

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

12-2018

Using smart card data to model commuters' responses upon unexpected train delays

Xiancai TIAN


Singapore Management University, shawntian@smu.edu.sg

Baihua ZHENG

Singapore Management University, bhzheng@smu.edu.sg

DOI: <https://doi.org/10.1109/BigData.2018.8622233>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), [Numerical Analysis and Scientific Computing Commons](#), and the [Transportation Commons](#)

Citation

TIAN, Xiancai and ZHENG, Baihua. Using smart card data to model commuters' responses upon unexpected train delays. (2018). *2018 IEEE International Conference on Big Data: Seattle, WA, December 10-13: Proceedings*. 831-840. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/4208

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Using Smart Card Data to Model Commuters’ Responses Upon Unexpected Train Delays

Xiancai Tian, Baihua Zheng

Living Analytics Research Centre, Singapore Management University, Singapore
{shawntian, bhzheng}@smu.edu.sg

Abstract—The mass rapid transit (MRT) network is playing an increasingly important role in Singapore’s transit network, thanks to its advantages of higher capacity and faster speed. Unfortunately, due to aging infrastructure, increasing demand, and other reasons like adverse weather condition, commuters in Singapore recently have been facing increasing *unexpected train delays (UTDs)*, which has become a source of frustration for both commuters and operators. Most, if not all, existing works on delay management do not consider commuters’ behavior. We dedicate this paper to the study of commuters’ behavior during UTDs. We adopt a data-driven approach to analyzing the six-month’ real data collected by automated fare collection system in Singapore and build a classification model to predict whether commuters switch from MRT to other transportation modes because of UTDs.

Index Terms—Mass Rapid Transit; unexpected train delays; smart card data; trip chains; individual travel patterns; clustering; DBSCAN; feature engineering; response modeling; feature insights

I. INTRODUCTION

Mass public transport is by far the most efficient mode of transport, in terms of both land and energy use. For land-scarce countries such as Singapore, it is critical and extremely important to improve the public transport in order to meet the increasing travel demands of a growing economy and population. The mass rapid transit (MRT) system, built by the Land Transport Authority (LTA) of Singapore, consists of 5 MRT lines, namely *North South Line (NSL)*, *East-West Line (EWL)*, *Circle Line (CCL)*, *North East Line (NEL)* and *Downtown Line (DTL)*, as shown in Figure 1. The MRT network is playing an increasingly important role in Singapore’s transit network, thanks to its advantages of higher capacity and faster speed. It has reached a daily average ridership of over 3 million by 2016 [1].

Unfortunately, due to aging infrastructure, increasing demand, and other reasons like adverse weather condition, commuters in Singapore recently have been facing increasing *unexpected train delays (UTDs)*, which has become a source of frustration for both commuters and operators. Table I outlines the number of UTDs happened each year from 2011 to 2017.

Service recovery is a top priority for transport agencies in the event of UTD, and there are many studies on the management of UTD [2, 3, 4, 5, 6]. In the literature, a few studies have examined travel behavior changes under severe UTDs, which have been shown to generate both short-term and

TABLE I: Number of Train Services Delays (> 30 minutes)

	Total Count	NSL	EWL	NEL	CCL	DTL
2011	9	4	1	1	3	-
2012	8	0	3	2	3	-
2013	7	1	2	3	1	-
2014	10	4	1	2	1	2
2015	15	5	3	4	2	1
2016	16	4	5	3	3	1
2017	16	6	5	3	2	0

^aData Source: Performance of Rail Service Reliability, LTA

long-term changes in commuter behavior [7, 8]. In the short term, depending on the trip purpose, commuters may change routes, modes or trip departure times. In the long term, more crucial changes may take place. For example, [9] states that major network disruptions may affect location decisions of residence or work; [10] reveals that the occurrence of major network disruptions provides good opportunities for travelers to experience alternative modes and possibly use them in the future.

Nevertheless, most UTD management studies do not consider commuter behavior explicitly and empirically. There are some studies on behavioral issues of commuters, based on certain assumptions instead of empirical studies. These assumptions, while generally accepted to be logical, are not verified or compared with behavioral studies to confirm the validity and accuracy. In other words, commuter behavior in the event of UTD has not been well studied. As a result, there is still room to further improve delay management and service recovery. In this article, we adopt a data-driven approach to analyzing smart card data collected in Singapore and develop a classification model to predict whether commuters switch from MRT to other transportation modes because of UTDs. The findings aim to help MRT operators to arrange more targeted remedial actions and to provide more personalized travel assistances to reduce commuters’ unhappiness and to help commuters resume their trips smoothly.

In order to achieve above objectives, we have to fully understand how commuters respond to a UTD. In other words, given one occurrence of UTD, we need to find out *who* are/will be affected, *how* they respond, and *why* they do so. To tackle these three challenging research questions, we strategically focus our study on *regular commuters* whose travel patterns are more predictable, e.g., a student who travels to school every weekday morning around 7:00 am and a working adult



Fig. 1: Singapore MRT Network Map

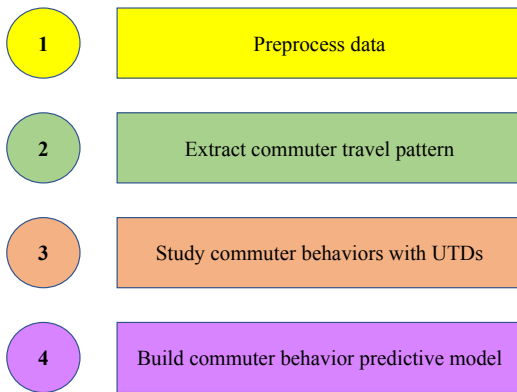


Fig. 2: Procedures of Our Methodology

who travels from office to home every weekday afternoon around 6:30 pm except Friday. We introduce the concept of regular commuters, and develop data mining approaches to extract travel patterns, if any, for each individual commuter by considering patterns in both *spatial dimension* as well as *temporal dimension*. The notion of regular commuter and his/her travel patterns provides a solution to the *who* question. Once we are able to locate the commuters who are/will be affected by the given UTD, we conduct a comparison study to compare their behaviors upon UTD and their normal behaviors to answer the *how* question. We further identify multiple features, including both global features and individual features, that affect the commuters' responses to UTD and build a classification model to perform the prediction. That answers the *why* question.

Accordingly, we propose a four-step procedure to build our predictive model, as depicted in Figure 2. In the first step, we perform data preprocessing, which will be detailed in Section II. In the following three steps, we perform various data analytics studies to answer *who*, *how* and *why* questions, to be presented in Section III, Section IV and Section V respectively.

II. DATA PREPROCESSING

The majority of exploratory travel pattern studies use travel survey data as the principal data source. However, survey data generally have limitations, such as small sample sizes, high cost, low response rates and inaccurate travel behaviour information [11, 12]. Therefore, alternative data sources are required to be able to more accurately and more comprehensively understand the spatial-temporal characteristics of travel patterns. The implementation of an automated fare collection (AFC) system allows public transport agencies to collect large quantities of data that record passengers' activities with detailed time and space information. It has been recognized that there are large potential benefits of using AFC data to improve public transport planning and operation [13].

EZ-Link card is the smart card used in Singapore for the payment of public transport. In our study, we analyze the data captured by EZ-Link from December 1st 2015 to May 31st 2016, which covers *all* the bus/MRT rides taken in Singapore within that six months' duration, in total 881,012,319 rides. The total number of cards is 8,557,776. Note the population of Singapore in 2016 was around 5.6 million.

As listed in Table II, each EZ-Link record is corresponding to one MRT/bus ride, including the boarding and alighting

TABLE II: Smart Card Data Sample

card id	type	mode	entry date-time	exit datetime	origin id	destination id
02***5F	adult	BUS	2016-01-25 08:00:59	2016-01-25 08:14:43	1001	1117
02***5F	adult	MRT	2016-01-25 08:20:04	2016-01-25 08:27:27	35	12
02***5F	adult	MRT	2016-01-25 18:13:57	2016-01-25 18:21:25	12	35
02***5F	adult	BUS	2016-01-25 18:26:57	2016-01-25 18:41:25	1118	1002
02***5F	adult	BUS	2016-01-26 07:57:24	2016-01-26 08:10:23	1001	1117
02***5F	adult	MRT	2016-01-26 08:13:51	2016-01-26 08:21:21	35	12
02***5F	adult	MRT	2016-01-26 18:31:45	2016-01-26 18:38:11	12	35
02***5F	adult	BUS	2016-01-26 18:47:45	2016-01-26 19:01:16	1118	1002

TABLE III: Smart Card Dataset Attributes

Attribute	Notation	Description
card id	c_{id}	unique identifier of a smart card
type	$type$	commuter type (i.e., child, adult, senior)
mode	$mode$	transport mode of the ride
entry date	$date_{in}$	starting date of a ride
exit date	$date_{out}$	ending data of a ride
entry time	t_{in}	starting time of a ride
exit time	t_{out}	ending time of a ride
origin id	id_{in}	unique identifier of the origin MRT station/bus stop
destination id	id_{out}	unique identifier of the destination MRT station/bus stop

MRT station/bus stop and the corresponding timestamps. Other information such as travel mode and passenger types are also recorded. Apart from that, each smart card is associated with an encrypted unique identifier, so that we can identify all the rides taken by one commuter with commuter's real identity being well protected. Table III lists the attributes captured by each EZ-Link record.

We mainly perform two tasks in this first step, namely *noisy data removal* and *trip chain generation*.

Noisy Data Removal. Due to AFC system deficiency and other technical limitations, some rides are not properly captured. Three types of noisy data are removed before we proceed with following analysis: 1) duplicate records for the same ride; 2) rides with impossible travel duration (e.g., rides with duration longer than 5 hours); and 3) records with missing values. In total, 44,050,616 records are removed which correspond to 5.0% of the complete dataset. As the noisy data is significantly smaller than the valid data, we assume that the removal of those noisy records will not bias our analysis.

Trip Chain Generation. Commuters take public transport

TABLE IV: Rides vs. Trip Chains

No. of rides	No. of trip chains	% of trip chains
1	459,374,676	72.5%
2	145,732,657	23.0%
3	24,711,189	3.9%
≥ 4	3,801,721	0.6%

TABLE V: Trip Chain Sample

card id	boarding time	origin id	alighting time	destination id
02***5F	2016-01-25 08:00:59	1001	2016-01-25 08:27:27	12
02***5F	2016-01-25 18:13:57	12	2016-01-25 18:41:25	1002
02***5F	2016-01-26 07:57:24	1001	2016-01-26 08:21:21	12
02***5F	2016-01-26 18:31:45	12	2016-01-26 19:01:16	1002

because of certain travel demands, e.g., traveling from home to office. However, not all the travel demands could be satisfied by a single ride. Consequently, there is a need to link consecutive rides that actually serve the same travel demand together. Accordingly, we introduce the concept of *trip chain* [14], which is defined as a series of rides taken by a commuter on a regular basis (e.g., daily) and is considered a useful way to demonstrate travelers' behaviors. Parameters τ and δ could be utilized to differentiate various trip chains in this study, where τ specifies a time interval and δ specifies a distance threshold (e.g., $\tau = 30$ minutes, and $\delta = 500$ meters in our study).

To be more specific, given a sequence of rides of a commuter sorted according to chronological order, let r_i and r_{i+1} be two consecutive rides taken by the commuter on the same day. If $r_{i+1}.t_{in} - r_i.t_{out} \leq \tau$ and $|r_{i+1}.l_{in} - r_i.l_{out}| \leq \delta$, then r_i and r_{i+1} are considered to be part of the same trip chain. Here, $r.t_{in}$ and $r.t_{out}$ represent the boarding time stamp and the alighting time stamp of a ride r respectively and $r.l_{in}$ and $r.l_{out}$ stand for the boarding location and the alighting location (i.e., the location of the corresponding MRT station/bus stop) respectively. Operation $|l_1 - l_2|$ is to return the distance between two locations. After this step, we have generated 633,620,243 trip chains from 836,961,703 valid rides. On average, each trip chain consists of 1.32 rides. The distribution of rides constituting trips is reported in Table IV.

For example, the first two rides in Table II, denoted as r_1 and r_2 , are merged into one trip chain, as listed in the first record of Table V. This is because $r_2.t_{in} - r_1.t_{out} = 08:04 - 08:14:43 \leq \tau$ and the physical location of bus stop 1117 (i.e., $r_1.l_{out}$) and that of MRT station 35 (i.e., $r_2.l_{in}$) are close to each other (within δ meters).

In the following steps, we study trip chains (in short trips) instead of raw rides. Given a trip chain T consisting of j rides r_i with $i \in [1, j]$, $T.l_{in}$ and $T.l_{out}$ record its origin and destination MRT stations or bus stops respectively, and $T.t_{in}$ and $T.t_{out}$ capture the starting time stamp and ending time stamp of the trip respectively. To be more specific,

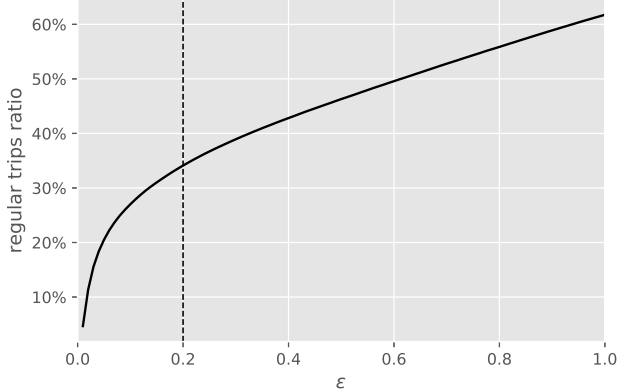


Fig. 3: Sensitivity Analysis of ϵ

$T.l_{in} = r_1.id_{in}$, $T.l_{out} = r_j.id_{out}$, $T.t_{in} = r_1.t_{in}$, and $T.t_{out} = r_j.t_{out}$. Rides shown in Table II are merged to generate four trip chains shown in Table V.

III. COMMUTER TRAVEL PATTERNS EXTRACTION

The word “commuting” by definition refers to the activity of traveling *regularly* between work and home. In other words, many periodically recurring commutings are performed because of certain purposes (e.g., work and school). As the travel purpose is hidden, we use the notion of *travel pattern* to summarize a sequence of regular trips made by one commuter that demonstrate similarity in both the temporal dimension and the spatial dimension. The temporal similarity and the spatial similarity are to guarantee that all the trips fallen within one travel pattern shall happen roughly in the same time window and have their origins (and their destinations) very close to each other if not identical. In this step, our objective is to extract the travel patterns of different commuters via data clustering. In the following, we first introduce the clustering algorithm we adopt and then present the clustering results. It is worth noting that we focus on the trips commuters made in weekdays only and exclude the trips made in weekends or public holidays from this study. This is because for most, if not all, commuters, their weekday travel patterns are very different from their weekend and public holiday travel patterns. However, our approaches, to be presented in the following, remain applicable when we study the travel patterns in weekend.

A. Clustering Algorithm

Given one trip, we know its starting time stamp t_{in} , ending time stamp t_{out} , the location of the boarding MRT station/bus stop in the form of latitude lat_{in} and longitude lon_{in} , and the location of the alighting MRT station/bus stop again in the form of latitude lat_{out} and longitude lon_{out} . Accordingly, we represent a trip via a 8-tuple vector $(t_{in}, t_{out}, t_{in}, t_{out}, lat_{in}, lat_{out}, lon_{in}, lon_{out})$. We purposely duplicate temporal attributes to make temporal attributes and spatial attributes both span four dimensions

and both are given equal importance when calculating the difference between two distinct trips. In addition, the eight dimensions are normalized to make attributes with various units and various scales comparable.

In term of clustering algorithm, we adopt DBSCAN [15]. Unlike most non-hierarchical clustering algorithms, DBSCAN algorithm does not require the input of the number of clusters which is unknown to us. DBSCAN relies on two parameters to perform the clustering, namely ϵ and $MinPts$. Distance ϵ defines a density-reachable range. Given an existing cluster C_i , if a sample record falls within the circular range defined by the center of C_i and ϵ , it will be included into the cluster C_i . Threshold $MinPts$ defines the minimum number of records to form a cluster; the final clusters with the number of records smaller $MinPts$ are marked as noise. The closer the records are to each other, the more the likely that those records are clustered into a cluster by DBSCAN. Outliers are often distant from other dense records, so DBSCAN is able to detect these outliers. We refer interested readers to [15] for more details about DBSCAN.

As the number of commuters is very large, we cannot afford to fine-tune $MinPts$ and ϵ for each individual commuter. Consequently, we set $MinPts$ and ϵ as global parameters. Larger $MinPts$ values are usually better for data sets with noise and will yield more significant clusters. As a rule of thumb, $MinPts = 2 \times D$ (i.e., 16) is used in our study [16]. As for ϵ , if its value is very small, a large part of the data will not be clustered; if its value is very big, the number of clusters will be small and majority of data records will be in the same cluster. ϵ of the “sparsest” cluster is a good candidate for this global parameter specifying the lowest density which is not considered to be noise. Consequently, we conduct a sensitivity analysis to select a proper value of ϵ .

To be more specific, we run DBSCAN on a subset of commuters for a range of values of ϵ , and record the ratio of regular trips to the total number of trips under each value of ϵ . We then plot a line graph of the ratio for each ϵ value, as shown in Figure 3. The ratio of regular trips grows very fast when ϵ increases from 0 to 0.2, after that the line becomes flatter. This observation implies that most of trip clusters have their density values within the range of 0 to 0.2. The upper bound of the density value is 0.2, under which 34.1% of the trips are considered as regular trips. Consequently, ϵ value of 0.2 would be a good candidate to represent the density of the “sparsest” cluster.

B. Clustering Results

By clustering historical trips of each commuter using DBSCAN, we extract travel patterns at an individual level. Take trip chains listed in Table V as an example. Based on clustering result, two travel patterns of this commuter are extracted, which very likely correspond to the home-to-work/home-to-school trip every morning and work-to-home/school-to-home returning trip in the evening.

We name a commuter with at least one identified travel pattern as a *regular commuter*, differentiated from commuters

TABLE VI: Commuters vs. Travel Patterns

No. of travel patterns	No. of commuters	% of commuters
1	730,676	14.11%
2	844,804	16.76%
3	328,619	6.62%
4	197,118	3.97%
≥ 5	133,051	2.65%

without any travel pattern (e.g., tourists); and we name a trip made by a regular commuter as a *regular trip* if it is captured by one identified travel pattern. Our following study only focuses on regular trips made by regular commuters but excludes all the irregular or adhoc trips. In a summary, 44.1% of commuters are identified as regular commuters and 34.0% of the trips are identified as regular trips. The number of identified travel patterns corresponding to each commuter varies, as reported in Table VI. Note the total number of regular commuters reaches 2.23 million, close to 40% of the population. Many regular commuters have either one or two identified travel patterns and very few commuters have more than four travel patterns. This is consistent with our assumption that traveling to and from work/school is the main reason for urban travel, and constitutes the primary demand for public transport services. In addition, commuters are different in terms of the degree of regularity. Some commuters travel very regularly with almost all the trips being regular, other commuters travel much less regularly with only a small portion of trips being regular.

IV. STUDY COMMUTER BEHAVIOURS WITH UTD

UTDs, especially those causing significant delays (e.g., ≥ 30 minutes), affect commuters’ travel experience. In this era of digital age, effective information dissemination provides one way to reduce commuters’ unhappiness and ease the crowds during UTD. However, it is critical to send the right information to the right commuters at the right time, otherwise irrelevant or not-so-useful information could easily overload the commuters which will only worsen the commuters’ frustration.

Our study presented in this paper tries to address the above issue via investigating the behaviors of regular commuters when their regular trips are affected by UTD. The question we want to answer is “*whether UTDs affect the regular MRT commuters’ choice of the means of transportation*”. For those commuters whose answer is NO, they still travel via MRT even during UTD, although the rides might take longer time and the travel experience might be less comfortable. Information such as the real-time update on the UTD, the schedule and boarding locations of the bridging buses that could bring them to the nearby MRT stations which are not affected by the UTD, and the estimation of the delays caused by the UTD becomes very relevant. On the other hand, for those commuters whose answer is YES, they will switch to other transportation modes such as buses, taxis, bike or even walking. Accordingly, information such as the availability of taxis, the locations

TABLE VII: Train Delay Dataset

Attribute	Notation	Description
MRT line	$line$	MRT line affected
starting station	s_{start}	starting station of the UTD
ending station	s_{end}	ending station of the UTD
one-way	$flag$	flag variable indicating whether a train delay is one-way
starting date & time	t_{start}	starting date & time of the UTD
ending date & time	t_{end}	ending date & time of the UTD
cause	$cause$	cause of the delay

of nearby taxi stand, alternate bus routes that can bring commuters to their destination (e.g., the schedules and the way to the right bus stops), the walking routes to the destinations if they are within walking distance, and the location of the parking zones where shared bikes are available becomes more useful. In other words, the relevance and usefulness of the information highly relies on how commuters respond to the UTD, and our study presented in this paper fills in the gap between the information and the commuters.

Our analysis is based on 34 MRT service delays happened in weekdays in Singapore from December 1st 2015 to May 31st 2016, the same time window as our EZ-Link dataset. Note that this number is larger than the number of UTD reported in Table I as we consider those minor delays (< 30 minutes) too. Table VII outlines the major attributes of each delay case.

As mentioned before, EZ-Link data does not explicitly capture the commuters who are affected by a UTD, and we identify regular commuters and their regular travel patterns to answer this “who” question. We define a confidence parameter ρ to measure the overlap between a UTD U and a regular pattern P that is formed by a sequence of trips T_i (i.e., $P = \cup_{i=1}^j T_i$). Given a trip T , we extract a sub-trip M that involves MRT rides only. Notations $M.t_{in}$ and $M.t_{out}$ refer to the start and end time stamps of the MRT rides M respectively, and notations $M.l_{in}$ and $M.l_{out}$ indicate the location of the tap-in station and that of the tap-out station respectively. For a trip T containing MRT ride M , if $[M.t_{in}, M.t_{out}] \cap [U.t_{start}, U.t_{end}] \neq \emptyset$ and $(M.l_{in}, M.l_{out}) \cap (U.s_{start}, U.s_{end}) \neq \emptyset$, we consider that trip T might be affected by UTD U . The first condition guarantees the temporal overlap between the trip and the UTD, and the second condition guarantees the spatial overlap between the trip and the UTD, i.e., the route taken by the MRT ride M overlaps with the segment of MRT lines that is disrupted. Let function $AFFECT(U, T)$ which returns 1 or 0 indicate whether T and U are affected. Parameter ρ stands for the ratio of the number of trips in P that overlap with U and hence are affected by U to the length of our analysis period (i.e., 125

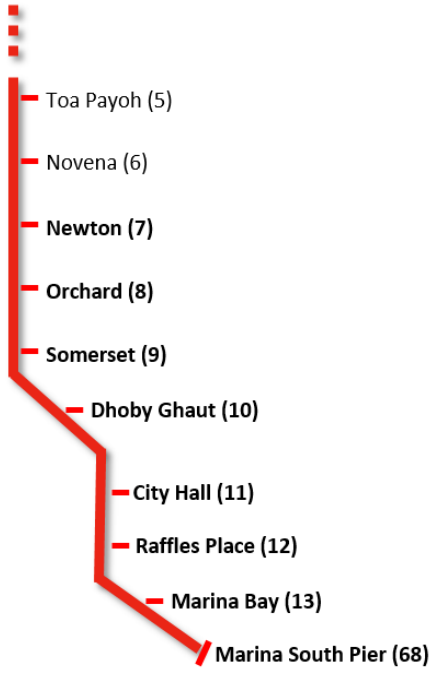


Fig. 4: A Segment of North South Line (number inside parentheses is MRT station id; affected MRT stations are highlighted in bold font)

weekdays), as defined in Equation (1).

$$\rho(P, U) = \frac{\sum_{\forall T_i \in P} \text{AFFECT}(U, T_i)}{125} \quad (1)$$

Apparently, ρ has its value ranging from 0 to 1, and it quantifies the likelihood that a commuter who owns this travel pattern P will be affected by the given UTD U .

In our study, we regard each regular commuter $user_i$ with at least one travel pattern P (denoted as $\exists P \in user_i$) such that $\rho(P, U) \geq 0.75$ as a commuter affected by UTD U . If a commuter has multiple patterns P with $\rho(P, U) \geq 0.75$, we only consider the pattern P with the largest $\rho(P, U)$ value. To facilitate following discussion, we introduce another function $\text{PATTERN}(U, user_i, v)$ that returns the pattern P among all the travel patterns P_j of a commuter $user_i$ with the largest $\rho(P, U)$ value and meanwhile $\rho(P, U)$ is no smaller than v , i.e., $\rho(P, U) \geq v$ and $\forall P_j \in user_i, \rho(P, U) \geq \rho(P_j, U)$. In other words, given a UTD U , a threshold value v and a commuter $user_i$, if $\text{PATTERN}(U, user_i, v)$ returns one pattern P , commuter $user_i$ is considered to be affected by U because of pattern P ; otherwise $\text{PATTERN}(U, user_i, v)$ returns NULL value and commuter $user_i$ is considered to be not affected by U .

Given 34 UTD and threshold value of 0.75 (i.e., v of $\text{PATTERN}(U, user_i, v)$ is set to 0.75), in total 21,832 regular commuters have been identified as affected. For each affected regular commuter $user_i$, we compare the trip she made on the day/time UTD U happened with her regular trips clustered into the regular pattern P (i.e., the output of $\text{PATTERN}(U,$

TABLE VIII: Real Examples of Commuters Behaviour (UTD: 6:09am to 6:57am, April 21st 2016, from Marina South Pier station to Newton Station on the NSL, as shown in Figure 4)

ID	card id	boarding time	origin id	alighting time	destination id
A.1	13***A3	2016-04-19 05:56:09	2918	2016-04-19 06:50:28	8
	13***A3	2016-04-20 05:52:14	2918	2016-04-20 06:47:04	8
	13***A3	2016-04-21 05:51:20	2918	2016-04-21 07:14:03	8
	13***A3	2016-04-22 05:53:02	2918	2016-04-22 06:43:18	8
A.2	2B***AA	2016-04-19 06:28:01	9	2016-04-19 07:26:16	5909
	2B***AA	2016-04-20 06:35:01	9	2016-04-20 07:31:24	5909
	2B***AA	2016-04-21 11:48:50	9	2016-04-21 12:47:26	5909
	2B***AA	2016-04-22 06:26:57	9	2016-04-22 07:28:05	5909
B.1	1F***A0	2016-04-19 06:25:08	109	2016-04-19 06:39:16	8
	1F***A0	2016-04-20 06:30:13	109	2016-04-20 06:47:42	8
	1F***A0	2016-04-21 06:26:43	109	2016-04-21 06:53:30	1214
	1F***A0	2016-04-22 06:33:37	109	2016-04-22 06:50:40	8
B.2	2C***2B	2016-04-18 06:37:43	7	2016-04-18 07:34:20	4649
	2C***2B	2016-04-19 06:49:03	7	2016-04-19 07:48:10	4649
	2C***2B	2016-04-20 06:36:33	7	2016-04-20 07:32:41	4649
	2C***2B	2016-04-22 06:43:57	7	2016-04-22 07:44:21	4649

$user_i, 0.75)$). In this study, we differentiate the commuters who switch to a different travel mode from those who still travel via MRT even during UTD, i.e., i) Type A: commuters still travel via MRT; and ii) Type B: commuters switch to other travel modes. Table IX reports the number of Type A commuters and that of Type B commuters.

For type A commuters, they may wait inside the MRT station for the MRT line to resume its service, especially common for UTD with short duration. Those commuters still take the regular route, exactly the same as all the trips they take in the normal days without UTD. However, the travel time has been extended. Commuter labeled as A.1 listed in Table VIII gives one example of such commuters. That commuter travels from bus stop 2918 to MRT station 8 almost every morning around 6:00 am, with the trip taking less than 55 minutes. On April 21st 2016 when the UTD happened, this commuter started her trip as usual, but the whole journey took her more than 80 minutes which is much longer than usual. Some commuters postpone their trips, especially those commuters who have not yet started their trips when the UTD happens. They travel via the regular route, but depart at a much later time point, usually after the resumption of train services, e.g., commuter A.2 shown in Table VIII. Her usually departure time is around 6:30 am but on the day UTD happened, she postponed her trip until 11:48 am.

TABLE IX: Commuters Response Distribution

Response Type	No. and % of commuters
Type A: commuters still travel via MRT	18,807 (86.1%)
Type B: commuters switch to other travel modes	3,025 (13.9%)

For type B commuters, some may change to buses with the alighting bus stops being close to the origin destination. An example of such commuters can be found in Table VIII, commuter B.1. She normally ends her trip at MRT Station whose ID is 8. On April 21st 2016 when the MRT service from Marina South Pier Station to Newton Station on the NSL was disrupted, she took the bus to complete her trip with the new alighting bus stop (ID 1214) being very close to the original alighting MRT station (ID 8). Some commuters may switch to taxis or bikes (or even walking) which are not captured by EZ-Link data. If those commuters switch to other transportation modes before they start the trip, they might not have any travel record on the day UTD happens, e.g., B.2 of Table VIII. If those commuters are in the middle of their journeys while the UTD happens, they may end their rides earlier at MRT stations/bus stops different from usual destinations and they switch to other transportation mode to complete their trips.

According to Table IX, most stranded commuters still use their original travel mode and route even during UTD, either by waiting or by delaying trip departure timestamps. Only 14 percent of affected commuters switch to other travel modes upon UTD.

V. COMMUTER RESPONSE PREDICTION MODEL

Based on the comparison study we perform in the previous step, commuters can be classified into two different types based on their reactions to a UTD. In this step, our target is to build a prediction model that is able to classify a commuter who will be affected by a given UTD to the right type. On the other hand, we all understand that the decision making process is extremely complex. The commuters' behavior during UTD could be affected by many factors and it is essential to have a good understanding of the impact of each factor on the way commuters eventually behave.

Based on literature, the following factors are believed to have a potential impact on the mode choice of travelers in response to UTD: cause of train service delays, stage of a trip, trip purpose, anticipated delay information, uncertainty of delay duration, and the weather. In addition to these factors, we study several other features as listed in Table X. Among those six features listed in Table X, the first two are related to UTD, which are shared by all the commuters; while the last four are related to each individual commuter. It is worth noting that the last three features listed in Table X are not directly available which require feature engineering.

To be more specific, given a commuter $user_i$ that is affected by a UTD U , $TripDuration$ captures the average duration of all the trips belonging to the regular travel pattern $PATTERN(U,$

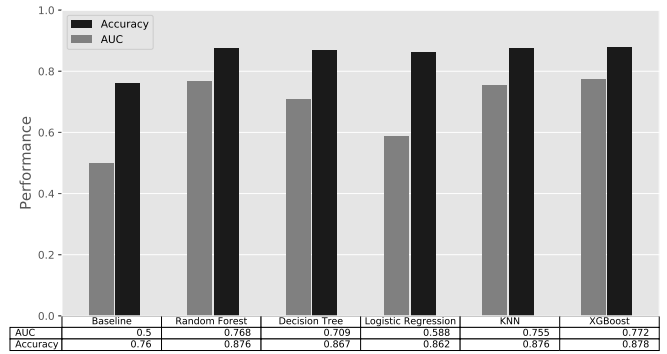


Fig. 5: Algorithm Performance Comparison

$user_i, v)$ with v set to 0.75 in our study. $Flag_{enroute}$ is inferred based on the starting time of UTD and the departure time of affected trips. It is set to one if UTD happens after the commuter has started the trip. $UrgencyLvl$ is defined as the reverse of the standard deviation σ of alighting times of all the trips belonging to $PATTERN(U, user_i, v)$. A large $UrgencyLvl$ is from a small σ value which corresponds to a regular commuter who reaches her destination via trips in $PATTERN(U, user_i, v)$ at roughly the same time stamp. In other words, commuters with large $UrgencyLvl$ values need to end their trips within a very small time window.

According to literature, there are other commuter-oriented factors that could affect commuters' decision, such as commuters' income, residential area, and their desire to experiment and habit [17]. However, such information is not available, which imposes a limitation on the accuracy of the prediction model built on top of the features that are captured by EZ-Link data. In the following, we first present the models we use, and then report the insights we learn.

A. Building Predictive Model

When doing supervised learning, a simple sanity check consists of comparing one's estimator against simple rules of thumb. *sklearn's DummyClassifier* module [18] implements several such strategies for classification. In this study, we use the stratified strategy as the baseline to generate random predictions by following the training set class distribution and it can achieve an accuracy score of 76.0% with AUC (area under the ROC curve) of 0.5.

Using a 5-fold nested cross validation, we implement a few state-of-the-art machine learning models, namely *decision trees* [19], *logistic regression*, *k nearest neighbors*, *random forest* [20], and *XGBoost* [21] and report their accuracy and AUC in Figure 5. It is noted that XGBoost outperforms other models and achieves the highest accuracy of 87.8% and the highest AUC of 0.772.

B. Feature Insights

In the following, we will present the feature insights we gain while building the predictive model and highlight the most important features that dominate commuters' decision making

TABLE X: Important Features

ID	Feature	Type	Description
1	UTDCategory	Categorical	cause of a UTD, e.g., signaling fault, track fault, maintenance, etc.
2	UTDPeriod	Categorical	depending on the time when a UTD happens, it is labelled as morning peak, evening peak or non-peak
3	CommuterCategory	Categorical	four categories of commuters: adult, student, kid, and senior citizen
4	TripDuration	Numeric	the duration of a trip based on historical records (unit: minute)
5	Flag _{enroute}	Boolean	value 1 indicates a UTD happens during a commuter’s trip and value 0 indicates it happens before the trip
6	UrgencyLvl	Numeric	a number measuring how urgent a trip is

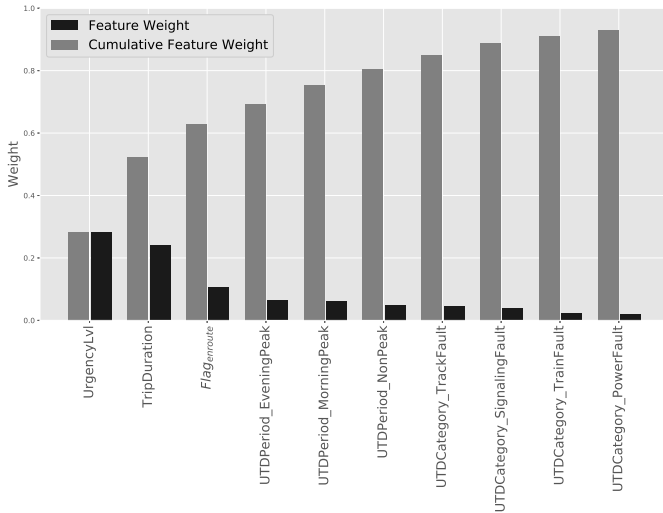


Fig. 6: Top 10 Predictors and Their Weights

process. We acknowledge that insights presented in the following are inferred from the association between the response and each individual feature variable, without accounting for other variables. Causal relationships between feature variables and responses are beyond the scope of this paper.

The top predictors ranked by feature importances from XG-Boost are: UrgencyLvl, TripDuration, Flag_{enroute}, UTDPeriod and UTDCategory, as depicted in Figure 6. These five features together contribute to more than 90% of the predictive power. The correlations between the most predictive features and commuters’ responses are depicted in Figure 7. Note that the bars correspond to the value distributions of each feature, and the red dots report the ratio of Type B commuters (i.e., switch to other travel modes) to the total number of affected commuters with feature value fallen within the corresponding range.

Consistent with our expectation, as UrgencyLvl becomes larger and larger, the standard deviation of the alighting time stamps is smaller and smaller. In other words, the commuters need to complete the trips within a smaller and smaller time window, and waiting inside the MRT station is no longer a practical option. Commuters who do not have much flexibility in changing the time when they need to complete the trips

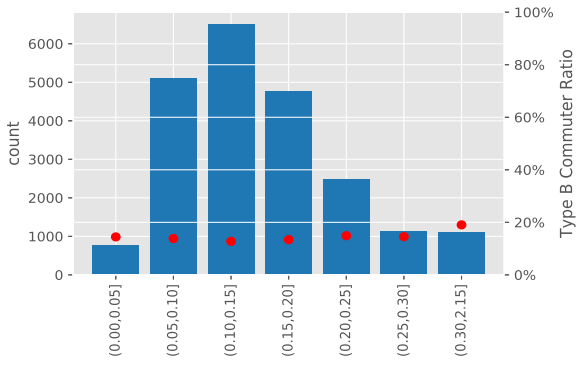
would just change travel modes to continue their journeys.

We notice that duration of trip has an impact on commuters’ responses too. The likelihood of changing travel mode declines with increasing trip duration when the trip is shorter than 30 minutes. This can be explained that when the trip is short, commuters usually have more options such as walking or taking buses. However, if the trip is 30 minutes or longer, the likelihood of changing travel mode increases. One possible reason is that the longer the trip, the greater the uncertainty about the time when the destination is reached. Consequently, commuters taking long trips may want to avoid extra uncertainty that could be caused by waiting for MRT to resume its services. This might be able to explain the relatively high rate of type B commuters as TripDuration becomes longer.

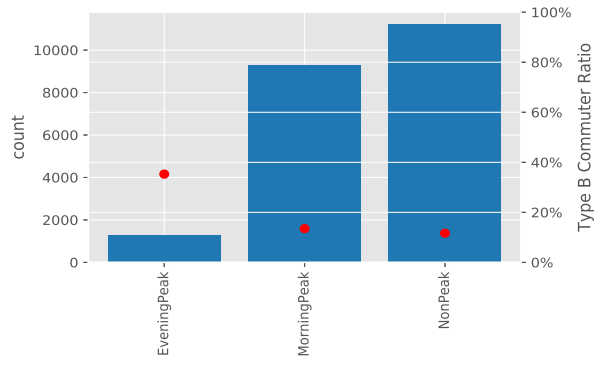
We also observe that the time a UTD happened (i.e., UTDPeriod) and whether commuters encounter UTD en-route or pre-trip (i.e., Flag_{enroute}) have a significant impact on commuters’ responses.

Recall that UTDPeriod is a feature generated by us via feature engineering. According to the time when a UTD starts, we set the feature UTDPeriod of each UTD to *morning-peak* (7am to 10am), *evening-peak* (5pm to 8pm), or *non-peak* (the rest). When a UTD happens during peak hours (including both morning-peak and afternoon-peak), commuters are more likely to change their travel modes. In particular, when a UTD happens during evening-peak, 35% of affected commuters would change their travel mode, while only 11% of stranded commuters would do so when the UTD happens in non-peak hours. One possible reason is that people need to reach their offices or schools at required time in the morning peak and they want to reach their destinations earlier (e.g., home, restaurants for appointments) in the afternoon after one busy day.

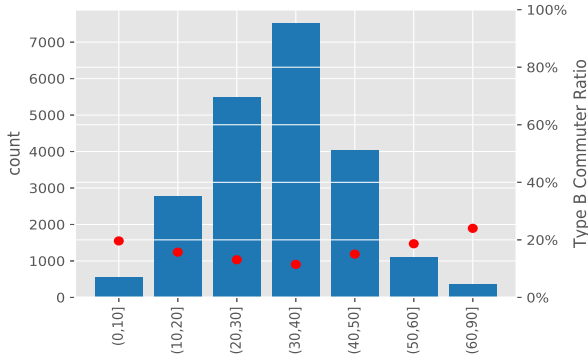
Commuters with en-route UTD have a higher chance (18%) of changing travel mode, while only 13% of commuters encountering UTD pre-trip change their travel mode. One possible reason is that being aware of the UTD before starting the trips opens up other options such as postponing the departure time without route change. On the other hand, commuters who encounter UTD in the middle of their journey may regard waiting as the easiest option because of multiple reasons such as unfamiliarity of the current MRT station or the unawareness of alternate routes.



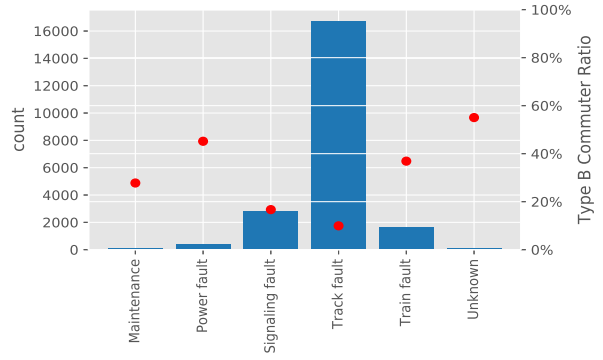
(a) UrgencyLvl



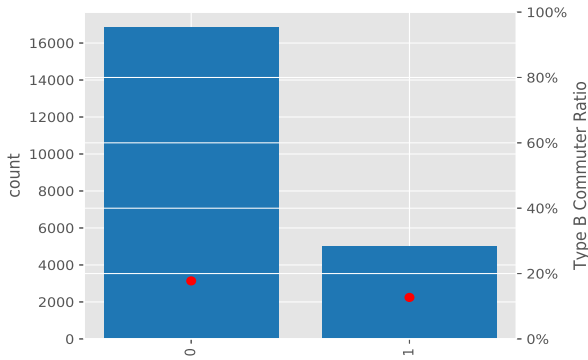
(d) UTDPeak



(b) TripDuration (unit: minutes)



(e) UTDCategory



(c) Flag_{enroute}

Fig. 7: Correlation Between Top Predictors and Commuters' Response

especially those making short trips still can bear with that.

VI. CONCLUSION

In this article, we conduct data-driven analysis on EZ-Link data to understand commuters' behavior upon UTDs and propose an effective predictive model to model commuters' responses upon UTD using massive smart card data.

We have summarized the main challenges of the predictive model to three research questions, mainly *who*, *how*, and *why* questions. The *who* question looks for commuters who are or will be affected by a UTD. As a solution, we introduce the notion of regular commuter and adopt a DBSCAN alike data mining approach to mine and summarize regular travel patterns for each individual commuter from EZ-Link data. The *how* question investigates how affected commuters behave when UTD happens. To be more specific, this paper studies whether affected commuters switch to other transportation modes because of UTDs. We introduce a quantitative measurement to identify those commuters who are affected by a given UTD, and perform a comparison study to examine their behavior with and without UTDs. The *why* question tries to find out the key factors that affect the commuters' decision on whether to change the transportation modes or not during UTDs. We perform feature engineering to identify a list of important features that have impacts on the decision-making process of

The cause of UTD (i.e., feature UTDCategory) has an impact on commuters' responses as well. UTD caused by power fault or unknown reasons incurs the highest Type B commuter rate, followed by UTD caused by maintenance and train fault reason. On the other hand, most of the commuters still use MRT as the travel mode when UTDs are caused by signaling fault or track fault. One reasonable explanation is that power fault and unknown reasons cause very serious faulty scenarios, under which trains may stop working persistently or intermittently. The severe faulty incidents exceed commuters' tolerance threshold and many commuters decide to look for alternative options. On the other hand, UTDs caused by signaling fault or track fault are minor, and trains are still able to move, but at a lower speed. Accordingly, commuters

a commuter in terms of whether to change the transportation mode or not during UTDs and build a classification model to differentiate commuters who still use MRT from those who switch to other transportation modes.

In the near future, we plan to extend our study to further investigate the behavior of type B commuters, those who switch to other transportation modes during UTDs, and understand their exact choices.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative. The authors would like to thank LTA for providing the data. However, we declare that all the findings shared in this paper represent the opinions of the authors but not LTA.

REFERENCES

- [1] Christopher Tan. *Bus, rail ridership soars to new high*. URL: <https://www.straitstimes.com/singapore/transport/bus-rail-ridership-soars-to-new-high>.
- [2] Brendan Pender et al. "Disruption recovery in passenger railways: International survey". In: *Transportation Research Record: Journal of the Transportation Research Board* 2353 (2013), pp. 22–32.
- [3] Jian Gang Jin, Kwong Meng Teo, and Lijun Sun. "Disruption response planning for an urban mass rapid transit network". In: *Transportation Research Board 92nd Annual Meeting*. 2013, pp. 13–17.
- [4] Jan-Dirk Schmöcker, Shoshana Cooper, and William Adeney. "Metro service delay recovery: comparison of strategies and constraints across systems". In: *Transportation Research Record: Journal of the Transportation Research Board* 1930 (2005), pp. 30–37.
- [5] Tim Darmanin, Calvin Lim, and H Gan. "Public railway disruption recovery planning: a new recovery strategy for metro train Melbourne". In: *Proceedings of the 11th Asia Pacific Industrial Engineering and Management Systems Conference*. 2010.
- [6] Julie Jespersen-Groth et al. "Disruption management in passenger railway transportation". In: *Robust and Online Large-scale Optimization*. Springer, 2009, pp. 399–421.
- [7] Genevieve Giuliano and Jacqueline Golob. "Impacts of the Northridge earthquake on transit and highway use". In: *Journal of Transportation and Statistics* 1.2 (1998), pp. 1–20.
- [8] Shanjiang Zhu et al. "The traffic and behavioral effects of the I-35W Mississippi River bridge collapse". In: *Transportation Research Part A: Policy and Practice* 44.10 (2010), pp. 771–784.
- [9] Sally Cairns, Stephen Atkins, and Phil Goodwin. "Disappearing traffic? The story so far". In: *Proceedings of the Institution of Civil Engineers-Municipal Engineer*. Vol. 151. 1. Thomas Telford Ltd. 2002, pp. 13–22.
- [10] Matthew G Karlaftis et al. "Public Transportation during the Athens 2004 Olympics: From planning to performance evaluation". In: *Transportation Research Board 85th Annual Meeting*. 2006.
- [11] Peter Rickwood and Garry Glazebrook. "Urban structure and commuting in Australian cities". In: *Urban Policy and Research* 27.2 (2009), pp. 171–188.
- [12] Ka Chu and Robert Chapleau. "Augmenting transit trip characterization and travel behavior comprehension: Multiday location-stamped smart card transactions". In: *Transportation Research Record: Journal of the Transportation Research Board* 2183 (2010), pp. 29–40.
- [13] Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. "Smart card data use in public transit: A literature review". In: *Transportation Research Part C: Emerging Technologies* 19.4 (2011), pp. 557–568.
- [14] Nancy McGuckin and Yukiko Nakamoto. "Trips, Chains and Tours-Using an Operational Definition". In: *National Household Travel Survey Conference*. 2004.
- [15] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, pp. 226–231.
- [16] Jörg Sander et al. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications". In: *Data Mining and Knowledge Discovery* 2.2 (1998), pp. 169–194.
- [17] Phil Goodwin. "Policy incentives to change behaviour in passenger transport". In: *May-2008.[Online]*. Available: <http://www.internationaltransportforum.org>. [Accessed: 14-Dec-2012] (2008).
- [18] sklearn. *sklearn.dummy.DummyClassifier*. URL: <http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html>.
- [19] J. Ross Quinlan. "Induction of decision trees". In: *Machine learning* 1.1 (1986), pp. 81–106.
- [20] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.
- [21] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 785–794.