Transportation Research Part C 49 (2014) 43–58

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Methods for pre-processing smartcard data to improve data quality

Steve Robinson [a,*], Baskaran Narayanan [b], Nelson Toh [b], Francisco Pereira [a]

[a] Singapore-MIT Alliance for Research and Technology, Singapore 138602, Singapore
[b] Public Transport Quality, Land Transport Authority, Singapore 575701, Singapore

ARTICLE INFO

Article history:
Received 22 February 2014
Received in revised form 11 October 2014
Accepted 14 October 2014

Keywords:
Smartcard
Automatic Fare Collection system
Automatic Vehicle Location system
Public transport travel patterns
Data quality
Trip identification

ABSTRACT

In recent years smartcards have been implemented in many transit systems around the world as a means by which passengers pay for travel. In addition to allowing speedier boardings there are many secondary benefits of smartcard systems including better understanding of travel patterns and behaviour of travellers. Such research is dependent on the smartcard correctly recording the boarding stop, and where available the alighting stop. It is also dependent on the smartcard system correctly aggregating individual rides into trips.

This paper identifies causes for why smartcard systems may not correctly record such information. The first contribution of the paper is to propose a set of rules to aggregate individual rides into a single trip. This is critical in the research of activity based modelling as well as for correctly charging the passenger. The second contribution of the paper is to provide an approach to identify erroneous tap-out data, either caused by system problems or by the user. An approach to detecting this phenomenon is provided. The output from this analysis is then used to identify faulty vehicles or data supply using the "comparison against peers approach". This third contribution of the paper identifies where transit agencies and operators should target resources to improve performance of their Automatic Vehicle Location systems. This method could also be used to identify users who appear to be tapping out too early.

The approaches are tested using smartcard data from the Singapore public transport network from one week in April 2011. The results suggest that approximately 7.7% of all smartcard rides recorded the passenger as alighting one stop before the bus stop that they most probably alighted at. A further 0.7% of smartcard rides recorded the passenger as alighting more than one stop before the bus stop that they most probably alighted at. There was no evidence that smartcards overestimated the distance travelled by the passenger.

1. Introduction

In recent years smartcards have been implemented in many transit systems around the world as a means by which passengers pay for travel. The primary benefits of smartcards are speedier boarding and reduced costs by the operator in handling cash. However there are many secondary benefits of smartcard systems including better understanding of travel patterns and behaviour of travellers. Such research is dependent on the smartcard correctly recording the boarding stop, and where possible the alighting stop. Such research is also dependent on the smartcard system correctly aggregating

* Corresponding author.

http://dx.doi.org/10.1016/j.trc.2014.10.006
0968-090X/

individual rides into trips. Much of the research assumes that this is the case. However there are many reasons why such basic information may not be correctly recorded by the smartcard system.

This paper identifies causes for why smartcard systems may not correctly record such information. The first contribution of the paper is to propose a set of rules to aggregate individual rides into a single trip. This is critical in the research of activity based modelling (Bowman and Ben-Akiva, 2001), as well as for correctly charging the passenger. The second contribution of the paper is to provide an approach to identify erroneous tap-out data, either caused by system problems or by the user. An approach to detecting this phenomenon is provided. The output from this analysis is then used to identify faulty vehicles or data supply using the "comparison against peers approach". This third contribution of the paper identifies where transit agencies and operators should target resources to improve performance of their Automatic Vehicle Location systems. This method could also be used to identify users who appear to be tapping out too early. These three areas of research are then brought together in a unified methodology for ensuring the quality of smartcard data. It is recommended that all users of smartcard data should implement this methodology before further analysing travel patterns and travel behaviour using smartcard data. The methods suggested in this paper could also be used by transit agencies to reduce certain revenue loss.

The paper is organised as follows. Section 2 provides a background to smartcard technology and a short literature review into research on the usage of smartcard data. Section 3 then lists various reasons why smartcard systems may output erroneous data. Section 4 describes some of the data that will be used later on in the paper. Section 5 suggests some rules for aggregating rides into trips. Section 6 then proposes a method for identifying faulty tap-out data. Section 7 then outlines a method for identifying the cause of erroneous smartcard data. Section 8 proposes a methodology for pre-processing smartcard data. Finally Section 9 presents conclusions and further work.

1.1. Clarification of terminology used in this paper

In this area of study there is often more than one word for the same meaning. This paper will use the definitions based upon Transmodel as given in reference (European Committee for Standardisation, 2006). The main ones used are defined below:

- **Ride:** This describes the movement of a passenger on a single vehicle, typically a bus or train. The ride begins at the stopping point where the passenger boarded the vehicle, and ends at the stopping point where the passenger alighted the vehicle.
- **Stopping Point:** This describes a location where a passenger may board or alight a public transport vehicle. A stopping point is typically a bus stop or a train platform.
- **Trip (shortened from PT Trip):** This describes the movement of the traveller from an origin location to a destination location. The origin location is assumed to be the first stopping point that the passenger entered the public transport network. Likewise the destination location is assumed to be the last stopping point from which the passenger exited the public transport network.
- **Vehicle Journey:** This describes the movement of a public transport vehicle through a defined sequence of stopping points.
- **Transfer (similar to TransXChange: Connection Link):** This describes the movement of a passenger from one public transport vehicle to another public transport vehicle. This may involve the passenger having to walk to a different stopping point. It should be noted that the term "transfer" has been used to generalise the fact that a passenger may not have to change stopping points, but only vehicles.

It should also be noted that there is inconsistency in the literature as to whether to use "smartcard" or "smart card". This paper uses the former.

2. Background and literature review

Smartcards have been used in the transport industry since the UPass was launched in Seoul in 1996 (Lee, 2010). This was followed by the launch of the Octopus card in Hong Kong in September 1997 (World Bank Group, 1999) and the Oyster card in London in June 2003 (BBC, 2013). This section firstly provides details on how smartcards work and the benefits in reduction of boarding time they provide. The section will then discuss how data from smartcards can be used by transport planners and academics to study OD and travel time estimation. The reader is also recommended to read the comprehensive literature review on smartcard data written by Pelletier et al. (2011). Surprisingly very little literature was found which investigates the quality of smartcard data.

2.1. How smartcards work

Smartcards are cards, typically shaped like a bank card, which have an embedded integrated circuit and are commonly called Proximity Integrated Circuit Cards (PICC) or proximity cards for short. They communicate with a reader. The reader generates an alternating magnetic current at a frequency of 13.56 MHz (Atmel, 2013). This induces an electric current in the antennas of the smartcard which allows two way communication to take place. Communication is undertaken between