Rail-to-Bus and Bus-to-Rail Transfer Time Distributions Estimation Based on Passive Data

Amr M. Wahaballa · Fumitaka Kurauchi · Jan-Dirk Schmöcker · Takenori Iwamoto

Abstract Transit travel time can be decomposed into walking, waiting, transfer and in-vehicle times, and some of these components can reasonably be assumed independent of one another. A practical transfer time distribution that is based on real-world data is difficult to estimate and has rarely been analysed in the literature. This may be due to the difficulty of collecting sufficient data, the dependency of transfer time and other journey time components on walking and waiting time and their variabilities. This paper utilises passive smart card data from Japan to estimate the distributions of the transfer time between bus stops and rail stations. The walking and waiting time distributions are concurrently estimated by the likelihood maximisation of a stochastic frontier model.

Keywords: Transit · Transfer time · Walking time · Waiting time · Stochastic frontier model · Smart card data
1 Introduction

Smart cards, utilised for automatic collection of transit fares, provide a rich and valuable data for many public transportation planning and management aspects. However, smart card data only provide the time of payment transaction which may be done at either boarding or alighting station or both (for tap-in and tap-out transaction system) without providing any information about the time spent on walking, waiting, in-vehicle or transfer among transit lines and/or modes. Specifically, a transfer time distribution from bus to rail and from rail to bus is difficult to estimate and there is accordingly limited literature. The difficulties are collecting sufficient data for all legs of the trip, the inter-dependency of transfer time with other journey time components such as walking and waiting time and their variabilities. In this regard, the objective of this research is estimating transfer time distributions between bus stops and rail stations including possible time fractions (walking time from bus stop to train, walking time from train station to bus stop and waiting time for the coming bus). The estimation of the time distributions of different transit trip components may help transit operators to understand passenger flow patterns and to better model impacts of schedule changes.

Some previous studies suggested splitting transit journey times into separate components assuming independence of these components (e.g., Sun and Xu, 2012, and Zhou et al., 2015). However, these studies ignored or simplified the walking time inside stations and between lines, and their variiances were not taken into consideration except Zhou et al. (2015) who estimated the walking time distribution by a moment estimation method. The approach by Zhou et al. (2015) relied on a field survey to calculate the entry walking time or the exit walking time of the first station. In addition, waiting time was not addressed explicitly in the previous studies, except for a recent study by the authors that estimated the platform waiting time distribution on London underground network based on smart card data (Wahaballa et. al., 2017). They assumed that walking times are normally distributed and validated this assumption by using the actual average and minimum walking time for the studied stations. They approximated walking time and its variability based on the archived data. All previous research lacked actual data for individual passengers’ walking time. Therefore also the variability of walking, waiting, in-vehicle and transfer times are not easily identifiable. In this research, considering the advantage of a smart card data for a medium-sized Japanese city (bus and rail transactions are tap-in and tap-out transaction system), we estimate the distributions of the times spent on the different journey components.

2 Network and Data Characteristics

The rail and bus network of Shizutetsu (Shizuoka Railway Co. Ltd.), located in Shizuoka Prefecture, has an arterial rail line and a bus network serving a wide area between east of Shin-Shimizu and west of Shin-Shizuoka, as shown in Figure 1. The main mode used is bus, however, it is expected that transfer probability exists between
bus lines and/or between rail and buses. The main reason is that there is no arterial or ring bus line that connects travelling from the area east of Shin-Shimizu to the area west of Shin-Shizuoka. Therefore, travellers may transfer from bus to rail to reach the nearest station to their destinations, and take another bus to complete their trip. Studying these complex itineraries is a challenge that requires sufficient and sophisticated data, and rich smart card data called LuLuCa may expedite the estimation of the distributions of different components of transit travel time. These cards collect tap-in and tap-out transactions of bus and rail in addition to shopping payments which facilitate the estimation of the different travel time distributions. The trip can be a pure bus trip, a rail and bus trip and/or including shopping activities.

We utilised panel LuLuCa data collected during the period between Oct. 2012 and Feb. 2017 that provides a reasonable sample size to expedite the estimation of aggregate travel time distributions. The data provides user ID, fare, bus boarding and alighting point and time, rail station entry-gate/exit-gate transactions including time and in addition time and location of shopping payments if any. To estimate the transfer times and their distribution we utilise the Stochastic Frontier Model that will be explained in the following section.

3 Model Formulation

3.1 The stochastic frontier model

The stochastic frontier model (SFM) is a regression model that has been used for estimating the economic efficiency (the production or the cost of a production unit). In SFM, a production (cost) function represents the ideal maximum output attainable (the minimum cost of producing that output) given a set of inputs (Green, 2008). The
The attractive characteristic of the SFM for our purposes is that it has an error term consisting of two different effects; firstly, a two-sided distribution representing the classical noise and secondly a one-sided disturbance error term representing the inefficiency (e.g., improper usage of a machine). Therefore, we suggest the SFM to represent the variability of different travel time components (as a classical noise) by a distribution different from the waiting time (related to transit inefficiency) as detailed below. Cost minimisation function of the SFM is:

\[ y_i = \beta^T x_i + \epsilon_i \]  

(1)

where \[ \epsilon_i = v_i + u_i \]  

(2)

where, \( y_i \) is the cost of the \( i \)th production unit, \( x_i \) is its input and \( \beta \) is a vector of the unknown frontier parameters (fixed for all \( i \)). The error term, \( \epsilon_i \), consists of a noise \( (v_i) \) and inefficiency \( (u_i) \) terms. The SFM parameters can be estimated by the maximum likelihood method assuming that \( v_i \) is a two-sided normal distribution \( N(0; \sigma_v^2) \), \( u_i \) is a nonnegative random term that follows a one-sided distribution that is independent and identically distributed (IID) across observations and which are independent of each other (Jondrow et al., 1982). Previous related studies have assumed a half-normal, exponential or truncated normal distribution for \( u_i \) and its parameters were estimated based on a vector of variables \( (Z) \) as follows:

\[ u_i = Z_i \phi + \omega_i \]  

(3)

S.t. \( \omega_i \geq -Z_i \phi \)

where \( \phi \) is a vector of parameters to be estimated, and \( \omega_i \) is a random variable which comes from the half normal or exponential distribution. The constraint is to satisfy a positive value for the random disturbance related to the inefficiency. Since \( \epsilon_i = v_i + u_i \), its probability density function is the convolution of the two component densities.

3.2 The proposed model

We assumed that a passenger’s walking time (his/her bus-to-rail transfer time) is the difference between the times of bus alighting transaction and rail station entry-gate transaction. While, a passenger’s rail-to-bus transfer time consists of his/her walking time from the exit-gate to the bus stop (assumed equal to his/her bus-to-rail walking time) in addition to the time he/she waited before bus boarding transaction time. We excluded all passengers who performed any shopping transaction during his/her transfer time. We further screened our data to only include passengers who have records for both bus-to-rail and rail-to-bus transfer time.

The relationship between the travel time components is represented by the cost frontier function (shown in equation 1) as follows. The average rail-to-bus transfer time of a passenger (consisting of his/her walking time from rail station exit-gate to bus stop plus his/her waiting time before boarding a bus) is the SFM output \( (y_i) \). The average of the observed walking times for a passenger is considered as the input of
the SFM (\(x_i\)) and the variance of his/her walking times as the classical noise (\(v_i\)). The bus stop waiting time is considered as the inefficiency of the system (\(u_i\)). To simplify SFM parameters estimation, the walking time variability is assumed to be normally distributed and the waiting time is assumed to follow an exponential distribution. These distributional assumptions have been followed in many studies (e.g., Zhou et al., 2015 and Wahaballa et al., 2017). The log-likelihood function for the normal/exponential cost frontier based on the output \(y_i\) can be obtained from the joint probability density function (pdf) of (\(u_i, v_i\)) using the transformation \(\varepsilon_i = y_i - \beta^T x_i\) as in Greene (2008). The maximum likelihood estimation for the stochastic frontier model allows \(\beta, \phi, \sigma_{v}^2\) and \(\sigma_{u}^2\) to be estimated jointly.

\[
\ln L(\beta, \sigma_u) = \sum_{i=1}^{n} \left\{ -\ln \sigma_u + \frac{\sigma_v^2}{2 \sigma_u^2} + \ln \Phi \left( \frac{\varepsilon_i - \frac{\sigma_v^2}{\sigma_u}}{\sigma_v} \right) - \frac{\varepsilon_i}{\sigma_u} \right\}
\]

Where
- \(\beta\) : the unknown parameters to be estimated (fixed for all passengers).
- \(\sigma_u\) : the standard deviation of the bus stop waiting time distribution to be estimated.
- \(\sigma_v^2\) : the variance of the walking time distribution to be estimated.
- \(\Phi()\) : the cumulative distribution function of the standard normal distribution.
- \(\varepsilon_i\) : the error term for observation \(i = y_i - \beta^T x_i\).
- \(N\) : the total number of the studied passengers.

4 Results

The transfer time between Kusanagi bus stop and rail station is modeled using SFM. Maximisation of the log-likelihood function is performed by iterating the numerical procedure of switching between three different methods up to the convergence of the maximisation using STATA software package (2013). A sample size of 68101 bus-to-rail and 68695 rail-to-bus transfers performed by 2277 passengers is used. We only analysed passengers who did both transfers to consider their actual walking times. As shown in Table 1, the P-values show that all parameters are statistically significant with an expected positive sign. We firstly note that the value of the average walking time parameter is close to one, showing low disturbance in the observed average walking time data. The mean waiting time at Kusanagi bus stop is nearly twelve minutes, which appears reasonable given a service headway, ranging from 20 to 60 minutes for different bus lines. The rate parameter of the exponential distribution of waiting time, \(\lambda\), is the inverse of the standard deviation of waiting time shown in the last row of Table 1.
### Table 1 Results of SFM for Kusanagi Station (2277 Observations)

| Parameter                                                      | Coefficient | z-value | P>|z| |
|---------------------------------------------------------------|-------------|---------|-----|
| Average walking time in seconds. ($\beta_{\text{walk time}}$) | 1.173       | 30.73   | 0.000 |
| Standard deviation of walking time seconds. ($\sigma_{x}$)    | 102.732     | 23.61   | 0.000 |
| Standard deviation of waiting time seconds. ($\sigma_{w} = \mu_{\text{wait time}}$) | 711.537     | 245.64  | 0.000 |

### 5 Conclusion

In this paper, we propose a concept to include information from smart card data to obtain distributions of the times required for different stages of a transit journey. We apply our analysis to a station in Japan as a proof of concept. We used actual observations for walking time and its variability to obtain the bus stop waiting time distribution of passengers. To the best of our knowledge, this study is one of the first attempts to do so. Walking time variability is assumed to be normally distributed (following previous related literature) and the waiting time distribution is assumed to follow an exponential distribution. We consequently utilise the normal/exponential SFM. The model parameters are estimated by the maximum likelihood method. The results show low running-cost estimation process and reasonable parameter values are estimated. These findings may help transit operators to make elaborated decisions for optimising transit schedules to maintain service reliability in an efficient manner. We further aim to expand our model to obtain distributions of transfer walking and waiting times for the different user classes. Our sample appears large enough to repeat the above analysis even after clustering passengers according to sociodemographic that are available for some of the smart card holders as well general travel patterns over the data period. We suggest this can give us information on estimation efficiency as well as insights regarding aggregation assumptions made in previous literature.

### References

StataCorp (2013). “Stata Statistical Software: Release 13” College Station, TX: StataCorp LP.
Based on Smart Card Data and Performance Reports. Transportation Research Record: Journal of the Transportation Research Board, (2652), 30-38.