Planning and Transport Research Centre (PATREC)

TECHNICAL REPORT

Travel Behaviour Patterns – Micro Analysis

<table>
<thead>
<tr>
<th>Project No</th>
<th>Project 4.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Rachel Cardell-Oliver, Travis Povey, Lidia Dokoucheva, Jianxin Li, Jordan Lilburne, Sharon Biermann</td>
</tr>
<tr>
<td>Steering Committee</td>
<td>Damien Martin (DOP), Wes Soet (MRWA), Graham Jacoby (MRWA), Charlotte Hayes (PTA), Sue Hellyer (DOT), Rosie Selby (DOT), Renlong Han (DOT), Laura Cook (Treasury)</td>
</tr>
<tr>
<td>Date</td>
<td>November 2017</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

List of Tables ........................................................................................................................................ 3

Project Overview ..................................................................................................................................... 4

2 Introduction ......................................................................................................................................... 5
   2.1 Project Initiation .......................................................................................................................... 5
   2.2 Project Components ..................................................................................................................... 5
   2.3 Aims and Objectives ..................................................................................................................... 5
   2.4 Project Scope ............................................................................................................................... 6
   2.5 Methodology ............................................................................................................................... 6
   2.6 Outputs ......................................................................................................................................... 7

3 Identifying Activity Hubs from Ticketing Logs ............................................................................... 7
   3.1 Methodology ............................................................................................................................... 8
   3.2 Results and Discussion ................................................................................................................ 10
   3.3 Where are the hubs? .................................................................................................................... 2
   3.4 Conclusions and Future Work ..................................................................................................... 2

4 Learning Passenger Usage Patterns From SmartRider Ticketing Logs ......................................... 4
   4.1 Methodology ............................................................................................................................... 4
   4.2 Results and Discussion ................................................................................................................ 6
   4.3 Conclusion and Future Work ....................................................................................................... 6

5 Passenger Analytics - Continuing Project ....................................................................................... 7
   5.1 Passenger Travel Profiles .......................................................................................................... 7
   5.2 Passenger Types .......................................................................................................................... 8
   5.3 Passenger Narratives .................................................................................................................. 10

6 Journey-centered Analytics - Continuing Project ........................................................................... 1

7 Spatial Hub analytics - Continuing project ...................................................................................... 2

8 Results Highlights ............................................................................................................................... 2

9 Conclusions and Future Work - What next? .................................................................................... 2

10 References ........................................................................................................................................ 3

11 List of Project Outputs ........................................................................................................................ 4
11.1 Decision-Support Software, Documentation and Visualisations .................. 4
11.2 Data Sets of Results ................................................................................... 4
11.3 Academic Reports ...................................................................................... 4
11.4 Workshops with Policy makers and Stakeholders ....................................... 5
11.5 Refereed Research Publications .................................................................. 5

12 Project Management ....................................................................................... 5
12.1 Budget ......................................................................................................... 5
12.4 Project 4 Steering Committee ..................................................................... 6

13 Acknowledgements ......................................................................................... 7

List of Figures

1. In-flow (red) and out-flow (blue) of four Perth stops by time of day ............... 9
2. Activities identified from NMF on the matrix of stops. Arrival Hour is separated into 24 one hour bins, and Duration into 30 half-hour bins (1 hour to 16 hours) . . 11
3. Location and intensity of different hub activities in the Perth public transport network. Clockwise from top left: residential, shopping (ad hoc short stay), school, work, university (ad hoc longer stay) ........................................................................ 13
4. Passenger types discovered by clustering passenger activity profiles .......... 16
5. Examples of passenger travel profiles ............................................................ 18
6. Passenger types discovered by partitioning (top) and their significance as percentage of passengers, journeys and fares (bottom) ........................................... 19
7. Some characteristics of one-way-only regular passengers (group R1A1) ........ 21

List of Tables

1. Activity composition of notable hubs.............................................................. 12
2. Frequency Band Statistics ............................................................................. 15
Project Overview

1 Executive Summary

Public transport is a critical aspect of any modern city. Many cities are moving to using paperless smart-card ticketing systems which provide a wealth of data about how the system is being used. This project aims to utilize SmartRider data from Perth’s ticketing system to develop an understanding of how passengers are utilizing the system.

The aim of project 4.2 is to develop a system for querying, analysis and data mining, to support a knowledge discovery process centred on passengers, hubs, and journeys. The objective is to be able to generate evidence-based answers from SmartRider ticketing logs to queries such as:

- What are the different types of passengers using the TransPerth network?
- How frequently and at what times do they travel?
- Where are the activity hubs in the Perth network?
- How intensively and with what stay patterns are the Perth hubs used?
- What are spatial catchments for different hubs?
- Which journey segments are heavily utilised during particular time periods?

Using TransPerth SmartRider ticketing logs and stop information, we have developed new data mining techniques that reveal latent information about passengers, activity hubs and heavy utilization of certain journey segments. Highlights of the findings of this project include:

- A total of 130 hubs were identified, 120 of which are located in the Perth metropolitan area, from Rockingham in the south, to Butler in the north. There were also hubs in Mandurah (4), Busselton (3), Albany (1), Geraldton (1) and Port Hedland (1). It was found that approximately half of all stays were at a single hub in Perth city, which covers Perth and Elizabeth Quay train stations.

- Identification of 5 distinctive ways in which hubs are used. Each of these uses is characterized by passengers’ arrival time and length of stay. The five usage patterns suggest: work day, school day, overnight stays, and variable arrival times followed by either a long or short stay.

- Description of each hub region by its unique mixture of the five activities. Dominant hub activities correlated well with points of interest such as schools, universities, business and shopping centres.

- Discovery of a new and flexible typology for passengers. We found several significant passenger types that are not considered in traditional transport models viz. ad hoc travellers, and one-way-only commuters.

- Creation of databases and visualisation software for automatically generating reports on hubs, passengers and journeys, as well as textual narratives for the discovered patterns.

- For PATREC participants a full table of all the discovered Perth hubs and their activity mixes is available. Section 11 details all software and data outputs for this project. These outputs are available in electronic form to PATREC participants by request.

Project 4.2 has met its agreed milestones and completed the project deliverables. Some research streams from the project will be continuing into 2018.
2 Introduction

Understanding urban mobility patterns is important for transport authorities, city councils, and businesses. Currently, TransPerth uses surveys conducted in person at high volume areas of the network, to assess performance and passenger satisfaction (Painted Dog, 2016). Perth’s SmartRider system was launched in 2007, and within a year, over 70% of public transport transactions were conducted through a SmartRider (PRIA, 2012) increasing to approximately 77% in 2014-2015 (PTA, 2015). As Perth plans growth towards a population of 3.5 million, the importance of reducing car trips and increasing active and public transport trips is critical. An aspirational target for Perth is 11% public transport patronage, well above the current 7% level.

Perth’s SmartRider logs offer a valuable, but largely untapped, source of knowledge on the patterns of use of Perth’s public transport network. Previous studies using passenger smart card data from cities around the world (Hasan 2013, Kieu 2014) have focussed on extracting explanatory statistics of global behaviours such as frequency and scale of travel. Another line or research uses feature extraction algorithms to discover hidden features in data (Poussevin 2016, Yuan 2014). Project 4.2 has utilised the latest computing techniques in automated knowledge discovery to reveal latent information about passengers, activity hubs and utilization of certain journey segments using the SmartRider logs.

2.1 Project Initiation

Project 4.2 commenced in July 2016 with the appointment of one UWA CEED Scholar. Two further scholars were appointed in February 2017 and a fourth in July 2017. To date, five undergraduate research students have worked on the project during 2016 and 2017 and one Engineering Masters project student from August 2017. Three UWA Faculty are supervising these projects: Rachel Cardell-Oliver, Jianxin Li and Wei Liu.

2.2 Project Components

Project 4 is an evidence based, data-driven project centred around the analysis of SmartRider data. The project is divided into two complementary projects, one focusing on macro analysis (4.1), investigating large scale patterns in the data, and the other on micro analysis (4.2), studying the travel patterns of individuals. Project 4.2 is organised around sub-projects centred on three key entities in the Perth public transport network, namely, passengers, hubs and journeys.

2.3 Aims and Objectives

The aim of project 4.2 was to develop a system for querying, analysis and data mining, to support a knowledge discovery process centred on passengers, hubs, and journeys. Our objective is to be able to generate evidence-based answers automatically from SmartRider ticketing logs to queries such as:

- What are the different types of passengers using the TransPerth network?
- How frequently and at what times do they travel?
- Where are the activity hubs in the Perth network?
- What are the activities that take place at those hubs, and for how many people?
- What are spatial catchments for different hubs?
Which journey segments are heavily utilised during particular time periods?

2.4 Project Scope

The scope of this project is defined by three constraints.

Project 4.2 focussed on certain types of knowledge that can be learned from the ticket logs: knowledge about passengers, places (hubs) and journeys.

The research is data-driven with results derived only from the SmartRider ticketing logs. The idea is to discover latent information from the ticket logs, without the requirement for external data such as passenger demographics from surveys or suburb profiles. In this way, project 4.2 complements the research in project 4.1 which links the SmartRider data with external data sources.

The third constraint is a focus on consistent behaviours. For this reason, we focus our analysis on short time intervals such as 4 weeks or one month. We chose short periods in order to capture consistent behaviours because over longer periods passenger behaviours change (e.g. term time vs holidays), and so averages can be misleading. Our methods can, however, be run automatically for any chosen month or selection of days (e.g. weekdays vs weekends, term time vs holidays).

2.5 Methodology

The following methodology was used to deliver the scope of work.

Data Discovery and Extract-Transform-Load Specify preliminary data requirements for each of the three sub-projects, and import data for phase 1 from NetBI in collaboration with PTA. The raw data is extracted monthly, transformed and loaded into a private and secure UWA Institutional Research Data Store (IRDS) for project participants.

Literature Reviews Review of existing representations and data mining algorithms for analysis of passengers and journeys. Assessment of the applicability of these methods for this project. Selection of the most promising approaches for identified query tasks.

Data Exploration Develop data mining algorithms to identify patterns in the micro-data. This task focused on micro-patterns, such as per-passenger, per-hub and per-journey features, complementing the macro-level analysis of Project 4.1. The first step is to develop algorithms for identifying per passenger, per hub, and per-journey features. The second step is clustering the data to identify common patterns, leading to a models for passenger types, hub types and journey segments.

Presentation of Preliminary Results This component investigated ways of packaging the results for policy makers. We used graphics, association rules and taxonomies to present the results of the data-mining supported queries.

Dissemination of Results In collaboration with PATREC domain experts, the results were packaged in usable forms as data sets, decision support software and written reports.
2.6 Outputs

Project 4.2 has delivered analysis methods for discovering patterns and software tools to facilitate that analysis and for visualisation of the results for policy makers. The project outputs are of these types:

1. Data sets (the results generated by data mining algorithms: for example passengers or hubs, their mix of activities and additional features that help to understand the activities),

2. Decision Support Software (implementation of our new data mining algorithms and prototypes that allow users to explore the data), and

3. Reports (literature surveys and research theses)


Project 4.2 has met its agreed milestones and completed the project deliverables. Some research streams from the project will be continuing into 2018.

3 Identifying Activity Hubs from Ticketing Logs

This section is based on 2017 CEED Seminar paper by Travis Povey

Smart-card ticketing systems are becoming ubiquitous in many cities, and they offer a wealth of data that can be analysed to better understand and improve a public transport network. Perth’s SmartRider system was launched in 2007, and within a year, over 70% of public transport transactions were conducted through a SmartRider (PRIA, 2012) increasing to approximately 77% by 2014-2015 (PTA, 2015). The popularity of smart-card systems continue to grow, among both users, and governments looking to replace paper ticketing systems.

Data collected from the SmartRider system provides accurate information about how the network is used, and can replace traditional survey methods. Currently, TransPerth uses surveys conducted in person at high volume areas of the network to assess performance and passenger satisfaction (Painted Dog, 2016). SmartRider data is a largely untapped resource that could go a long way towards improving understanding of Perth’s transport network.

This project focuses on understanding Perth at a regional level, by understanding travel patterns associated with activities, and how that drives traffic between hubs. An activity may be something such as working, going to school, or shopping, and part of these activities involves travelling. A hub is an area that facilitates one or more activities. In the infrastructure planning field, it is often referred to an activity centre.

Previous research has analysed smart-card data to understand travel patterns, and how this can be used to improve transport infrastructure. Yuan et al. (2014) identified functional zones (or hubs) by analysing smart-card ticket logs, and taxi trips in Beijing, supplemented with points of interest data, which describes the number of services (such as shops, cafes, theatres etc.) in a zone. The points of interest data is used to determine the type of zone, and the transport data is used to determine which zones should be aggregated, and the relationship between zones. Our project aims to identify activities without using external points of interest data, and to infer the activity type from its characteristics.

Poussevin et al. (2016) focused on individual patrons of the Paris Metro network, and identified common temporal travel patterns. The goal was to characterize each trip by an activity, which
defines the reason the trip occurs. One important aspect highlighted in this paper is the separation of the patrons into frequency bands, based on how commonly they use public transport. Frequent users are likely to travel to and from work every day, whereas someone who uses public transport only once a month is much less likely to be going to work on that trip.

3.1 Methodology

3.1.1 Stop Profiles

Considering an individual bus, train or ferry stop, we can map the in-flow (tag-offs) and out-flow (tag-ons) from that stop by time of day. Figure 1 shows the profiles of four highly used stops in the Perth network. For this example the y-axis shows the average number of tags per 15 minutes on weekdays. It can be seen that each of the stops have characteristic profiles suggesting activities such as work (in AM, out PM), bus stops (eg all arrivals), shopping and entertainment (in and out all day) and residential (out AM, in PM).

A dictionary of profiles for the busiest 110 stops in the network is available as a project output. However, considering the pattern of a single stop can be misleading. For example, bus stops to and from a given location are located on opposite sides of the road, and so we need to consider at least pairs of bus stops. On the other hand, grouping many stops using external entities such as STEM or ROM zones, SA census zones, or suburbs, does not adequately capture the idea of an activity centre. Either these models place too many stops into one group. Or, in the case of STEM and ROM which are based on the road network, many activity centres such as railway stations are split between two or more zones. The challenge is, how do we discover groups of stops that characterise an activity centre? Recall that our goal is to use only the SmartRider ticket logs for this discovery.

3.1.2 Defining Features

The first step is to identify features extracted from the data that are representative of an activity. We propose using a stay as the main feature, which is the time a patron arrives, and how long they stay within a local area. This is determined by examining trip pairs from the data the first of which describes the arrival time and location, and the second being the departure. If the time difference between arrival and departure is 16 hours or less, and the distance between the two stops is less than 500m, it is assumed that the person undertook some activity in that area. These values were determined by examining the distribution of stays and stop pairs in the dataset.

Stays are characterized by two values: arrival to nearest hour, and duration to nearest half-hour. Each stop in the network is then described by a 56-component vector, the first 24 values representing the number of arrival times of stays originating at that stop, and the following 32 values representing the number of each duration of the stays (0 to 16, in half-hour bins). This vector represents the probability that a person arriving at that stop arrives at a particular hour, and stays for a particular duration.

3.1.3 Identifying Hubs

A goal of this project is to identify hubs where certain activities are undertaken. This involves grouping together stops of close proximity and similar type into hubs. In order to determine which stops should be grouped together, we first consider stop pairs. These are pairs of stops used by at
Figure 1: In-flow (red) and out-flow (blue) of four Perth stops by time of day
least one person for a particular stay. We assume that if a person arrives at one stop, and departs from another, satisfying the conditions for a stay, then both those stops should be associated to the same hub. In order to identify highly utilized hubs strongly associated with an activity, we also require that the stop pair is utilized by a large number of patrons.

Clustering stop pairs into hubs is done spatially, grouping nearby stop pairs based on distance and density, forming coherent groups into hubs. Not every stop is sorted into a hub, only highly utilized stops. Each hub can be associated with one or more activities that describe the reason that a person would visit it.

### 3.1.4 Determining Activities

We define the feature-vector of a hub, as the sum of all feature-vectors of all its stop pairs. Activities will be represented as common patterns exhibited among the set of all hub’s feature-vectors. In order to identify these patterns, non-negative matrix factorization (NMF) is used. This method breaks down a matrix V into a set of weights, W, and basis vectors, H, such that V=WH. The input to this process is the number of basis vectors to identify, and V, the matrix of all hub feature-vectors. Each row in the matrix can be represented as a sum of the basis vectors, which are common patterns that appear in the matrix. This process isn’t exact; the resulting matrix WH will be an approximation of the original.

The result of this process is a set of basis vectors, H, which represent the activities, and a set of weights, W, which represent the degree to which each hub is comprised of each activity. By normalizing the weights, we can determine the percentage mix, and dominant activity. There is no way to determine the exact number of “correct” activities to find, instead a heuristic approach is used, where the number of activities should be high enough so that the approximation is sufficiently accurate, but there should be few, if any, duplicate activities.

### 3.2 Results and Discussion

#### 3.2.1 Extracting Stays

The results in this section are derived from analysing one month of travel data from October, 2016, which contains approximately 8 million transactions. Each transaction provides the following information:

- Card ID Number
- Location of Tag-On (Stop ID Number)
- Location of Tag-Off (Stop ID Number)
- Time of Tag-On
- Time of Tag-Off
- Type of transaction (standard, concession, senior)

The threshold used to determine if a pair of transactions should be used to derive a stay is based on the Tag-Off data, and the Tag-On data of the subsequent trip. A time threshold of between 1 to 16 hours was chosen because this is the typical length of an overnight residential stay. Longer durations might mean that the person travelled without using public transport, meaning we cannot know what activities are associated with that stay. 1 hour was chosen as the lower bound to eliminate transfers, as we are primarily interested in activities undertaken at a patron’s final destination. The distance threshold of 500m was chosen by calculating the 95th percentile distance between
Arrival Hour Duration

Figure 2: Activities identified from NMF on the matrix of stops. Arrival Hour is separated into 24 one hour bins, and Duration into 30 half-hour bins (1 hour to 16 hours).

stops for all valid transaction pairs (below the time threshold, and same Card ID). This is relatively close to a commonly cited value of 400m as the maximum distance a person will walk for a bus (Daniels 2006). From the 8 million records, approximately 2 million stays are derived.

3.2.2 Identifying Activities

The arrival times of each stay is rounded to nearest hour, and duration to nearest half hour. In this analysis, there are only 30 half hour duration bins because 0 – 0.5, and 0.5 – 1 were excluded. Each stop is represented by two vectors: a 24 component vector for arrival hour bins, and a 30-component vector for duration.

Each vector representing a stop is a row in the matrix used as input to the NMF process. Using the heuristic outlined in Section 3.1.4 for determining number of activities, it was found that there are 5 distinct activities. Figure 2 shows the identified activities, separated into Arrival Hour (column 1), and Duration (column 2). There are some recognizable patterns in these activities. Activity 1 is arrival at 7 to 8am, for a duration of 8 – 10 hours, which is a typical pattern for a work day. Activity 2 is arrival strongly at 8am, and a duration of 6.5 – 7.5 hours. This likely corresponds to a school day activity. Activity 3 includes mid to late morning arrivals, for a spread of durations.
Table 1: Selected hubs and their percentage composition of the identified activities

<table>
<thead>
<tr>
<th>Region</th>
<th>Description</th>
<th>Uni (%</th>
<th>School (%)</th>
<th>Resi. (%)</th>
<th>Shop. (%)</th>
<th>Work (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mt Claremont</td>
<td>Residential Zone</td>
<td>18</td>
<td>20</td>
<td>41</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Osborne Park</td>
<td>Industrial Zone</td>
<td>10</td>
<td>4</td>
<td>31</td>
<td>2</td>
<td>53</td>
</tr>
<tr>
<td>Karrinyup Shops</td>
<td>Shopping Mall</td>
<td>18</td>
<td>7</td>
<td>7</td>
<td>55</td>
<td>13</td>
</tr>
<tr>
<td>Perth CBD</td>
<td>Business District</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>84</td>
</tr>
<tr>
<td>UWA</td>
<td>University</td>
<td>69</td>
<td>8</td>
<td>3</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

This activity is commonly associated with stops around universities. Activity 4 represents daytime, short duration outings. This could be activities such as shopping, lunch, and general errands. We can see residential stays (overnight) in activity 5: people arriving at 4-6pm, and staying for 13-15 hours.

3.2.3 Activities Associated with Hubs

Since we are not using points of interest directly, inferring the real activity associated with each derived basis vector is mainly done using knowledge about the area around stops associated with each particular activity. In order to gain some insight into these activities, and validate our understanding, we examine the weights associated with some identified hubs. Table 1 outlines a selection of hubs with predictable results, and their corresponding composition of discovered activities. The results for Karrinyup Shops, Perth CBD, and UWA strongly match what would be expected of these regions.

3.3 Where are the hubs?

Figure 3 show the location of hubs associated with the five activity types. The size of each circle demonstrates the number of passengers involved in that activity. It can be seen that residential hubs are spread along the main train lines. University and school activities align with known locations of universities and schools. Short trip (shopping) hubs also alight with known locations. The work activity is very strongly focussed on the Perth CBD.

3.4 Conclusions and Future Work

We have identified hubs, which are areas of that attract a large number of visitors via public transport, by analysing the SmartRider ticketing logs. These hubs have a socio-economic function, which determines the demand for transport to and from these locations.

Activities associated with hubs have also been derived from the SmartRider data. These are identified as common arrival time and stay lengths for a large number of hubs. Five distinct patterns were discovered, which we reason are associated with the following activities: work, school, shopping, university, and residential. Each hub is comprised of a mix of activities, which is representative of its function. This mix was examined for several hubs, and compares well with expected results.

A significant amount of traffic in the public transport network in Perth is purely for travel to and from Perth city. It was found that Perth accounted for approximately 80% of all work activities, and about half of all stays. One difficulty when analysing Perth as a hub, is that we cannot know
Figure 3: Location and intensity of different hub activities in the Perth public transport network. Clockwise from top left: residential, shopping (ad hoc short stay), school, work, university (ad hoc longer stay).
if passengers utilize the free transit area, as there is no tag-on event, and thus no transaction in the
data set. Because of this, the true catchment area of the Perth hub is significantly larger than the
distance threshold around the centroid.

In future work we plan to further refine the process of identifying hubs. Data from different months
will be analysed to test the robustness of the algorithm, and determine to what extent hub locations
and activities are consistent or change over time.

4 Learning Passenger Usage Patterns From SmartRider Ticketing Logs

This section is based on the 2017 CEED Seminar paper by Lidia Dokuchaeva

Understanding urban mobility patterns is important for making modern cities liveable and pro-
ductive. Studying public transport in particular is critical for understanding urban mobility as
a whole. This project aims to reveal latent information from smart card data related to passenger
segmentation for public transport users. In this project, the latest data mining and machine learning
techniques were used and evaluated in order to develop and test a system that exploits SmartRider
logs to characterise the spatial and temporal habits of individual passengers. The paper presents
the results obtained for passengers with high frequency travel.

Public transport is a vital part of urban infrastructure, and studying the way that people use it
is critical for understanding urban mobility as a whole. The paper describes a project focusing
on evaluation based on data mining methods in order to develop and test a system that exploits
SmartRider logs to characterise the spatial and temporal habits of individual passengers and to
address the issue of passenger segmentation for public transport users.

Clustering is a machine learning technique for identifying similar groups of individuals in a pop-
ulation. For travel logs, it can be used to identify types of passengers (as in this project), types of
stops or types of journeys. By clustering the ticketing logs, it is possible to find the common ways
that public transport is used, and by finding how public transport is being used, it is possible to
evaluate the effectiveness of the public transport system overall - whether by performing further
analysis on the data set or by combining this knowledge with other knowledge such as household
tavel surveys. As such, clustering is an important way of understanding public transport, and thus
urban mobility as a whole.

Non-negative matrix factorization (NMF) was used for clustering transport passengers. NMF gen-
erates a multi-scale representation of the user activities. NMF has been used to analyse traffic
flow (Peng et al. 2012) and profiles of Paris metro passengers (Poussevin et al. 2016). Poussevin
focused mainly on using the resulting model to cluster stations, not users. Although there have
been some studies using smart card data, there is still a lot of potential for mining the data to better
understand urban mobility.

4.1 Methodology

Currently, Perth’s SmartRider data is sent by the Public Transport Authority (PTA) to a third party
data warehousing company specialising in ticket log data (netBI). netBI provide a data warehousing
service. They clear and archive the log data, and provide statistical reports to PTA. The SmartRider
data used for this project has been sourced from netBI under an agreement with PTA and PATREC.
Table 2: Frequency Band Statistics

<table>
<thead>
<tr>
<th>Frequency</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Trips</td>
<td>12,257</td>
<td>12,070</td>
<td>30,942</td>
</tr>
<tr>
<td>Most Frequent (15min)</td>
<td>07:30am</td>
<td>09:00am</td>
<td>09:00am</td>
</tr>
</tbody>
</table>

The dataset used includes records from October 2016. Each months’ data has three parts: a listing with the bus stops, a listing of the train stops, and the listing of SmartRider transactions. This project will examine the SmartRider transactions only.

A transaction begins with the user tagging on at a particular location, and ends with the user tagging off. The user ID, the time and location they tagged on, as well as the time and location they tagged off, are all recorded. Additionally, the card type (e.g. student, adult, etc.), the distance travelled and the fare paid are recorded, as well as some other metrics.

The dataset used for this project was the SmartRider ticketing logs for October 2016. For analysis, a set of roughly 75,000 card IDs were chosen from the dataset of 420,000 cards. This was done as it is the maximum amount of data the clustering algorithms could handle. These cards were chosen purposefully randomly in order to best approximate the dataset. In order to have a random selection that is similar to the full dataset, the card IDs were chosen to have the same proportion of tokens as the dataset as a whole.

First, we pre-process data into an acceptable format, by removing the errors and converting the logs to appropriate formats for analysis. Next, starting with a database of smart card data, we construct the set of boarding times associated with each user. This representation depends on the events, where an event is a single boarding trip.

Using the boarding times, a user model vector is constructed of probabilities of usage in three frequency bands: high frequency (for events occurring more than twice a week), medium frequency (for events occurring at least once each 10 days), and low frequency (for unusual events). Frequency bands are summarised in Table 2. This representation was experimented with to find the best representation for our task. The resulting model will allow the construction of a vector of probabilities for each event.

We then collected the vectors for all users with activities in the same frequency band. Using non-negative matrix factorization the significant activities were extracted. Using these activities, a clustering algorithm was applied. Specifically, the CLARA algorithm was used, which is a k-means hierarchical method that clusters around medoids rather than means and is optimized for large datasets (Maechler et al 2017). The non-negative matrix factorization used in this particular case is the non-smooth NMF method (Pascual-Montano et al. 2006). The clusters revealed were evaluated by considering real-world interpretations, considering clustering metrics e.g. distance and coherence, and associations between clusters and other passenger information.

The database of trips was filtered so that only unique user, and location couples remained. After this reduction, 355,943 trips from the original dataset of 1,048,540 trips remained. The resulting dataset was then divided into the three different frequency matrices (high, medium, and low) described above. Note that in future studies, the parameters for the frequency bands may be adjusted for different definitions of 'high', 'medium', and 'low'. After the frequency bands were separated, the NMF algorithm was repeated, only per frequency band and aggregating the trips by user (thus dropping the Location data). Activities were extracted by running the nsNMF algorithm on the frequency bands (Gaujoux and Seoighe 2010). After pre-processed the ticket log data, filtered it into three frequency bands and using NMF clustering to identify the core activities in each band,
the final step is to cluster the passenger activities to find common types of passengers. The CLARA clustering algorithm was used to cluster passengers by their activity weights.

4.2 Results and Discussion

The processing procedure we have developed comprises a pipeline: pre-processing the data, building per-passenger time-of-day profiles, using NMF to identify common activities, and then clustering with CLARA to group passengers with similar activity mixes. We applied this processing pipeline to the October 2016 SmartRider data. The high frequency activities generated ten clusters; the medium frequency activities generated seven clusters; the low frequency activities generated nine clusters. All are dominated by one or more of the activities extracted by NMF. For instance, the dominant activity of high frequency cluster D, shown in Figure 4 (top), is a going to work activity combined with a going home activity. This implies that, as might be expected, a common high frequency passenger type is the daily commute activity. However, this passenger type also has hints of other activities as well – a little of a slightly earlier going to work activity, the dominant going home activity, and even less of the other going home activities. In general, each activity corresponds to roughly one cluster that is dominated by this activity. Cluster B, shown in Figure 4 (bottom), is the only high frequency cluster that does not have obvious peaks around commuter peak hours. Cluster B can perhaps be explained when considered with its card types, since it is heavily dominated by Senior SmartRider passengers.

4.3 Conclusion and Future Work

This paper examines some of the passenger clusters that arise in Perth SmartRider data. The results presented here include some of the common high frequency activities found across the dataset. Using non-negative matrix factorization, the dataset was decomposed into a series of activities, and the passengers were clustered based around those activities. This project discovers that decomposing the ticketing logs into frequency matrices and running NMF on these matrices produces a
viable set of activities that can be used to improve the quality of discovered clusters. This approach can be used to run over different sets of the data, or over different frequency matrices, in order to discover interesting passenger classes, some of which have been discussed in this paper. Overall, this has been yet a preliminary study of the user clusters in Perth’s SmartRider data. There is still work that can be done on this dataset and other smart card data of this type. For instance, the effects of other clustering algorithms can be studied in order to compare the various types of clusters found. Additionally, comparison to different types of NMF, such as the Brunet method or the ALS method, can be undertaken. Perhaps, other information than just boarding time and location can in included, such as alighting time and location, in order to create more finely grained clusters. Additionally, boarding location can be expanded to include boarding zone, whether it is the STEM zones used by the Department of Transport or some other zones. Other information, such as the total kilometers travelled, can also be considered. In short, full use can be made of the rich dataset available.

5 Passenger Analytics - Continuing Project

When validating the passenger clusters discovered in the research described in Section 4, we found some unexpected passenger behaviours such as one-way-only commuting. So in an ongoing project we are investigating additional ways of characterising passenger types to better understand these patterns.

5.1 Passenger Travel Profiles

Figure 5 shows the travel profiles for a sample of 6 individuals from the database. Each bar in the graphs represents the number of journeys by this passenger that started at a certain hour of the day. Journeys are counted over a 4 week period. The horizontal line on each group, at 6 journeys per 4 weeks, is the cut off we have chosen to distinguish between ad hoc travel, where the passenger occasionally starts a journey during a particular hour of day, and regular travel, where the passenger often starts a journey during a particular hour of the day. All weekdays and weekends are counted over a 4 week period, so a passenger who travels every week day at 8am would have a bar for 20 trips at 8am. It can be seen that individual passengers may have a mix of regular and ad hoc travel patterns or only one. Each passenger was classified by two profiles: their ad hoc travel profile and their regular travel profile, each filtered from the passenger’s full travel profile. Passengers may have an empty ad hoc or empty regular profiles. In fact, 73% of passengers have no regular travel times; they only have ad hoc travel times.

The following features were used to characterise each passenger. Selected features are those that capture interesting characteristics of passengers, particularly those features that varied for different SmartRider token groups.

**token** The 15 types of Transperth token from the dataset were aggregated into 5 groups: Concession (9), School (2), Senior (2), Standard (1), Tertiary (1).

**railpct** Percentage of all trips taken by rail. Bus trips are usually 100 minus railpct since only 2.4% passengers have any ferry journeys.

**tripsperjourney** The number of trips per journey. Trips are each individual on/off transaction for a passenger and journeys are a sequence of trips taken with 2 hours. This metric is averaged over all trips and journeys in a 4 week period.
minsperjourney Total travel time divided by the number of journeys. The results are grouped into four categories: 30, 60, 90 or 120 (any >90) minutes.

wkendpct Percentage of journeys that occur on a Saturday or Sunday.

starttime The hour of day when the first regular travel time occurs (only for regular travellers)

stayhours Number of hours between the two most frequent travel hours (only for travellers with 2 or more regular travel times)

5.2 Passenger Types

Using the features described above, we created a database of passenger characteristics. By partitioning this database in different ways, we arrive at a flexible model for passenger types. That is, passenger types are not a fixed typology, but rather a flexible typology that can be adjusted according to different modelling needs.

The top row of Figure 6 (left) shows three main types of passengers: those with only ad hoc travel times (ie no regular travel), those who have only one regular travel time per day, and passengers with two (or more) regular travel times per day: that is, traditional commuters. Each box represents a group of passengers, with P being the percentage of passengers and J the percentage of journeys for each group. All figures are rounded to the nearest percent. It can be seen that by far the largest passenger group is ad hoc (only) travellers, who make up 74% of all passengers who used their SmartCard at least once in the month, whilst traditional twice a day commuters were only 13% of all passengers and 37% of all journeys. The surprising group of one-way-only regular travellers are 13% of all passengers and 23% of all journeys. This high level typology of 3 types of passengers
Finer grained analysis of passenger behaviours are shown in the lower level boxes of Figure 6 (left). These provide further partitioning based on additional features associated with each passenger. For example, the ad hoc only passengers can be partitioned into rare or frequent travellers: rare have 4 or fewer different travel hours and frequent have more than 4 different travel hours. The regular once-a-day passengers are partitioned by their start times. Two groups are shown: those whose single regular journey is in the morning (7% of all passengers), and those whose journey is in the afternoon (4% of all passengers). Of the twice a day commuters most (6% of all passengers) have an 8 or 9 hour stay while 3% have shorter, 6-7 hour stays, and 4% stay 10 or more hours.
5.3 Passenger Narratives

Partitioning passenger types can give aggregated statistical summaries of each group: the number of passengers and their significance. However, it may be more helpful to describe passenger types using a narrative of their characteristics. The challenge is, how can we learn meaningful narratives for a given passenger type using automated analysis?

In a preliminary study, we have used association rule mining to generate narratives about different passenger types. The database of passenger features was mined for interesting associations using the Apriori method. This resulted in a set of 11,202 rules. Our current work is looking at how to select the most interesting rules from this database: “moving beyond the deluge of the obvious”. Table 7 shows a set of rules that characterise one of the surprising passenger groups R1A1: once-a-day regular passengers who rarely travel at other times. The rules are read from left to right. For example, row one is read: Passengers who rarely travel on the weekend and whose regular travel time is usually 3pm or 4pm are likely R1A1 passengers. Similarly, row three is read: Passengers with School token SmartRiders and whose regular travel time is usually 3pm or 4pm are likely R1A1 passengers.

The last three columns give some metrics for the rules. S is support: the proportion of all passengers that this combination of attributes applies to. For example, rule 1 applies to 2% of all Transperth passengers. Since the SmartRider data is dominated by ad hoc, rare travellers, most of the regular patterns have relatively low support in terms of percentage of passengers. However, they are more significant in terms of percentage of trips and fares.

C denotes confidence: how certain the rule is. That is, how many passengers who satisfy the left hand side of the rule also satisfy the right hand side. For example, rule 1 tells us that 90% of passengers who do not travel on weekends and do travel once a day at 3pm or 4pm are also R1-A1 passengers. The confidence for rule 5 in row 5 tells us that 70% of all passengers who travel at 3pm or 4pm are in group R1-A1.

L denotes lift: how surprising or interesting the rule is. Interestingness can be assessed either subjectively or objectively. The lift metric is intended as an objective measure of interestingness. A rule, A => B, is said to capture a spurious correlation between attributes if the probability of A or B occurring is the same as the probability of A times the probability of B. L=1 means that events A and B are independent (in the probability sense) so the rule could have occurred by chance and it is not surprising or interesting. Values of L>1 means that A and B are positively correlated, with higher values indicating higher “surprisingness”. Figure 7 shows only rules that are strongly surprising by this objective measure.
Understanding Group  R1-A1  P=7% J=10%
(once a day regular & rare ad hoc)

<table>
<thead>
<tr>
<th>RULES</th>
<th>S</th>
<th>C</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>{wkendpct=0,star/mes=1516} =&gt; {gp=R1-A1}</td>
<td>0.02</td>
<td>0.90</td>
<td>13</td>
</tr>
<tr>
<td>{railpct=0,starMmes=1516} =&gt; {gp=R1-A1}</td>
<td>0.01</td>
<td>0.90</td>
<td>13</td>
</tr>
<tr>
<td>{token=School, star/mes=1516} =&gt; {gp=R1-A1}</td>
<td>0.02</td>
<td>0.81</td>
<td>12</td>
</tr>
<tr>
<td>{tripspertjourney=1,starMmes=1516} =&gt; {gp=R1-A1}</td>
<td>0.02</td>
<td>0.77</td>
<td>11</td>
</tr>
<tr>
<td>{star/mes=1516} =&gt; {gp=R1-A1}</td>
<td>0.02</td>
<td>0.70</td>
<td>10</td>
</tr>
<tr>
<td>{minspertjourney=60,starMmes=1516} =&gt; {gp=R1-A1}</td>
<td>0.01</td>
<td>0.65</td>
<td>9</td>
</tr>
</tbody>
</table>

Figure 7: Some characteristics of one-way-only regular passengers (group R1A1)

6 Journey-centered Analytics - Continuing Project

This section is by CEED student Jordan Lilburne and supervisor Jianxin Li

The goal is to provide PATREC with the ability to query SmartRider logs efficiently concerning the regions of the Transperth transit network with the greatest levels of passenger mobility for a specified time interval.

The main challenge is how to compute heavy subgraphs for a specified time-window of transit activity efficiently, for the purposes of querying for regions of heavy usage during a defined time window from the datasets of SmartRider journeys. The activity of the transit network will be modelled by generating the passengers’ position in the network by combining the dataset of tag on/off time and location, and the dataset of the bus timetable.

The transit activity network is represented as graph which supports a stream of trip segments represented by transit stops as nodes, transit stops which are connected as consecutive stops in at least one trip as the edges, and the edge weight for a given time represents the number of tagged on passengers.

There is also a sub-problem of how to maintain a representation of the changes in edge weights over time so that multiple queries for different time windows does require the entire network to be reconstructed from the dataset. Another sub-problem being investigated is how the inclusion of spatial information can be included as a constraint to the pattern found in dense graph discovery, in order to vary the bias with respect to region size.
7 Spatial Hub analytics - Continuing project

This project is being undertaken by Masters of Professional Engineering student Xiaolin Hu, commenced August 2017, supervised by Rachel Cardell-Oliver. The goal is to study the spatial properties of inward and outward journeys for a specified hub region, such as a shopping centre or business district. Spatial hub analytics will provide insights about hub supply vs. demand from a local viewpoint. This project builds on the hub-analytics project of Travis Povey, and will deliver ways to visualise answers to queries such as: For a given hub location, what is the total in-flow and out-flow of passengers from that hub? What are the most common next hop and last hop locations from a given hub? What are the most common routes used in and out of a given hub? What shape is the catchment area (trips or journeys) for a given hub? How do each of the above characteristics vary between hubs and over time?

8 Results Highlights

This section highlights some of the most interesting findings from project 4.2 to date.

• A new method was developed for learning spatial hubs from in the Perth network. These hubs are areas of the network with high passenger numbers staying in the area. Passenger transfers are excluded. Each hub has a unique mixture of activities. The activity mixes at each hub were found to correlate well with known nearby points of interest. We also found that existing spatial models such as STEM and ROM zones were not suitable for hub activity analysis.

• The hub discovery method found 5 types of hub activities in the network. Activities are characterised by the time of arrival and stay time there. Correlating these activities with hub location and known points of interest, we named the hub activities: short outings, university, school, work, and residential.

• Although hub use may be dominated by an obvious activity, the mixture of additional activities can be surprising. For example, at the University of Western Australia hub area, ad hoc long (3-10 hour) stays (student day) comprise 69% of stays. The remaining stays comprise ad hoc short (2-4 hour) stay (11%), work day (9%), school stay (7 hours) (8%) and overnight stay (3%).

• A new, flexible model for classifying passengers was learnt from the SmartRider data. As well as standard visual representations, narrative descriptions can be generated for each passenger type.

• Passenger types discovered from the SmartRider data include some new and significant behaviours: ad hoc travellers dominate the data (74% of all passengers), and there is a small but significant group of only-once-a-day regular travellers (13%). Understanding these types of passengers offers new opportunities for increasing the patronage of public transport in Perth.

9 Conclusions and Future Work - What next?

Project 4.2 has delivered methods for querying, analysis and data mining, to support a knowledge discovery process centred on passengers, hubs, and journeys. Using 2016 SmartRider data we have
discovered some surprising usage patterns. Our systems enable policy makers to able to generate evidence-based answers automatically from SmartRider data.

This project has demonstrated that there is great potential for knowledge discovery from SmartRider data. An important avenue for future research is integrating SmartRider knowledge discovery into models for integrated land use and transport futures. SmartRider knowledge can also be used in the design of new behaviour change instruments and for evaluating the results of behaviour change interventions. Another future direction is to improve the usability of our results through integration with decision support tools such as the web portal developed in Project 4.1.

10 References


11 List of Project Outputs

Project 4.2 has met its agreed milestones and completed the project deliverables (see Section 12.2). Some research streams from the project will be continuing into 2018. The following outputs are available from PATREC on the dates shown. If desired by the Steering Committee, a project web page could be developed to archive these outputs for PATREC members.

11.1 Decision-Support Software, Documentation and Visualisations

2. Station and Stop usage profiles (R code) and Top 120 stations (pdf), R Cardell-Oliver, (Mar 2017)
3. Passenger Analytics (R code, R data files), R Cardell-Oliver and Anh V, (Sep 2017)
4. Hub Analytics (R code, R data files) and user manual (pdf), T Povey, (Nov 2017)

11.2 Data Sets of Results

5. Passengers and their features (Nov 2017)
6. Hubs and their temporal features (Nov 2017)
7. Hubs and their spatial features (June 2018)
8. Journeys and their features (June 2018)

11.3 Academic Reports

9. Literature Review on passenger-centered analytics, L Dokuchaeva, (Nov 2016)
10. Literature Review on Hub Analytics, T Povey, (May 2017)
12. CEED paper on passenger Analytics, L Dokuchaeva, (Sep 2017)
13. CEED paper on Hub Analytics, T Povey, (Sep 2017)
17. Spatial Analytics, Masters Thesis, X Hu (July 2018)
11.4 Workshops with Policy makers and Stakeholders

18. UWA CEED student internships with PATREC partners
   Lidia Dokuchaeva. Jan-Feb 2017 in the Transport Modelling group in Integrated Transport Planning group of the Department of Transport mentored by Alan Kleidon and Renlong Han.
   Travis Povey Jan-Feb and July 2017 in the Network Operations Analysis group of the Network Operations Directorate at the Department of Mainroads mentored by Graham Jacoby.

19. Workshop with the Department of Transport Team, 18 September 2017

20. PATREC Forum, 27 November 2017 (date TBC), project presentation

11.5 Refereed Research Publications

*In progress.*

12 Project Management

12.1 Budget

<table>
<thead>
<tr>
<th>Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEED project on passengers</td>
<td>$18,000</td>
</tr>
<tr>
<td>CEED project on hubs</td>
<td>$18,000</td>
</tr>
<tr>
<td>CEED project on journeys</td>
<td>$18,000</td>
</tr>
<tr>
<td>Summer student scholarships</td>
<td>$10,000</td>
</tr>
<tr>
<td>Publication scholarships</td>
<td>$11,000</td>
</tr>
<tr>
<td><strong>Total budget</strong></td>
<td><strong>$75,000</strong></td>
</tr>
</tbody>
</table>
12.2 Timeline and Milestones

<table>
<thead>
<tr>
<th>Milestones</th>
<th>Summary of Activity</th>
<th>Key deliverables</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project management</td>
<td>Finalise details of project; Agreement on use of data sources; 3 progress meetings</td>
<td>Project plan to be agreed by PATREC and P4 Steering Committee August 2016</td>
<td>✔</td>
</tr>
<tr>
<td>Data management</td>
<td>Data specification, discovery and collection of sample data sets for analysis; Data pre-processing; Data storage</td>
<td>Progress report to Steering Committee December 2016</td>
<td>✔</td>
</tr>
<tr>
<td>Data mining algorithms development</td>
<td>Literature surveys on existing algorithms; Feature engineering to derive relevant per-passenger and per-hub and per-journey features for analysis; Population data mining algorithms to identify clusters of similar passengers / journeys / hub activities; Presentation of preliminary results: visualisations and reports</td>
<td>Workshop on results and their presentation April 2017</td>
<td>✔</td>
</tr>
<tr>
<td>Dissemination of results</td>
<td>Analysis and reporting results on extended data sets; Delivery of data-mining software; Embed the data-mining supported queries software into the Project 4.1 framework where possible</td>
<td>Draft report June 2017; Final report July 2017; Dashboard prototypes with Project 4.1; Dashboards were only developed for project 4.1; 4.2 outputs are as data tables and visualisations</td>
<td>✔</td>
</tr>
</tbody>
</table>

12.3 UWA Research Team and Expertise

Rachel Cardell-Oliver, Assoc. Professor Computer Science: sensor networks and data mining
Jianxin Li, Senior Lecturer Computer Science: databases and graph algorithms
Wei Liu, Senior Lecturer Computer Science: data mining and text mining
Lidia Dokuchaeva, Xiaolin Hu, Jordan Lilburne, Travis Povey, CEED scholars
Robert Banks, Kit Buckley, Yan Ji, Bryan Trac, Andrew Briscoe, Summer scholars

12.4 Project 4 Steering Committee

Damien Martin, Department of Planning
Renlong Han, Department of Transport
Sue Hellyer, Department of Transport
Tom Pacy, Public Transport Authority
Wes Soet, Main Roads
Laura Cook, Treasury
13 Acknowledgements

Author LD would like to extend a thank you to my supervisor, Rachel Cardell-Oliver, the staff at CEED and PATREC for supporting me with this project. Additionally, I would like to thank the Department of Transport for their support, in particular Renlong Han and Alan Kleidon.

Author TP would like to thank the rest of the research group in this project, the sharing of ideas and results has helped progress greatly, Chao Sun for his advice and assistance in acquiring census data that has been invaluable to the project, and Graham Jacoby who helped me gain an understanding of transport modelling and his mentorship during my internship. Lastly, I would like to thank PATREC for supporting the project, and the Public Transport Authority and Transperth for providing the data that made this research possible.

All the project team would like to thank PATREC for supporting project 4.2 and so testing the potential for Computer Science techniques to help to solve challenging problems in the cross-disciplinary field of transport modelling.