Activity identification and primary location modelling based on smart card payment data for public transport

Working Paper

Author(s):
Chakirov, Artem; Erath, Alexander

Publication date:
2012

Permanent link:
https://doi.org/10.3929/ethz-a-007328823

Rights / license:
In Copyright - Non-Commercial Use Permitted
Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport

Artem Chakirov
Dr. Alexander Erath

Future Cities Laboratory, Singapore – ETH Centre
Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport

Artem Chakirov  
Singapore – ETH Centre  
Future Cities Laboratory  
1 CREATE Way  
#06-01 CREATE Tower  
Singapore 138602  
Mobile: +65 9638 9737  
Email: chakirov@ivt.baug.ethz.ch

Dr. Alexander Erath  
Singapore – ETH Centre  
Future Cities Laboratory  
1 CREATE Way  
#06-01 CREATE Tower  
Singapore 138602  
Mobile: +65 9107 6574  
Email: erath@ivt.baug.ethz.ch

Abstract

The introduction of electronic payment systems for public transport in many cities all over the world enabled collection of detailed and comprehensive data records of public transport journeys. Processing and mining of these data opens new opportunities in transport modelling and travel behaviour research. In particular its potential for identification and analysis of activities, which cause regular travels, represents a highly interesting research question.

In this paper, opportunities for detection of primary activities, as home and work activities, and their locations based on records of smart card fare payment system for public transport are investigated. Thereby the city-state of Singapore is used as a case-study. In order to gain information about country specific activity characteristics and allow accurate model calibration, data from household travel survey is used. In particular, two major activity detection models are considered: a simple rule-based model using only activity duration for detection of work activities and more involved discrete choice modelling approach using additional variables as activity start-time and land-use information, and distinguishing between home, work and other activities. In particular, the benefit resulting from including land-use information into the discrete choice model is evaluated and models with and without it are compared.

The developed models are applied to the full record of public transport journeys in Singapore and outcomes are compared against each other. In conclusion, challenges associated with analysis of longer time periods consisting out of journey data for several days are discussed.

Keywords

Smart card data, activity location modelling, public transport, working locations

Preferred Citation

1 Introduction

The emergence of information technology during the last two decades is predominantly responsible for the strongly increased interest in the activity-based approach for transport demand modelling and travel behaviour analysis. Travel can be considered as one of the essential attributes of many major activities, as it is often inevitable for accessing the location where the desired activity will be performed. At the same time obtaining detailed information on activities of people and their spatial and temporal distribution in a certain region is usually a very laborious and expensive endeavour.

However, the rapidly increasing number of electronic automatic fare collection (AFC) systems for public transport provide a new substantial source of information, which can be used as a supplement for the traditional travel surveys. An important aspect of an electronic fare collection system is the detailed data records that are continuously generated and archived. Analysis of these data can provide valuable insights into the usage of public transport and help to better understand people’s travel behaviour and the underlying motivation for their journeys in the form of performed activities.

In this context, the example of Singapore represents an interesting study case as a highly dynamic city with a centralized, long-term oriented transportation planning policy and fast growing commuter population. The combination of a distance-based fare scheme for public transport and the use of contactless stored-value smart cards for payment of public transport fares, provide a nearly comprehensive data record of public transportation usage for the entire city. Together with the traditionally available survey data on travel behaviour as well as the land-use information, such record offers potential for the identification of mobility and activity patterns within the area served by public transport.

This paper focuses on detection and identification of primary activities, as home and work activities based on public transport smart card records. It starts with a brief overview of the recent research on analysis and enrichment of smart card data as well as published approaches to the activity location modelling. Afterwards, in Section 3, the AFC system implemented in Singapore is introduced and the available data sets as the smart card data record, the Household Interview Travel Survey 2008 (HITS) and the Master Plan 2008 containing the land-use information are described. In the following Section 4 two concept for the detection of the primary activity type and location, based on the record of all public transport journeys are presented. First of them is using a simple rule based model, second is applying a discrete choice modelling approach with additional variables as start time and land-use information. Results of application of both models to the smart card journey record, are described in Section 5. In conclusion, an outlook on application of developed model on multiday data is given as well as the direction of future research is indicated.

2 Related Work

In recent years the number of published articles and papers on smart card data use in public transport have increased. A comprehensive literature review on this topic is offered by Pelletier et al. (2011), where three categories of studies have been identified: strategic level studies with focus on long-term planning, tactical level studies related to schedule adjustment, load profiles and transfers and operational-level studies analysing performance indicators of public transport.
network and smart card operations itself. Research, primary focused on the spatial travel information provided by AFC systems, has been conducted by Li et al. (2011) and Munizaga et al. (2011), where Origin-Destination Matrices are estimated using smart card data. Moreover recent work with the focus on commercialization by Páez et al. (2011) presents a geodemographic analysis, which links the smart card data to the household survey data and business data points to identify potential commercial partnerships near metro stations in Montreal.

However, only few research papers address modelling of activity types and locations implicitly based on travel data. Ahas et al. (2010) used positioning data from mobile phones to identify activity locations meaningful to users. Chu et al. (2010) presented a methodology for characterization of trips based on socio-demographic characteristics, multiday travel patterns and association of travel with specific locations. More recently Hasan et. al (2012) present a simplistic mobility model for prediction of people’s location selection, based on the number of visits to certain places and using subway smart card data from London. Strategies for extension of these concepts towards more general activity detection and location modelling based on smart card data form the goal of this work.

3 Used Datasets

3.1 Smart card fare payment system in Singapore

The AFC system for public transport based on contactless, stored value smart cards was introduced in Singapore in April 2002. Today, smart cards can be used island wide for payment of all modes of public transport, regardless of operator. Though cash payment of single fares at higher rates is still possible, e-payments using smart cards account for 96% of all trips, which makes the smart card data records highly comprehensive and the missing cash paying travellers negligible (Prakasam, 2008).

The implementation of an uniform smart card AFC system allowed the introduction of a distance-based fare scheme for all modes of public transport in Singapore. The fare charge for each customer is based exactly on the distance travelled, transport mode and demographic attributes as there are prioritized rates for children, students and senior citizens. Customers have to tap their smart card on the reading device every time they enter and leave a train station or a bus. Thus, besides information on boarding time and location, the data collected from smart cards contains detailed records of alighting times and destination location for both bus stops as well as Mass Rapid Transit (MRT) and Light Rail Transit (LRT) stations. These attributes distinguish the Singaporean smart card data from those collected by the majority of other AFC systems and allow more detailed assessment of mobility patterns. Furthermore, as the smart cards are easily rechargeable, people tend to continuously use a single, uniquely identifiable card for all their public transport journeys for substantial periods of time. As the technical setup of the system doesn’t allow more than one person to travel on a single card, it can be assumed that each card represents a single person. This enables highly disaggregated analysis of individual itineraries and opens new ways for understanding people’s travel behaviour on short as well as long term scales.

3.2 Available smart card dataset

The public transport journey dataset used in this paper was recorded for one complete week, Monday till Sunday, in April 2011. It contains all trips paid by smart cards in Singapore during
this time period. Each trip record includes, among other attributes, an encrypted unique card number, boarding station and boarding time, trip duration, alighting station as well as the passenger type, which allows differentiation between Child/Student, Adult and Senior Citizen. Original dataset was further processed using a MySQL open source database in combination with the programming languages R and Java. Furthermore all train stations and bus stops were geocoded using location information of MRT/LRT stations and bus stops provided by LTA. New stops, which were not yet included into the geo-location database, were geocoded manually using Google Maps.

In the context of this research, only trips during working days - Monday to Friday – are considered as relevant, since the identification of work activities is a primary goal. Furthermore, 2.8% of all trips didn’t have any alighting time and location due to the missing tap-out and were removed from the data set.

In this paper we define a journey as a one way travel from one activity to another. Each journey consists of one or several consecutive journey stages or trips on same or different modes. As this work focuses on the activities between the journeys, a rather simple rule to distinguish between activities and journey stages is applied. If a time span between last alighting and consecutive boarding is equal or longer than 1 hour, we consider the new trip to be a start of a new journey.

3.3 Singaporean Household Travel Interview Travel Survey (HITS)

The Household Interview Travel Survey (HITS) is conducted by Land Transportation Authority in Singapore every 4 years and involved over 10,500 qualifying households during its latest conduct in 2008 (Land Transportation Authority, 2008). With qualifying households including Singaporean citizens, permanent residents and legal immigrants residing in Singapore, it represents a comprehensive sample of more than 38’000 participants out of 4.839 million people legally residing in Singapore in 2008 (Media Research and MVA, 2009).

In the context of analysis of smart card fare payment data and detection of activity locations, HITS data can be used for model estimation and calibration as well as serve as a valuable source for enrichment of the AFC smart card data and supplement it with detailed information on usage of different transportation modes.

3.4 Master Plan 2008

The Master Plan 2008 (Urban Redevelopment Authority, 2008) developed by Urban Transportation Authority of Singapore, contains detailed information on the current and planned land-use on the island. Thereby, it divides the island into about 11,000 single zone, comparable with a size of a parcel, and assigns each of them one of 31 different land-use categories. Furthermore, for each zone it contains the total area size and the maximal allowed gross plot ratio and hence enables the estimation of maximal gross floor area, which is the product of the area size and the gross plot ratio, within each zone. By aggregating these estimates in proximity of certain locations, as for example bus stops or train/subway stations, maximal floor area for major land-use categories, as for instance total residential or commercial floor space around these locations, can be obtained.
4 From public transport journeys to activity locations

4.1 Activities as reason for travelling

People usually travel from one place to another because they want to perform an activity, which cannot be performed in the desired way at their current location. The activities can be of various types and durations as there are work, education, leisure and social activities for example. Hence, in order to understand peoples travel behaviour and therewith flows, mobility patterns and traffic volumes, it is necessary to look at people’s daily activities and their locations. This approach is also used in modern agent-based transport simulations like MATSim, which include spatial and temporal information about activities in its model (MATSim-T, 2011). In absence of comprehensive data sources such as a whole population census, one of the main challenges in modelling activities results from lack of verified data of high spatial resolution on people’s home and particularly work locations. However such information is not only important for transport planning and implementation of agent-based transport models, but is also very valuable in the areas of urban development and land-use planning. Travel patterns observed in smart card data from public transport can provide valuable information on people’s primary activity locations and help to verify and refine existing models and assumptions.

In the broader context, primary activities are the major activities performed by a person on a regular working day with significant distinction from other activities in terms of duration. In this work the term primary activity is used as a synonym for home and work activity as these are dominant activities and locations for the majority of the population during a regular work week.

4.2 Consistency of public transport journey chains

In order to make statements about activities of a particular person, the recorded daily journey chain of this person needs to be consistent. Consistency in the context of AFC smart card data record means that the person who arrived to the activity location by public transport, has to leave it after ending the activity also by public transport, otherwise the activity duration can’t be extracted. The assumption of consistency is hard to verify based on AFC data only, as the use of any other means of transport except from public transport, e.g. walking, cannot be detected. However, the most obvious cases of inconsistency can be identified by analysing the distances between the alighting location of last journey and the boarding location of the following journey. Using this method, in-between trips by other means of transport such as taxi or car can be recognized. For the rest of the paper detectable activities within the consistent public transport journey chain will be referred to as consistent PT activities.

The analysis of distances between public transport journeys on typical workday (Tuesday) is shown as a cumulative relative frequency graph in Fig. 1. Only persons with more than one journey recorded in the smart card data record for this day were evaluated. The graph shows that 90% of journeys, which are following a preceding one, start less than 1 km away from previous alighting location. This indicates that the majority of public transport users don’t switch to other modes of transport between public transport journeys and therefore have consistent journey chains.
A slightly different picture is obtained when including also persons with only one observed trip per day and investigating distances between first boarding and last alighting station of the day, which in case of them being close by, gives a strong indication for home location. As shown in the cumulative relative frequency graph of these distances for all persons in the data record for this day (Fig. 1), 63% return to the station, which is located in radius of 1 km around their first departure station of the day. This is for the major part a logical consequence of the fact, that about one third of all public transport smart cards are used for a single journey per day. Furthermore, based on analysis of HITS 2008, this aspect of travel behaviour in Singapore can be also explained by high popularity of pick-ups and drop-offs by a family member or a friend driving a private vehicle. Strong tendency to avoid the use of public transport for the first or last mile of the trip, combined with high costs of owning a car, which encourages carpooling are the major reasons for this behaviour.

These observations have strong implications on the potential of detecting work activities based on public transport smart card records. When inflating work activities detected by analysing public transport smart card data to total expected work activities, one needs to consider not only the public transport mode share, but also include inconsistent PT activities (see Section 5.2).

### 4.3 Rule based approach - detection of work activities and locations based on activity durations

As mentioned in Section 4.1, detailed information about the location of workplaces in a city is highly important for transport and land-use planners. Unfortunately such spatially highly resolved information is usually not available and only rough models exist to estimate these numbers. In this section a simple method based on the activity duration threshold is investigated in order to detect working locations and their spatial distribution in Singapore based on the record of public transport journeys.

One of the most important indicators for type of activity performed between two consecutive journeys is its duration. Also in this case HITS provides a valuable source of information for understanding the typical durations of work activities in Singapore.
Fig. 2. Distribution of activity durations in HITS 2008. Dark grey coloured is the share of consistent PT activities, as defined in Section 4.2. Light grey is the share of consistent PT work activities from all public transport activities.

Fig. 2 shows the histogram of activity durations for activities equal to or longer than 1 hour, which were reported in HITS in form of trip purpose and destination. The dark grey coloured part illustrates the share of consistent activities involving public transport return journey. The light grey coloured part is the share of work activities. The majority of performed activities is either short, with duration ranging from 1 to about 5 hours or longer, with duration from 7 to 12 hours. The share of consistent PT activities increases with longer activity duration. For shorter activities transportation to these activities seem to predominantly involve other modes of transport as taxi, private car, motorcycle etc. More interesting in the context of activity detection though, is the fact that consistent PT activities with duration more than 6 hours are predominately work activities, which for their part account for the vast majority of all consistent work activities reported in HITS. Based on this observation, a simple rule based model is formulated: all consistent activities with duration between 6 and 16 hours are work activities. Based on this rule, the number of work activities with duration 6-16 hours is overestimated, but as the work activities shorter than 6 hours are not accounted for at all, the total number of detected work activities equals to 94% of all work activities reported in HITS.

4.4 Discrete choice modelling approach with overnight activities

Another possibility for activity identification based on public transport journey record, is the application of the discrete choice modelling approach. Thereby, a multinomial logit model with the choice set consisting out of three alternatives: work activity, home activity and other activity is developed. As utility variables duration and start time of each activity as well as spatial variable of land-use in the neighbourhood of the particular location are used. A model based on these parameters can be calibrated with HITS data and is expected to provide a higher accuracy and reliability of work activity detection, which is the primary goal in this context. As HITS data provides a journey record only for one day, the duration of overnight activity for the night prior as well as the night past the recorded day, is unknown. However the overnight activity is particularly important for the detection of home activities as well as shift-work activities. In order to overcome this difficulty, it is assumed that the type and
location of the overnight activity corresponds to the journey purpose, arrival time and destination location of the last reported journey of the day. Therefore, the duration and consistency of the overnight activity is determined by start time and location of the first activity of the reported day.

After removing journey records containing obvious errors and inconsistencies, total of 40’113 activities could be identified within the HITS record. Out of these activities, 15’459 activities were found to be PT consistent, longer than 1 hour and contain the information on activity location or alighting public transport station or stop. Table 1 shows distribution of major activity types within these activities:

<table>
<thead>
<tr>
<th>Activity Type</th>
<th>Total Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>7539</td>
<td>48.8 %</td>
</tr>
<tr>
<td>Work</td>
<td>6143</td>
<td>39.7 %</td>
</tr>
<tr>
<td>Other</td>
<td>1777</td>
<td>11.5 %</td>
</tr>
<tr>
<td>Total</td>
<td>15459</td>
<td>100 %</td>
</tr>
</tbody>
</table>

Table 1. Distribution of all PT consistent activities in HITS 2008, including overnight activities

In following three major utility terms used in utility functions of the model are analysed:

**Activity Duration**

The activity duration is the most important indicator of the activity type. Fig. 3 and 4 show the number of all activities, consistent PT activities as well as consistent only home and only work activities, dependent on the activity duration extracted from HITS. The majority of home activities have durations longer than 9 hours, with the peak of the distribution located between 12 and 13 hours. Compared to home activities, work activities are significantly shorter, with the duration range mainly between 7 and 11 hours and with distinct distribution peak between 8.5 and 9 hours. Fig. 5 and 6 present the share of consistent home and work activities from all consistent activities respectively, plotted against activity duration and aggregated for per hour.

Distribution of this share over activity duration serves as the model for design of duration dependent term within the utility function. In case of home activity the distribution of share from all activities resembles a sigmoid curve. Thus, a simple logistic function is used to estimate the utility term:

\[
U_{\text{durationHome}} = \beta_{H1} \frac{1}{1 + e^{\beta_{H2} \cdot t + \beta_{H3}}} 
\]

(1)

with 3 estimated parameter: \( \beta_{H1-H3} \) and the activity duration \( t \).
Fig. 3 Number of all reported activities (white), consistent activities (dark grey) and consistent home activities (light grey) dependent on activity duration [HITS 2008].

Fig. 4 Number of all reported activities (white), consistent activities (dark grey) and consistent work activities (light grey) dependent on activity duration [HITS 2008].

Fig. 5 Share of home activities from all consistent activities dependent on activity duration and estimated utility function for home activity duration.

Fig. 6 Share of work activities from all consistent activities dependent on activity duration and estimated utility function for work activity duration.

For work activities, a piecewise linear, continuous function with total of 2 parameters $\beta_{W1-W2}$ was chosen in order to account for activity duration within the utility function of work activity.

\[
U_{durationWork} = \beta_{W1} \cdot t_1 + \beta_{W2} \cdot t_2
\]

with

\[
t_1 = \begin{cases} 
0 & \text{for } 0 \leq t < 9 \\
9 & \text{for } t \geq 9
\end{cases}
\]

\[
t_2 = \begin{cases} 
0 & \text{for } 0 \leq t < 9 \\
t - 9 & \text{for } t \geq 9
\end{cases}
\]

As can be seen from the Fig. 6, the utility of very long activities does not decrease as strongly as it would be expected based on the share of longer work activities with durations over 15 hours. This can be attributed to the small number of observations of work activities for these durations. The few one, which are observed though, start at the time of day with low start time utility for
work activity and/or are located in residential areas. Therefore, they can be only detected based on their duration, which prevents the duration utility to decrease further.

**Activity Start Time**

As public life and business activities commonly follow a regular daily schedule, time of day when a certain activity is performed can be an important indicator for its type. More than 90% of employed HITS participants reported to have fixed working hours. If an activity starts in the morning, the probability of it being a work activity is higher, than for activity which starts towards the end of the day, which is more probable to be associated with the arrival at home. Fig. 7 and 8 confirm this statement, showing number of all reported activities, all consistent PT activities and only home or only work activities respectively, dependent on the start time of each activity. Almost all work activities start as assumed in the morning, between 6am – 11am. In contrast, vast majority of home activities start between 16pm – 23pm.

Fig. 9 and 10 show the shares of consistent home and work activities from all consistent activities, plotted against activity start time and aggregated by hour. For home activities 3 dummy variables with corresponding estimated parameters $\beta_{H4–H6}$ are included into the utility function. In case of work activity start times 4 dummy variables with estimated parameters $\beta_{W3–W6}$ are used. Number of dummy variables thereby is partially limited by low number of observations for particular activities within certain time bins. The contributions of start time variable to utility functions is therefore as follows:

$$U_{\text{startTimeHome}} = \beta_{H4} \cdot \text{StartTime}_{2–11\text{Home}} + \beta_{H5} \cdot \text{StartTime}_{12–15\text{Home}} + \beta_{H6} \cdot \text{StartTime}_{16–1\text{Home}}$$

with

$\text{StartTime}_{2–11\text{Home}} = 1$ for $2 \leq t < 12$, otherwise 0

$\text{StartTime}_{12–15\text{Home}} = 1$ for $12 \leq t < 16$, otherwise 0

$\text{StartTime}_{16–1\text{Home}} = 1$ for $16 \leq t < 2$, otherwise 0

$$U_{\text{startTimeWork}} = \beta_{W3} \cdot \text{StartTime}_{5–8\text{Work}} + \beta_{W4} \cdot \text{StartTime}_{9\text{Work}} + \beta_{W5} \cdot \text{StartTime}_{10–16\text{Work}} + \beta_{W6} \cdot \text{StartTime}_{17–4\text{Work}}$$

with

$\text{StartTime}_{5–8\text{Work}} = 1$ for $5 \leq t < 9$, otherwise 0

$\text{StartTime}_{9\text{Work}} = 1$ for $9 \leq t < 10$, otherwise 0

$\text{StartTime}_{10–16\text{Work}} = 1$ for $10 \leq t < 17$, otherwise 0

$\text{StartTime}_{17–4\text{Work}} = 1$ for $17 \leq t < 5$, otherwise 0.
Fig. 7 Number of all reported activities (white), consistent activities (dark grey) and consistent home activities (light grey) dependent on activity start time [HITS 2008].

Fig. 8 Number of all reported activities (white), consistent activities (dark grey) and consistent work activities (light grey) dependent on activity start time [HITS 2008].

Fig. 9 Share of home activities from all consistent activities dependent on activity start time and estimated parameters $\beta_{HA-H6}$ for home activity start time dummy variables.

Fig. 10 Share of work activities from all consistent activities dependent on activity start time and estimated parameters $\beta_{W3-W6}$ for work activity start time dummy variables.

Land-use

The characteristics of the land-use around the location, where an activity is presumably performed can be used as an additional indicator for the activity type. If the neighbourhood mostly consists out of office or industry buildings, it is highly probably that mostly work activities are performed in this area. In contrary, home activities will be mainly performed in residential areas, illustrated on the map in Fig. 11. Using land-use data for entire island of Singapore, it is interesting to investigate to which degree can its usage improve the activity detection model and compare obtained results from application of the land-use enriched model with simpler model, without any land-use information.

In order to match the land-use information from Singapore Master Plan 2008 with the record of public transport journeys, locations of each activity identified in HITS has to be defined. As the developed model will be applied to the record of public transport journeys, the activity location is defined by the alighting MRT station or the alighting bus stop of the journey preceding the
activity. In case a person arrived to an activity by bus, only bus service number and destination location were reported in HITS. In such cases bus station of the particular service closest to the destination was assigned as the activity location. Furthermore, a kernel density map of estimated total gross floor area for each land-use category in the Master Plan with cell size of 100m x 100m is created using ArcGIS software package. Thereby, the gross floor area is estimated by multiplying area of each master plan zone with the maximum allowed gross plot ratio given in the Master Plan. Spatially matching each station and bus stop with the kernel density values for each land-use category provides a dataset of all stations/stops with corresponding estimated floor areas of each land-use category. Map of all stations/stops overlaid with kernel density map of one land-use category is shown Fig. 12. In this way values of land-use variables for each activity can be looked up in the generated data set. By gradually adding each category to the utility functions of discrete choice model, land use categories significant for home or work activity choice can be identified.

Fig. 11 Locations of three different activity types extracted from HITS 2008 plotted on kernel density map of residential areas in north-east part of Singapore.

Fig. 12 MRT stations and bus stops plotted on kernel density map of zones with mixed commercial & residential use.

From 31 land-use categories present in Master Plan 2008, total of 6 turned out to be significant. In case of utility function of home activity estimated gross floor areas from Commercial, Hotel and Residential were used.

\[
U_{Home-land-use} = \beta_{H7} * Commercial + \beta_{H8} * Hotel + \beta_{H9} * Residential
\]  

(5)

For the utility function of work activity Business Park, Business 2 and Residential variables were included.

\[
U_{Work-land-use} = \beta_{W7} * BusinessPark + \beta_{W8} * Business2 + \beta_{W9} * Residential
\]  

(6)
4.5 Parameter estimation and model comparison

With utility terms for activity duration, activity start time and land-use as defined above, for the utility function for each choice alternative follows:

\[ U_{\text{HOME}} = U_{\text{durationHome}} + U_{\text{startTimeHome}} + (U_{\text{land-useWork}}) \] (6)

\[ U_{\text{WORK}} = U_{\text{durationWork}} + U_{\text{startTimeWork}} + (U_{\text{land-useWork}}) + \beta_{\text{constW}} \] (7)

\[ U_{\text{OTHER}} = \text{const}_O \] (8)

Total of 4 discrete choice models are estimated. Two models are estimated with 100% sample of HITS activities, one with and another without the land-use variables. In order to be able to verify the estimated model, another pair of models, again one with and another without the land-use variables, was estimated using only 80% of the HITS sample. The remaining 20% were then used for model testing and verification.

For the model estimation the software package Biogeme (Bierlaire, 2003) and the CFSQP solver algorithm were used. The estimated values for model with and without land-use variables are presented in Table 2.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
<th>Model without land-use, 100% sample</th>
<th>Model without land-use, 80% sample</th>
<th>Model including land-use, 100% sample</th>
<th>Model including land-use, 80% sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>Robust Std</td>
<td>Robust t-test</td>
<td>Value</td>
</tr>
<tr>
<td>Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>Activity duration coefficient</td>
<td>8.18</td>
<td>0.27</td>
<td>30.31</td>
<td>8.13</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>Activity duration - logistic fcn. parameter 1</td>
<td>-0.384</td>
<td>0.0254</td>
<td>-15.11</td>
<td>-0.416</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>Activity duration - logistic fcn. parameter 2</td>
<td>2.82</td>
<td>0.174</td>
<td>16.23</td>
<td>2.99</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>Coefficient for start time 2am - 12pm</td>
<td>-3.39</td>
<td>0.124</td>
<td>-27.25</td>
<td>-3.54</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>Coefficient for start time 12pm - 16pm</td>
<td>-1.39</td>
<td>0.133</td>
<td>-10.47</td>
<td>-1.48</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>Coefficient for start time 16pm - 2am</td>
<td>0 fixed</td>
<td>0 fixed</td>
<td>0 fixed</td>
<td>0 fixed</td>
</tr>
<tr>
<td>( \beta_7 )</td>
<td>Coefficient for commercial land-use</td>
<td>-6.43</td>
<td>1.03</td>
<td>-6.23</td>
<td>-8.32</td>
</tr>
<tr>
<td>( \beta_8 )</td>
<td>Coefficient for hotel land-use</td>
<td>-28.6</td>
<td>4.46</td>
<td>-6.42</td>
<td>-25.9</td>
</tr>
<tr>
<td>( \beta_9 )</td>
<td>Coefficient for residential land-use</td>
<td>2.14</td>
<td>0.331</td>
<td>6.46</td>
<td>2.52</td>
</tr>
</tbody>
</table>
## Table 2. Estimated model parameters

As already suggested by descriptive analysis, t-test values indicate the activity duration being the most significant parameter in all 4 models. Together with the start time information, a model with reasonably good fit can be estimated without adding land-use parameters to it. However, the chosen 6 kernel density variables of land-use have minor, but still significant influence and are able to slightly increase the Rho-square (see Table 3).

Functions of estimated utility terms for activity duration and activity start time were presented in Fig. 5,6 and Fig. 9,10 respectively, and show reasonably accurate fit with the descriptive analysis of these variables (Fig. 3,4 and Fig. 7,8).

In case of utility function of home activity, commercial and hotel land-use has a negative contribution. For the commercial use this is intuitive, as it indicates the land-use for commercial purposes as offices, shopping, entertainment facilities etc. In case of hotels, the negative sign can be explained by the fact, that in Singapore majority of hotels are large, high-capacity establishments, occupying significant parcel of land and being located in proximity of commercial areas. The residential term has a positive contribution, which is in case of home activities to be expected. For work activities business parks, which indicate a large-scale business or science complexes as well as business 2 areas, mainly indicating manufacture and heavy industries have a positive utility, which is again in line with a common reason. Utility term of residential land-use on contrary has a negative sign, as usually no other land-use is permitted within these zones. Commercial land-use turns out to be not significant for work activities, due to the large number of other activities in the same
areas and therefore is not taken into account within the work activity utility function. Other land-use variables, as for example educational, religious, civic & community institutions or commercial/residential mixed use, were not taken into the discrete choice model at all. This is in most cases due to the very small number of observations of HITS activity within these zones or a strong correlation with other variables, as it is in case of residential and educational land-use, as schools are mainly located close to or even surrounded by residential areas.

In order to compare the two models, with and without land-use information, the likelihood – ratio test is applied. Using number of parameters and final log-likelihood values presented in Table 3, models including land-use parameters pass this test for 100% as well as 80% samples. This provides a good indication that land-use parameters significantly improve the quality of the estimated model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model without land – use, 100% sample</th>
<th>Model without land – use, 80% sample</th>
<th>Model including land – use, 100% sample</th>
<th>Model including land – use, 80% sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of parameters</td>
<td>11</td>
<td>11</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Rho-square</td>
<td>0.771</td>
<td>0.769</td>
<td>0.788</td>
<td>0.789</td>
</tr>
<tr>
<td>Final log-likelihood</td>
<td>-3886.673</td>
<td>-3292.566</td>
<td>-3604.605</td>
<td>-3009.125</td>
</tr>
</tbody>
</table>

Table 3. Model performance indicators

To get a better feel for the quality improvement, simulation of the estimated models is conducted. Thereby the differences in calculated probabilities to make a correct choice for each activity are compared between the two models. Fig. 13 and 14 show histograms of differences in these probabilities, calculated for each choice as follows:

\[ O = P_{\text{choiceWithLandUse}} - P_{\text{choiceWithoutLandUse}} \] (9)

Thereby in slightly more cases the probability of correct choice increases, but the majority of changes lies in the range between -0.1 and 0.1. This applies for simulation of 100% sample as well as application of the model estimated based on 80% sample to the 20% sample.

Table 4 shows the results of the comparison of rule-based approach and the simulation two discrete-choice models estimated with 80% sample and tested of remaining 20% sample. Though there is small improvement in the mean probability of the right choice, the increase is rather small. From this we can conclude that the impact of land-use variables within the model is indeed rather low and the primary activity type is to a great extent characterised by the activity duration and the start time.
Fig. 13 Histogram of probability differences for correct choice for each activity between MNL-Models with and without land-use variables, estimated and applied with 100% HITS sample.

Fig. 14 Histogram of probability differences for correct choice for each activity between MNL-Models with and without land-use variables, estimated with 80% and applied to remaining 20% of HITS.

<table>
<thead>
<tr>
<th>Probability for correct choice</th>
<th>Rule-based model</th>
<th>Model without land-use</th>
<th>Model with land-use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.867*</td>
<td>0.890</td>
<td>0.893</td>
</tr>
<tr>
<td>Median</td>
<td>-</td>
<td>0.969</td>
<td>0.975</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>-</td>
<td>0.212</td>
<td>0.214</td>
</tr>
<tr>
<td>Min.</td>
<td>-</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>Max.</td>
<td>-</td>
<td>0.995</td>
<td>0.999</td>
</tr>
<tr>
<td>Total observations</td>
<td>7936**</td>
<td>2479</td>
<td>2479</td>
</tr>
</tbody>
</table>

* absolute share of correct detected activities between work or other
** all HITS observations of consistent PT activities, without overnight activities

Table 4. Comparison between rule-based and discrete choice models with and without land-use. Discrete choice models estimated based on 80% HITS activity sample and applied to remaining 20% HITS activity sample

5 Model application to public transport smart card record

5.1 Detection of work activities from public transport journeys

In following developed models are applied to the smart card data record of public transport journeys in order to derive a first spatially highly resolved map of work places in Singapore. Thereby the location of the work place is identical with alighting station of the trip preceding the work activity. As soon as certain person performed at least one work activity at the particular location, it is identified as a work place. Therefore, a single person can have multiple work places during the one day. In practice, 4% of workers identified in HITS and 3% of workers identified from public transport journey record had more than one work place.
Corresponding to the activity duration in HITS, duration of activities extracted from smart card data is defined as the time span between two journeys of a consistent journey chain as described in Section 4.2. For overnight activities again the first boarding station and last alighting station of the same day are compared and in case being not more than 1000m away from each other, the time between both is taken as overnight activity duration.

Such distribution of activity durations between journeys recorded in the smart card dataset on a typical weekday, with and without overnight activities is shown in Fig. 15. It features similar pattern as observed in HITS. Again only activities with duration equal to or longer than 1 hour are considered. The slight shift to the right for the peak of longer activities in comparison with HITS (Fig. 3,4) is due to the fact that in HITS participants report the actual beginning and ending of a journey, while only the times of entering or leaving a bus or station are recorded by the smart card. Other parts of the journey, such as walking stages are not captured.

As the discrete choice models were estimated based on the one workday journey records from HITS, they don’t contain any information on the weekly activity dynamics as for instance frequency of the same activity during the week. Therefore, we first evaluate the result from application of the model to the weekly smart card dataset separately for each week day and compare the obtained results among each other. Table 5 shows this comparison with number of home, work and other activities detected for each workday with rule based and discrete choice approach based models. As rule-based model only differentiates between work or non-work activity, only number of detected work activities can be directly compared.

<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rule-based</td>
<td>DCM, no land-use</td>
<td>DCM with land-use</td>
</tr>
<tr>
<td>Home</td>
<td>1'067'905</td>
<td>1'069'899</td>
<td></td>
</tr>
<tr>
<td>Work</td>
<td>652'678</td>
<td>707'976</td>
<td>709'744</td>
</tr>
<tr>
<td>Other</td>
<td>617'580</td>
<td>525'137</td>
<td>521'375</td>
</tr>
<tr>
<td>Total</td>
<td>1'270'258</td>
<td>2'301'018</td>
<td>2'301'018</td>
</tr>
</tbody>
</table>
Table 5. Number of activities of each type for each day of smart card data record, detected with rule-based model as well as discrete choice models estimated based on 100% HITS sample.

<table>
<thead>
<tr>
<th></th>
<th>Thursday</th>
<th></th>
<th>Friday</th>
<th></th>
<th>Average</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rule-based</td>
<td>DCM, no land-use</td>
<td>DCM with land-use</td>
<td>Rule-based</td>
<td>DCM, no land-use</td>
<td>DCM with land-use</td>
</tr>
<tr>
<td>Home</td>
<td>1'082'148</td>
<td>1'083'379</td>
<td></td>
<td>1'079'767</td>
<td>1'080'036</td>
<td></td>
</tr>
<tr>
<td>Work</td>
<td>656'412</td>
<td>716'852</td>
<td>718'774</td>
<td>635'706</td>
<td>708'032</td>
<td>710'747</td>
</tr>
<tr>
<td>Other</td>
<td>650'693</td>
<td>551'023</td>
<td>547'870</td>
<td>776'422</td>
<td>650'534</td>
<td>647'550</td>
</tr>
<tr>
<td>Total</td>
<td>1'307'105</td>
<td>2'350'023</td>
<td>2'350'023</td>
<td>1'412'128</td>
<td>2'438'333</td>
<td>2'438'333</td>
</tr>
</tbody>
</table>

5.2 Mode inflation

With the approach presented in the previous section, only activities from consistent public transport journeys were detected. Work places of people arriving to or leaving from work with other modes of transport, such as car, taxi or privately operated shuttle busses are not included in numbers presented in Table 5. As a next step, mode share data from HITS 2008 is used to inflate the obtained work place numbers.

In order to account for spatial differences in mode choice of work journeys, mode shares are calculated on a zone level. Thereby, 1092 traffic planning zones as used in the transport demand model of LTA serve as a basis. However, as HITS counts only about 13,000 work activities, many of which are accumulated in the central business district, the HITS sample does not provide sufficient observations for the reliable mode share calculation in each of the 1092 zones. Therefore, adjacent traffic planning zones are merged together, using an iterative procedure and a threshold of at least 50 HITS work activity observations per zone. As a result, 162 aggregated zones are obtained, of which almost all feature at least 50 work activities observed in HITS. The few exceptions containing smaller number of observations are zones without any other adjacent zones, as in case of some islands.

For defining a mode of each work activity, journeys to and from work are considered jointly and only one mode is assigned to each work activity. Thereby, only two different modes are differentiated. Public transport mode includes all work travels, which form a consistent public transport journey chain encompassing a work activity, as defined in Section 4.2. All other work travels consisting of journey chains which would be identified as inconsistent or would not be visible at all within the public transport smart card record are considered to be private or other mode of transportation. This should allow to correct by mode inflation for the majority of work activities which cannot be detected within the public transport journey record.

Furthermore, before calculating the mode share, an inflation factor is assigned to the each work trip in HITS, as it was derived by Fourie and Mueller (2011). After all work trips, weighted by the inflation factor, are aggregated within each zone, final mode share is calculated for each of these zones. The result of this detection and inflation process is shown in the Fig. 17. The figure shows, that the work journeys to the central business district in the south-east part of the island are predominantly made by public transport. In contrast, journeys to the western, industrial part are mainly conducted by other means of transport than a consistent public transport journey chain.
5.3 Assessment of result feasibility

The next step would be the comparison of results obtained with the described method with spatial distribution of work locations derived from models currently used by the Urban Redevelopment Authority (URA) of Singapore. Unfortunately at this point in time these models are not accessible for researchers, but will become available in the near future. Instead a rough verification can be made based on the comparison of the total number of detected work locations with the employed population in Singapore reported by the Ministry of Manpower. The total number of work locations in Singapore detected from the public transport journey record using a discrete choice model with land-use information and inflated with mode shares obtained from HITS 2008 is 1.70 million. The employed population of Singapore in 2010 amounted to 3.1 million people (Ministry of Manpower, 2011). Out of these 3.1 million about 0.83 million are work permit holders (Teo, 2012). The majority of work permit holders works in the construction and manufacturing sector or are domestic helpers. These workers are usually brought to work from their dormitories with private buses or trucks or do not need to travel to work at all, as in case of domestic helpers. Hence, it is supposed that these 0.83 million workers cannot be detected based on smart card and HITS travel data. Subtracting these workers from the total employed population, 2.27 million employed people are left. This results in a substantial difference of 0.53 million between employed people according to Ministry of Manpower and workplaces detected from smart card data. There are several reasons for this discrepancy.

First, it is important to note that HITS is a travel survey, not an activity survey. Short journeys, especially walking journeys, tend to be under-reported. As a result, work activities based at or near the home, as is often the case with self-employed business owners, are not recorded and therefore not accounted for in the process of the mode inflation. Fourie and Mueller (2011) used iterative proportional fitting (IPF) techniques to inflate or “gross-up” HITS records such that they match trip flows by time of day, and produce at least the number of households recorded in the 2010 census. As this procedure could only be applied to the motorized trip-making population of Singapore (trip flows for walking trips are not available as a control in grossing-up), it estimates the number of workers, excluding domestic workers,
to be 2.0 million, which is less than the estimation based on official employee numbers and in line with the argumentation above. Comparing this number, with the number of work places obtained from the smart card data record, smart card based detection still results in about 0.3 million less detected work places, which is equal to 15% of all work places. As the absolute number of work places detected based on public transport journeys is highly dependent on the mode share inflation, one reason for underestimation is the increase in car and motorcycle ownership between HITS 2008 and public transport smart card record from 2011, as a continuation of a persistent trend (Choi and Toh, 2010). Hence also the mode share of other than public transportation increased and is not accounted for in the process of inflation. Furthermore, underrepresentation of employment pass (EP) holders in HITS, which in 2011 accounted for 0.17 million work places (Teo, 2012), leads to underestimation of private or alternative transport mode share, as majority of EP holders belong to high income groups and preferably travel by car or taxi. Additional effect arises from the 2.8% not considered public transport trips due to missing alighting station or stop information. Furthermore, as noted before, for work locations on islands like e.g. Jurong Island the number of HITS observation is too low to obtain reliable mode share information.

In summary, it can be said, that the presented method of work place detection leads to reasonable number of workplaces, but is vulnerable to the data on mode share and in case of Singapore results in underestimation of total number of workplaces. Nevertheless, it allows statements about work places locations and its local distribution. Using proportions of number of work places in different areas, it enables accurate assignment on work locations to synthetic population with already fixed number of workers as mentioned above.

5.4 Challenges in assessment of multiday public journey records

As described earlier, HITS 2008 provides only journey records for one workday per household. The lack of any information on continuity and frequency of certain activity types through the week makes application of models estimated based on this data to multiday public journey records a challenging task. In particular important in this context is the treatment of different work or home locations detected for the same card during the analysed time period. Same applies to several activities detected at one location for the same card. One option would be to address this issue by limiting only one activity type per location and person. Thereby one would detect all activities of one person at each location and for each activity calculate the utilities based on discrete choice model, as estimated above. Before calculating the probabilities, one would sum utilities of each activity choice from all identified utilities and then use these aggregated utilities, to calculate the probabilities for each activity type choice. Though this would assure only one single activity per location, still several home or work locations per person would be possible. Another problem arising with assessment of multiday journey records, is the inapplicability of mode inflation from HITS, as described in Section 5.2. Using multiday data, the chance to detect a consistent work activity is higher as in case of one day observation. Therefore, inflation factors obtained from HITS would lead to overestimation of number of work activities.

6 Conclusion

In this paper an approach for usage of public transport smart card fare payment data for characterization of primary activities of people and identification of activities locations by example of the city-state Singapore is presented. It was shown how data from a household
travel survey can be used to supplement and enrich the smart card data recorded by an AFC system and provide information for model estimation. The feasibility of identifying the number of work activities and their locations based on public transport journey record and using inflation factors based on the mode share information from HITS 2008 was demonstrated. Simple rule-based method as well as advanced discrete choice modelling approach delivered reasonable results. Use of land-use information had positive but not very substantial effect on the model fit. Activity duration and activity start time were identified as dominant variables for identification of primary activities.

In order to apply the obtained activity and work place information within an agent-based model, one needs to distribute the detected work places to single buildings around each stop or station. This represents another challenge, as each workplace needs to have certain properties, as type of work or working hours, in order to be match with the agent population. Furthermore, additional effects as for instance private shuttle bus services, which expand the number of buildings being reachable from certain public transport station or stop, need to be taken into account.

Concluding, it can be said that smart card data from public transport offers significant potential for study of travel behaviour and activity identification and can provide important additional data input for development of highly disaggregated, agent-based transport simulations.

References


Fourie, Pieter and Kirill Mueller (2011) “Multi-level weighting of travel survey results”, 16th International Conference of Hong Kong Society for Transportation Studies, Hong Kong, December.


