these two CWPs are considered to be similar and no new main pattern is created. This procedure of comparing \( \text{CWP}_1^S \) and \( \text{CWP}_1^T \) and creating a new main pattern (if found) is repeated for all weeks \( i \) by looping from week 1 to week \( n \). Figure 3-4 shows examples of main patterns that are identified from this procedure and Figure 3-5 shows the pseudocode for implementing this procedure in Module 2.

Figure 3-4: The Main Pattern Samples
The number of main patterns added to the main pattern database at the end of the loop depends on the SR value used. The lower the SR value is, the smaller the number of main patterns remaining in the database. To facilitate the detection of the most significant change between these main patterns, we can identify the SR threshold value that produces the two most aggregated main patterns. In other words, we can detect the most significant change by reducing the SR threshold value from the highest to the lowest value until only two main patterns remain. In this study, we are interested in identifying passengers who exhibit clear and significant changes in their weekly boarding patterns and detecting the single most significant change in each individual (if any) to facilitate the identification of such passengers with long-term behaviour changes. The detection of the single most significant change in a passenger’s weekly boarding patterns is done by repeating the Module 2 procedure shown in Figure 3-4 at each value of SR by changing the SR from 100% to 0% until two of the most aggregate main patterns remain. The point in time where the pattern changes from the first main pattern to the second main pattern is where the most significant change in the passenger’s trip pattern occurs. If a passenger does not have such a clear shift in the trip pattern, the algorithm will not stop but continue until the SR reaches 0% without finding any two remaining main patterns. Such passengers will end up having one aggregate main pattern cluster at the SR = 0%.

Module 3. Assigning Each Week to the Closest Main Pattern Class
Module 3 assigns a class label to each week by using the most similar main pattern. Module 3 uses the most similar main pattern by comparing the actual data of each week and the main patterns in the main pattern database. Module 3 uses the similarity metrics that were generated in Module 2-2. Module 3 functions to classify the main pattern that is closest to each real weekly pattern \( W \) by using the similarity metrics. Module 3 assigns all the weekly records in the 168-hour table to a class number from the main pattern database. As shown in Figure 3-6(a), the weekly data do not have an assigned class number. However, in Figure 3-6(b), module 3 assigns each class number to that week.
Module 4. Classifying Passengers with Long Term Changes

Module 4 determines if a given passenger shows a long-term behaviour change by looking at the assigned main patterns across all weeks. For instance, when the SR is at 9%, as shown in Figure 4-3, two long-term behaviour patterns of an individual passenger are identified, where the long-term behaviour pattern is defined as the weekly pattern of a passenger’s movements that lasts more than $k$ weeks and until the class pattern changes. In this study, $k=4$ is used indicating that a pattern is considered as a long-term behaviour pattern if it lasts longer than 4 weeks (1 month).

4 Experiments and results

4.1 Case study

In this section, we demonstrate how the proposed long-term change detection framework works using the results obtained from a case study.

Case 1. Adult Passenger (ID = 175)

Figure 4-1 shows the results of our model, which identifies a single passenger with long-term trip pattern changes. Figure 4-1(a) presents the passenger’s weekly boarding pattern that is obtained from the actual data over a one-year period. The horizontal axis represents days and the vertical axis represents time-of-day. Blue vertical lines are added to indicate the start of each week, showing a total of 52 weeks over the study period. Black dots represent the times of boarding. This passenger is a regular PT user who uses the system twice a day, almost every day (the first around 9-11am and the second around 8-10pm). Figure 4-1(b) shows the corresponding compressed patterns (main patterns). The main patterns are obtained by removing noise from the original data and by extracting representative weekly boarding patterns by clustering and classifying the original data. The algorithm detects two distinct boarding pattern classes, namely Main Pattern 1 from Week 1 to Week 15 and Main Pattern 2 from Week 18 onwards. The representative weekly boarding patterns for these two classes are also presented in Figure 4-1. Based on these simplified patterns, our framework detects automatically that this passenger changes his PT usage patterns at Week 17. The visual inspection of the original data allows us to understand that the passenger becomes more regular in his movement. That is to say, the variations in departure times across different days of the week decrease at Week 17. The departure times are also moved to earlier time windows (i.e., 9-11am $\rightarrow$ 9am and 8-10pm $\rightarrow$ 6pm).

As illustrated in this case study, the proposed framework scans millions of users’ data within the smart card database. In addition, the framework identifies passengers who made systematic and lasting changes in their PT usage patterns. The framework
identifies these changes automatically. The ultimate goal of using the framework is to generate a labelled dataset that contains detailed information on the locations and times of any systematic changes in passengers’ long-term trip patterns. The labelled datasets can support further investigation into the causes for trip pattern changes. As a result, PT officials will better understand passengers’ responses to implemented policies. Consequently, PT officials will be able to design better customer retention strategies.

Figure 4-1: An illustrative example of a passenger’s long-term behaviour change and the extracted patterns that were identified by the proposed framework.

The results shown in Figure 4-2 were found from Case 1 with several similarity methodologies (Jaccard, Euclidean, Dynamic Time Warping Symmetric and Dynamic Time Warping Asymmetric). The representative weekly boarding pattern 1 (Main Pattern 1) is presented on the left of Figure 4-2. Main Pattern 2 is presented on the right-hand side of Figure 4-2. Figure 4-2(b) shows the results of using Jaccard Similarity in our model. Figure 4-2(c) shows the results of using the Euclidean-based model. Figure 4-2(d) and (e) show the results of using the symmetric DTW and the asymmetric DTW respectively.

As shown in Figure 4-2(b) and (c), the model that uses Jaccard Similarity and Euclidean Similarity detects that a change in long-term pattern occurs at Week 17. On the other hand, although the DTW-based models detect two remaining main patterns, those models do not detect long-term pattern changes. The DTW Similarity compares the similarities of a given day with the preceding and following days. On the other hand, Jaccard Similarity and Euclidean Similarity compare similarities of a given day only. As a result, the model that uses Jaccard Similarity and Euclidean
Similarity is more effective than the model using DTW because the former methods detect long-term behaviour changes.

![Figure 4-2: Detection Result of Case 1](image)

Table 4-1 shows the percentage of similarity between the actual weekly behaviour and its corresponding main pattern of a transport user. The first column in Table 4-1 shows the similarity metric that is applied in Module 2. The second column shows the SR that produced two main patterns. The third column shows the similarity metric that is used in Module 3. From the fourth to the seventh column, the similarity measures between the actual weekly pattern data and their corresponding main pattern (the representative pattern of their assigned main pattern classes) are presented. The higher this value is, the better the corresponding main pattern represents the underlying data to which this main pattern is assigned. The last column of the table shows the average of the four similarity values in column 4 through 7.

Based on the last column of Table 4-1, it is found that the Jaccard Similarity metric produces the highest similarity between the actual data and the extracted main patterns. Case 1 shows that the Jaccard Similarity has the best performance because it has the highest percentage average of similarity between the actual week patterns and the main patterns among the used similarity methods. In Case 1, Jaccard Similarity detects long-term behaviour changes. The Jaccard and Euclidean similarity metrics detect two remaining main patterns at a low similarity ratio (9% SR at Jaccard, 4% SR at Euclidean). On the other hand, the DTW-based model detects two main patterns at a high SR. The identified two main patterns suggest that the DTW-based model does not detect a long-term behaviour pattern change because it fails to detect two lasting patterns but instead captures a short-term fluctuation as a main pattern.
Similarity metric used in Module 2

<table>
<thead>
<tr>
<th>Similarity metric used in Module 2</th>
<th>Similarity metric used in Module 3</th>
<th>Similarity between actual weekly patterns and assigned main patterns</th>
<th>Average of (4) – (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>Jaccard</td>
<td>Jaccard</td>
<td>Euclidean</td>
</tr>
<tr>
<td>9 %</td>
<td>54 %</td>
<td>36 %</td>
<td>88 %</td>
</tr>
<tr>
<td>Euclidean</td>
<td>Euclidean</td>
<td>49 %</td>
<td>34 %</td>
</tr>
<tr>
<td>DTW(s)</td>
<td>DTW(s)</td>
<td>26 %</td>
<td>17 %</td>
</tr>
<tr>
<td>DTW(a)</td>
<td>DTW(a)</td>
<td>26 %</td>
<td>16 %</td>
</tr>
</tbody>
</table>

Table 4-1: The similarity between the actual week patterns and the main patterns

Figure 4-3 shows the process of detecting the two remaining main patterns according to the Jaccard Similarity. The figure shows how the noise is removed according to the value of the Jaccard-based Similarity Ratio (JSR). A JSR between 40% and 10% detects more than two main patterns while a JSR of 9% detects two remaining main patterns. Although the figure shows that the same long-term change is detected with a similarity ratio between 9% to 30%, there is a more stable long-term behaviour pattern change at a 9% similarity ratio.

Figure 4-3: Detection Result of Case 1 according to Jaccard-based Similarity Ratio (JSR)

Case 2. Adult Passenger (ID = 99)

Case 2 shows the detection of a long-term change at a higher JSR (20% JSR) than Case 1. As shown in Figure 4-4(a), we confirm visually that the long-term behaviour pattern of a public transport user changes at week 6. In Figure 4-4(b) and (c), Jaccard and Euclidean-based models detect the same date to each other of the long-term change. The shape of each main pattern is the same as well. However, as shown in Table 4-2, the Jaccard Similarity shows a higher similarity than the Euclidean Similarity. Figure 4-4(a) and (b) show that the pattern changes temporarily from Week 24 to Week 26 in the shape of Main Pattern 1. The Jaccard Similarity detects this intermediate change more effectively than the Euclidean
version. Figure 4-4(d) shows that Main Pattern 1 and Main Pattern 2 are repeated multiply without showing long-term behaviour stability. Figure 4-4(e) shows that the long-term behaviour pattern continues in the same shape as Main Pattern 2.

![Figure 4-4: Detection Result of Case 2](image)

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity metric used in Module2</td>
<td>Similarity Ratio (SR)</td>
<td>Similarity metric used in Module3</td>
<td>Similarity between actual weekly patterns and assigned main patterns</td>
<td>Jaccard</td>
<td>Euclidean</td>
<td>DTW(s)</td>
<td>DTW(a)</td>
</tr>
<tr>
<td>Jaccard</td>
<td>20 %</td>
<td>Jaccard</td>
<td>46 %</td>
<td>30 %</td>
<td>74 %</td>
<td>87 %</td>
<td>59 %</td>
</tr>
<tr>
<td>Euclidean</td>
<td>11 %</td>
<td>Euclidean</td>
<td>40 %</td>
<td>27 %</td>
<td>61 %</td>
<td>94 %</td>
<td>56 %</td>
</tr>
<tr>
<td>DTW(s)</td>
<td>70 %</td>
<td>DTW(s)</td>
<td>35 %</td>
<td>23 %</td>
<td>84 %</td>
<td>89 %</td>
<td>58 %</td>
</tr>
<tr>
<td>DTW(a)</td>
<td>80 %</td>
<td>DTW(a)</td>
<td>23 %</td>
<td>15 %</td>
<td>76 %</td>
<td>75 %</td>
<td>47 %</td>
</tr>
</tbody>
</table>

Table 4-2: The similarity between the actual week patterns and the main patterns

**Case 3. Adult Passenger (ID = 101)**
The similarity ratio in both Case 3 and Case 2 resemble each other (21% SR with Jaccard Similarity). This resemblance shows a long-term pattern change for only a few weeks. In Figure 4-5, the public transport user has a pattern of boarding at 5am and 6pm from week 1. The boarding pattern changes at week 16 where the passenger boards mainly at 5am and 4pm. From week 17, the boarding pattern returns to its previous state of boarding at 5am and 6pm. In Figure 4-5(b) and (c), Jaccard and Euclidean Similarity Ratios are used to detect the intermediate long-term changes, while in Figure 4-5(d) and (e), DTW Similarity Ratios do not detect these intermediate changes.
Table 4-3 shows that the result of Jaccard Similarity is more similar to the actual week patterns than the Euclidean Similarity.

<table>
<thead>
<tr>
<th>Similarity metric used in Module2</th>
<th>Similarity Ratio (SR)</th>
<th>Similarity metric used in Module3</th>
<th>Similarity between actual weekly patterns and assigned main patterns</th>
<th>Average of (4) – (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>21 %</td>
<td>Jaccard</td>
<td>62 % 43 % 88 % 98 % DTW(s) DTW(a)</td>
<td>73%</td>
</tr>
<tr>
<td>Euclidean</td>
<td>11 %</td>
<td>Euclidean</td>
<td>59 % 41 % 78 % 99 % DTW(s) DTW(a)</td>
<td>69%</td>
</tr>
<tr>
<td>DTW(s)</td>
<td>80 %</td>
<td>DTW(s)</td>
<td>55 % 38 % 90 % 98 % DTW(a)</td>
<td>70%</td>
</tr>
<tr>
<td>DTW(a)</td>
<td>90 %</td>
<td>DTW(a)</td>
<td>64 % 47 % 95 % 95 % DTW(a)</td>
<td>75%</td>
</tr>
</tbody>
</table>

Table 4-3: The similarity between the actual week patterns and the main patterns

**Case 4. Adult Passenger (ID = 5)**

Case 4 shows Jaccard Similarity has the highest ratio (30% SR) among the cases mentioned above. As shown in Figure 4-6, the public transport user is boarding mainly at 7am and 3pm from Week 1 to Week 41. This pattern changes at Week 42. Case 4 has more noise than the previous cases. Therefore, Main Pattern 2 of the model that uses Jaccard Similarity, Figure 4-6(b), and the model that uses Euclidean Similarity, Figure 4-6(c), are different. However, both models detect the same long-term change. In addition, as shown in Table 4-4, the model that uses Jaccard Similarity shows a higher correspondence with the actual week patterns than the model that uses Euclidean Similarity.
Figure 4-6: Detection Result of Case 4

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity metric used in Module 2</td>
<td>Similarity metric used in Module 3</td>
<td>Similarity between actual weekly patterns and assigned main patterns</td>
<td>Jaccard</td>
<td>Euclidean</td>
<td>DTW(s)</td>
<td>DTW(a)</td>
<td>Average of (4) – (7)</td>
</tr>
<tr>
<td>Jaccard</td>
<td>Jaccard</td>
<td>Jaccard</td>
<td>45 %</td>
<td>28 %</td>
<td>82 %</td>
<td>94 %</td>
<td>62%</td>
</tr>
<tr>
<td>Euclidean</td>
<td>Euclidean</td>
<td>Euclidean</td>
<td>45 %</td>
<td>29 %</td>
<td>68 %</td>
<td>94 %</td>
<td>59%</td>
</tr>
<tr>
<td>DTW(s)</td>
<td>DTW(s)</td>
<td>DTW(s)</td>
<td>40 %</td>
<td>25 %</td>
<td>88 %</td>
<td>94 %</td>
<td>62%</td>
</tr>
<tr>
<td>DTW(a)</td>
<td>DTW(a)</td>
<td>DTW(a)</td>
<td>42 %</td>
<td>26 %</td>
<td>87 %</td>
<td>86 %</td>
<td>60%</td>
</tr>
</tbody>
</table>

Table 4-5: The similarity between the actual week patterns and the main patterns

Case 5. Regular Long-Term pattern without Long-Term change

Case 5, as shown in Figure 4-7, does not have a long-term change but shows a regular boarding pattern. In this case, two remaining main patterns are not detected by any similarity ratio. However, the similarity ratio can distinguish temporary passengers from regular passengers.

Figure 4-7: Weekly Data of Case 4

In the case studies described in the report, the most effective method for detecting a long-term pattern change among the four Similarity metrics is the Jaccard Similarity. Furthermore, the DTW-applied models detect two main patterns. However, the DTW versions are not efficient as they do not detect two long-term behaviour patterns.
5 SUMMARY AND CONCLUSIONS

This study proposes a model that recognizes automatically a passenger's weekly pattern from smart card data and detects significant long-term changes in a passenger's weekly pattern. Our model uses an autoencoder so as to recognize the characteristics of a passenger’s weekly pattern. Additionally, the autoencoder extracts the long-term weekly pattern behaviour of the passenger. Following this, our model uses a similarity measurement method and similarity comparison criteria in order to detect significant long-term changes. We compare the performance of the model in its ability to detect long-term changes when four different types of similarity metrics are used. The similarity metrics used are Jaccard Similarity, Euclidean Distance, Dynamic Time Warping Symmetric and Dynamic Time Warping Asymmetric. The result of the performance comparison shows that the Jaccard Similarity and the Euclidean Distance are able to detect effectively long-term behaviour changes. In particular, Jaccard Similarity shows the best detection performance out of the four methods. Consequently, our model detects automatically the long-term changes of individual passengers’ movements and removes a passenger’s short-term, transient fluctuations. Further research is planned as an extension to this analysis in order to investigated long-term behaviour changes in boarding and alighting locations.

References


