Estimation of train dwell time at short stops based on track occupation event data

Dewei Li [a,1], Winnie Daamen[b], Rob M.P. Goverde [b]
[a] State Key Lab of Rail Traffic Control& Safety, Department of Traffic and Transportation, Beijing Jiaotong University, China
[1] E-mail: lidw@bjtu.edu.cn, Phone: +86 (10) 51684197
[a,b] Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, The Netherlands.

Abstract
Train dwell time is one of the most unpredictable components of railway operations mainly due to the varying volumes of alighting and boarding passengers. For reliable estimations of train running times and route conflicts on main lines it is however necessary to obtain accurate estimations of dwell times at the intermediate stops on the main line, the so-called short stops. This is a big challenge for a more reliable, efficient and robust train operation. Previous research has shown that dwell time is highly dependent on the number of boarding and alighting passengers. However, the latter numbers are usually not available in real time. This paper discusses the possibility of a dwell time estimation model at short stops without passenger demand information, by means of a statistical analysis of track occupation data from the Netherlands. The analysis showed that the dwell times are best estimated for peak and off-peak hour separately. The peak hour dwell times are estimated using a linear regression model of train length, dwell times at previous stops and dwell times of the previous trains. The off-peak hour dwell times are estimated using a non-parametric regression model. There are two major advantages of the proposed estimation model. The model does not need passenger flow data which is usually impossible to know in real time in practice. Also, detailed parameters of rolling stock configuration and platform layout are not required, which eases implementation.

Keywords
Prediction; dwell time; short stops; track occupation; data mining

1 Introduction

Model predictive control has recently been widely used in railway traffic control research, especially in the field of rescheduling (Hansen et al. 2010; Caimi et al. 2012; Quaglietta et al. 2013; Cacchiani et al. 2014; Kecman 2014). Prediction of train dwell times at stations is one of the most important inputs in solving the problem. It provides the predicted trains' trajectories and conflicts to the train dispatchers, and is thus an important input to adjust the timetable in order to resolve the conflicts between train paths. The estimation of dwell time, especially at short stops on main lines may have a big influence on the result of conflict detection. Short stops are stops on the open track where sidings are not available and where trains only dwell for alighting and boarding after which they immediately continue their journey. These dwell times are thus an integrated part of the overall running time over the open tracks between stations. A good estimation of these dwell times is thus

required to be able to predict headway conflicts on the open tracks and arrival times at the main stations at the end of the open tracks.

So far, compared to the running time and dwell times at large stations, dwell times at short stops are not well estimated. Previous researches (Wiggenraad 2001; Daamen et al. 2008; Buchmueller et al. 2008; Yamamura et al. 2013) show that the number of the boarding and passengers is the main determinant of the dwell times especially at small stations which have no passenger connection from one a train to another. However, due to the difficulty to obtain passenger information in real time, most of the existing models, which represent dwell time as a function of the number of boarding and alighting passengers, cannot be used for real time rescheduling when the passenger flow is not available. This is a big challenge for a more reliable, efficient and robust train operation. Li et al. (2014) analyzed the influence of available factors on dwell times other than the number of alighting and boarding passengers. These factors are based on track occupation data of Dutch railways. They found that the dwell times at short stop stations are different from large stations. Moreover, the dwell times at short stops are influenced by different weekday, peak hour, train length in addition to the number of alighting and boarding passengers. This motivates this research: to examine the possibility of building a dwell time prediction model based on predictors without passenger demand.

Based on the assessment of methods of train dwell time estimation by comparing their strengths and weakness, independent variables will be selected that can be used for estimation. This paper gives a more generic and practical dwell time estimation model using the selected variables. The model does not include passenger demand and the detailed parameters of rolling stock configuration, which cannot be obtained in real time.

The remaining paper is organized as follows: Section 2 contains the literature review. Section 3 presents the dwell time prediction model; Section 4 validates the proposed model and describes a case study. We end this paper with conclusions and discussions of further research in section 5.

2 Literature review

There are many factors influencing the dwell times. They can be classified into five categories: passenger, rolling stock, station, operation and external factors (Figure1). Passenger factors include both the amount of passengers and passenger characteristics (gender, luggage, handicap). These factors influence the alighting and boarding time. The influence from rolling stock are threefold: first, different types of rolling stock have a different door control system, which would influence the door unlocking time, door open time and door closing time. Second, The number and width of doors, as well as the horizontal and vertical gap between train and platform determine the capacity of the doors which influence passenger boarding and alighting time. Third, the interior layout of the train (seat arrangements, aisle width, space near the door) would limit the speed of alighting and boarding, thus influence the alighting and boarding time. Station factors include the position of access facilities on the platform and the layout of the yard. The former have an impact on the alighting and boarding time of a door by influencing the distribution of passengers on the platform. The latter influences the dwell time by influence headway between consecutive trains. The railway operation, such as train delay, train overtake or meet, train couple and decouple, time for passenger connection, and operation margin can also bring extra waiting times to a train. External factors include weather conditions, traffic conditions at level crossings near the platform would also have an influence on train dwell times.

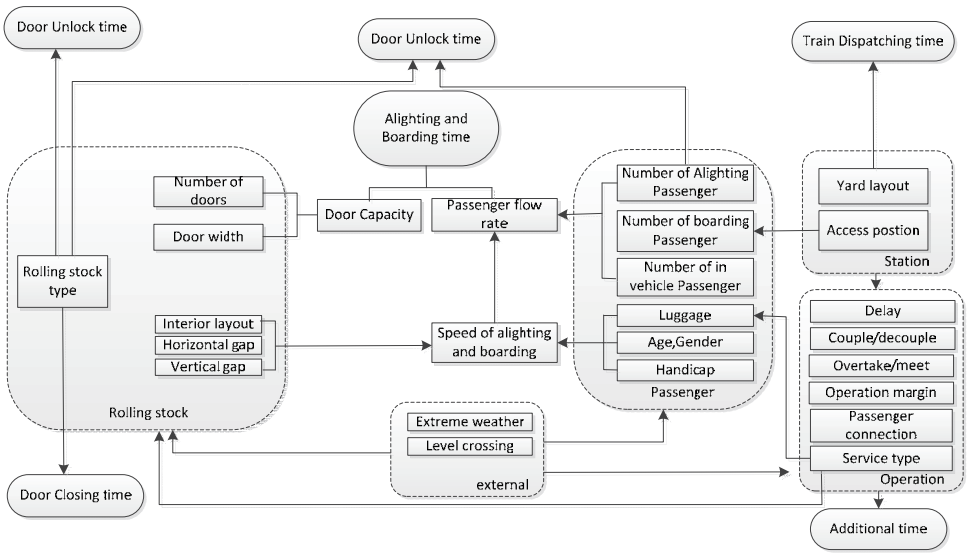


Figure 1: Influencing factors of train dwell time

Based on how different factors are input into the model, a dwell time estimation model may be divided into deterministic and stochastic models. A deterministic estimation model tries to quantify a series of factors, which have an influence on dwell time, and establish the relationships between the dwell time and these factors by a set of fixed parameters. A stochastic prediction model can take into account a certain degree of randomness or uncertainty of the system by estimating the expected probability distribution of the dwell time.

2.1 Deterministic models

In the earlier literature, the dwell time was estimated as the sum of a constant (door opening, closing time and departure preparation time) and the alighting and boarding time (passenger service time). Lam et al. (1998) developed a linear regression model to predict the dwell time as a function of the number of alighting and boarding passengers per train. The assumption of the study is a uniform distribution of boarding passengers on the platform, which may be not true for many stations. According to the investigation from the Dutch railway, there are clear concentrations of waiting and boarding passengers around platform accesses (Wiggenraad 2001). Wirasinghe and Szplett (1984) developed a linear regression dwell time estimation model considering a non-uniform distribution of boarding passengers, calculating the passenger service time of each door respectively, and estimating the dwell time as the maximum passenger service time over all doors. Lin and Wilson (1992), Parkinson and Fisher (1996), and Puong (2000) took the number of standing passengers in the vehicle and their interactions with boarding and alighting passengers into account and developed nonlinear estimation models. The problems of those regression models are that many background variables are not included, such as the composition of the passenger population (e.g. with or without luggage, mobility), configuration of rolling stock, the type of station and so on, which have an irrefutable impact on dwell time (Wiggenraad 2001; Heinz 2003; Daamen et al. 2008). So, these models can hardly be used widely for other trains and stations. In order to estimate the

effect of the configuration of the train on dwell time, Weston (1989) introduced the door width factors of the train into a nonlinear regression model. Weston's model is the most comprehensive model in these deterministic models. It considers the number of alighting and boarding passengers, the interaction between alighting, boarding and standees, and the width of the doors. According to Weston's model, the dwell time will be the same given the door width of the train and the number of passengers. This may not be true because these models neglected some essential factors such as interior layout of train, and horizontal and vertical gaps between the train and platform, which are obviously different from train to train and also depend on the platform where train stops. Harris (2006) tested Weston's model, and found the interior layout of the train should be considered to improve the model accuracy. Jone (2011) estimated the alighting and boarding time at a specific station as a function of the number of alighting and boarding passengers, and different train services that imply the influence of rolling stock. However, the occupancy of the train and the interaction between passengers are not considered.

The aforementioned studies have demonstrated that the dwell time of a train depends on the passengers, rolling stock, station, operation and external factors. However, none of the existing models could fully take all these factors into account. Although these models fits 87% of the data, most of the relative percentile errors are not reported. According to Puong's model, the standard error is 4.04s, the mean value of the dependent variable is 27.76s, which hints an error of 14.55% .

2.2 Stochastic models

It should be noted that many stochastic factors influence the train dwell times. The stochasticity includes the temporal and spatial distribution of passengers, passenger behavior, train driver behavior and train delay. This indicates that a deterministic model could hardly explain the dwell time difference under such uncertain conditions. A stochastic model would be a more appropriate model. There are two types of stochastic dwell time estimation models. One type is a model based on statistical techniques. Buchmueller et al. (2008) proposed a dwell time calculation model for regional trains in Swiss Federal Railways (SBB). The dwell time is estimated as an aggregation of different sub-process times. The distribution of the sub process time depends on vehicle types and the number of boarding and alighting passengers is analyzed based on the sensors' data in the trains. The dwell time is calculated as the sum of these sub-process times. This model is the most generic one. However, the disadvantage of the model is that the occupancy of the train is not considered. It is also not clearly stated how the distribution times of sub processes are aggregated. Besides, it is very expensive to install the detectors to each door of each running train. Hansen et al. (2010) and Kecman and Goverde (2013) found that there is a strong relationship between train dwell time and train arrival delay (or earliness). They estimate the dwell time of a train as a function of its arrival delay which is derived from track occupancy data of the Dutch Railway. It means that no matter how many passengers board and alight, or whatever rolling stock type, the dwell time of a train is determined mainly on whether it is delayed. This model is very applicable for real-time use. However, a later research shows the error of the model on dwell time estimation is even larger than the corresponding scheduled dwell times (Kecman 2014). This may because the linear dependency between dwell time and delay may be true for early arrival trains at big stations where the train should wait until the scheduled departure time. However, there is no evidence whether it is appropriate for shortstop stations where the dwell time is not scheduled explicitly and the train driver locks the doors and departs as

soon as the alighting and boarding process is finished. The other type of stochastic model is the microscopic simulation models. These models focus on passenger alighting and boarding behavior, and estimate the dwell time of the train by repeated simulation of the passenger alighting and boarding process, and record the dwell time as the average passenger alighting and boarding time of each round. Zhang (2008) proposed a microscopic simulation model to estimate the dwell time as a function of alighting and boarding passenger and the width of the door. Yamamura (2013) developed a multi-agent simulation model which also considered the effect of layout of the rolling stock. These models can describe train layout and the behavior of passengers in a very flexible and detailed way. However, these models need to be improved because some factors like horizontal and vertical gap between the train and the platform are not considered. The applicability of these models in real time use is also doubtful due to their time consuming calculation.

A comparison of existing models is shown in table 1. In summary, all these models could hardly be used in real time estimation and prediction, because of the lack of passenger data, the low accuracy or time consuming problems.

Table 1: Main features of existing dwell time estimation models

Source	Model Type	Input variables												
		Passenger				Rolling stock			Station		Operation			External
		Number of A&B	Interaction of A&B	Number of S	Passenger on the platform	Number of doors	Door width	Interior layout	Horizonal &Vertical gap	Heterogeneous stations	Peak& off peak time	Service type	Delay	
Lam,1988	Linear Regression	√	×	×	×	×	×	×	×	×	NA	NA	NA	×
Weston, 1989	Non-linear Regression	√	√	√	×	√	√	×	×	×	NA	NA	NA	×
Lin, 1992	Non linear Regression	√	√	√	×	NA	×	×	×	×	NA	NA	NA	×
Parkinson,1996	Non linear Regression	√	×	√	×	√	×	×	×	×	NA	NA	NA	×
Puong, 2000	Non linear Regression	√	√	√	×	NA	×	×	×	×	NA	NA	NA	×
Buchmueller, 2008	Distribution	√	×	×	×	√	√	√	×	×	NA	NA	NA	×
Hansen, 2010	Linear Regression	×	×	×	×	×	×	×	×	×	×	√	√	×
Kecman &Goverde 2013	Linear Regression	×	×	×	×	×	×	×	×	×	√	√	√	×
Jone 2011	Linear Regression	√	×	×	×	NA	NA	NA	√	×	NA	√	NA	×
Zhang, 2008	Microscopic simulation	√	√	×	×	×	√	×	×	×	NA	NA	NA	×
Yamamura,2013	Microscopic simulation	√	√	√	×	√	√	√	×	×	NA	NA	NA	×

Note: A- Alighting passenger; B-Boarding passenger; S- Standee in the train/vehicle; "√"- included; "×" – excluded; "NA"-not applicable

3 Methodology

The train dwell time estimation for rescheduling can be described as estimating the dwell time of a train (target train) at a station (target station) given real time information related to that train and historical data. In most cases, the number of alighting and boarding passengers, which is the most important independent variable, is unknown in real time. So existing dwell time estimation models, which heavily rely on the actual passenger demand, cannot be used effectively. The main idea of this paper is to find substitute variables which can reflect the passenger demand and to predict the dwell times by using these substitution variables. Most importantly, these variables should be obtained in real time.

The modelling approaches to similar estimation problem include parametric regression model and non-parametric regression model. The former could provide a clear way to show the effect of each predictor on the dependent variable. However, it is difficult to use parametric regression when there are uncleared and complicated non-linear relationships between different variables. To some extent, Non parametric regression model can solve this problem. This paper first select predictors based on the data availability, then it tries to find the relationship between the dwell time and the predictors by applying a parametric model. When the parametric model cannot fit the data, this partial data are estimated by applying a non-parametric model.

3.1 Predictors' selection

Given the dependent variable $\hat{DT}_k^s$, which indicates the dwell time of target train k at target station s, this paper initially selects 10 independent variables as possible predictors, which can be directly get in practical. The main variables and their meanings are shown in table 2.

Table 2: Possible predictors

NO	Variables	Meaning	NO	Variables	Meaning
1	W_k	Weekday or weekend	6	DT_k^{s-2}	Dwell time at second previous station
2	P_k	Peak or off-peak	7	DT_{k-1}^s	Dwell time of preceding train
3	L_k^s	Train length	8	DT_{k-2}^s	Dwell time of preceding train in same train line
4	D_k^{s-1}	Departure delay at previous station	9	L_{k-1}^s	Train length of preceding train
5	DT_k^{s-1}	Dwell time at previous station	10	$DT_{k\text{-hist}}^s$	Dwell time in last week

In table 2, variables refer to weekday or weekend and peak or off-peak 2 and 3 reflect the time variation of the dwell time. The peak period is determined based on the passenger demand of railway in the Dutch railway network as $AT_k \in [6:30, 9:00) \cup [16:00, 18:30)$ at weekdays (NS Group 2013). Statistics (Li et al. 2014) show that dwell times of the peak is significantly different from the off-peak hour (p-value = 0). P_k can be considered as a vector that contains one dummy variable, which indicate the peak and off-peak respectively. Remaining variables are possible predictors which could be derived from track occupation data and timetable data. Target train length is set, because different train lengths require different stop positions, which have great impact on the dwell time.

"Departure delay at previous station" is selected due to the assumption that train delays may increase the number of passengers on the platform, and the dwell time; variables refer to the dwell time at previous station and the second previous station are based on the assumption that there are some relationships of dwell times between consecutive stations. In other words, if the dwell time at one short stop is longer than normal, this may also hold for other trains at other short stops. Variables refer to the dwell time of preceding train and preceding train in same train line are based on the assumption that there are some relationships between the dwell times of consecutive trains. Because the length of preceding train may be different from the target train, variable refers to the train length of previous train is also chosen. Variable refers to dwell time of the same train number in the same day of the last week is a historical variable. We also tested other variables such as headways including headway between the target train and preceding train, head way between the target train and the follower train, and more complicated time series variables including historical average dwell times of the target train, historical average dwell times of the target train at previous station and second previous station, historical average dwell times of the preceding trains at target station. However, these variables have very weak relationships with the dwell times of target train at target station.

Based on the selected variables, the initial model can be described as follows

$$\hat{DT}_k^s = f(W_k, P_k, L_k, D_s^{k-1}, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k-2}^s, D_{k-1}^s, L_{k-1}^s, DT_{k\text{-hist}}^s) \quad (1)$$

3.2 Estimation models

Parametric regression model

A parametric regression method is introduced to build the estimation model. The independent variables are fitted by using a stepwise regression process: First, we started fitting the regression model from a simple linear model with the minimum number of variables (Model 1), and add more variables gradually (Model 2 - Model 6) to see whether a better result can be obtained. The decision about the order in which variable is entered into the model depends on the significance of the relationship between the new variable and the dependent variable as well as the improvement of estimation accuracy by adding the new variable. Some non-linear items are also added to examine whether they can improve the accuracy of the model. The non-linear items include both quadratic items and interactive items (Model 7, Model 8 and Model 9). Due to the earlier finding that the dwell time fits the log-normal distribution (Li 2014), there is an additional model (Model 10), based on the logarithm of the dwell time instead of dwell time.

The significance of an independent variable is different from the synthesis effect of multiple variables. After a model is selected, the significance of each parameter is estimated by using t-test. The corresponding variable with large p-value, which indicate the parameter is not significantly different from zero, is removed from the model. By these steps, a final model is obtained.

Model 1: W_k, DT_k^{s-1}

Model 2: $W_k, DT_k^{s-1}, DT_k^{s-2}$

Model 3: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s$

Model 4: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k\text{-hist}}^s$

Model 5: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k\text{-hist}}^s, DT_{k-2}^s$

Model 6: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k\text{-hist}}^s, DT_{k-2}^s, D_s^{k-1}$

Model 7: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k\text{-hist}}^s, DT_{k-2}^s, D_s^{k-1}, \left(DT_k^{s-1}\right)^2, \left(DT_k^{s-2}\right)^2$

Model 8: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k\text{-hist}}^s, DT_{k-2}^s, D_s^{k-1}, \left(DT_k^{s-1}\right)^2, \left(DT_k^{s-2}\right)^2, DT_k^{s-1} * DT_k^{s-2}$

Model 9: $W_k, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^s, DT_{k\text{-hist}}^s, DT_{k-2}^s, D_s^{k-1}, \left(DT_k^{s-1}\right)^2, \left(DT_k^{s-2}\right)^2, DT_k^{s-1} * DT_k^{s-2}, DT_{k-1}^s * DT_{k-2}^s$

Model 10:

$$\ln(DT_k^s) = f(W_k, \ln(DT_k^{s-1}), \ln(DT_k^{s-2}), \ln(DT_{k-1}^s), \ln(DT_{k\text{-hist}}^s), \ln(DT_{k-2}^s), D_s^{k-1}, \left(\ln(DT_k^{s-1})\right)^2, \left(\ln(DT_k^{s-2})\right)^2, \ln(DT_k^{s-1}) * \ln(DT_k^{s-2}))$$

Non-parametric regression model

A non-parametric regression model is also used to estimate the dwell times, especially on part of the dataset where the parametric model gets low accuracy. The reasons are twofold: firstly, the relationship between dwell time and the independent variables might not be linear. Taking the delay factor as an example, if the delay is small, the effect of delay on dwell times is not significant. However, large delays do have great impact on dwell times due to the accumulation of the boarding passengers. Secondly, the dwell times at shortstops does not fit normal distribution, which is a compulsory condition of linear regression models. In this case, linear regression would be likely to fail. An alternative is to use a non-parametric regression. The basic approach of non-parametric regression is influenced by its roots in pattern recognition (Karlsson 1987).

The non-parametric regression has been widely used in urban traffic estimation and prediction (Davis and Nihan 1991; Smith et al. 2002), where particularly the method of *k*-nearest neighbor (*k*-NN) was applied. This approach will be used in this paper for its fast calculation and relatively good performance accuracy. In the *k*-NN method, it is assumed that the dwell time DT_i depends on a series of variables x_i,i=1,2,3,…,n. Given the measurement of x_i at the moment of prediction, one can find similar cases (called nearest neighbors) from historical data based on the distance between the historical data points $x_{hist,i}$ and the current observation xi. The smaller the distance, the more likely the DT_i equal to $DT_{hist,i.}$ More generally, the forecast of DT_i can be computed as the mean of the dwell times with *k*-th nearest neighbors.

$$DT_i = \frac{1}{k} \sum_{hist-i=1}^{k} DT(x_{hist-i})$$

The core problem is to define the distance function and the choice of *K*. The simplest way to define this distance is to use the absolute sum of differences of independent variables $d = \sum |x_i - x_{hist}|$. Other functions include non-weighed Euclidean distance $d = \sum \sqrt{(x_i - x_{hist})^2}$, and weighted Euclidean distance $d = \sum w_i \sqrt{(x_i - x_{hist})^2}$ to show the importance of each variable. Different values of *K* will be tested to get the minimum estimation error.

The prediction error of *k*-NN method is related to the size of historical data *n* and the neighbor *k*. it can be shown that as $n \to \infty$ and $k \to \infty$ the *k*-NN procedure yields asymptotically minimum risk decisions (Devijver 1982).

3.3 Performance measure

The estimation accuracy is evaluated in terms of the performances of two indicators, the mean absolute percentage error (MAPE) and the root mean square error (RMSE). The MAPE is used to measure the estimation accuracy. The RMSE is also selected to show the actual error when the result is used as inputs of real time rescheduling model, where the total error is calculated based on the combination of the running time and dwell time.

$$\mathrm{MAPE} = \frac{1}{N}\sum\left|\frac{\hat{DT}_k^s - DT_k^s}{DT_k^s}\right| \times 100\%$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N-p}\sum(\hat{DT}_k^s - DT_k^s)^2}$$

Where $\hat{DT}_k^s$ and DT_k^s indicate the predicted and observed dwell times of train k at stop s respectively. N is the total number of trains observed. p indicates the number of degree of freedom.

4 Case study

4.1 Data Collection

The Dutch railway Utrecht – Eindhoven area is selected. Utrecht and Eindhoven are the fourth and fifth largest city in the Netherlands. The railway connected the two cities has a length of 45 kilometers, and contains 13 stations. Utrecht, Eindhoven and Tilburg are the main stations: most trains depart and terminate in these stations. Geldermalsen and Boxtel are basic stations which allow trains to merge, diverge and cross. The remaining stations are shortstop stations. Stations in the corridor are distinguished based on three principles: first, only shortstop stations are selected; second, consecutive short stop stations are selected, so that the relationship of the dwell times between two successive shortstops can be examined; third, stations at which at least 4 trains stop per hour are selected, this ensures as many data as possible for a station. Based on these principles, Houten, Houten Castellum and Culemborg are selected (see Figure 2).

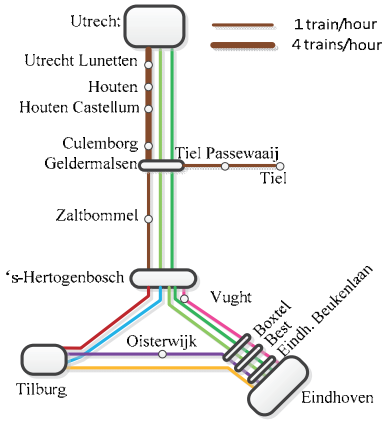


Figure 2: Selected Dutch railway corridor for dwell time estimation

At these selected stations, two train lines S6000 and S16000 have a stop, both have a train interval of 30 minutes. Thus, a train stops at these stations every 15minutes.

The dwell times at selected stops and trains are estimated based on the track occupation data. In the Netherlands, track occupation data are collected using a train describer system (TNV), which provides the exact time of occupation and clearance of track sections. By using a dwell time estimation algorithm (Li 2014) in total 17306 trains running from 1 Sep. 2012 to 30 Nov. 2012 are processed and analyzed.

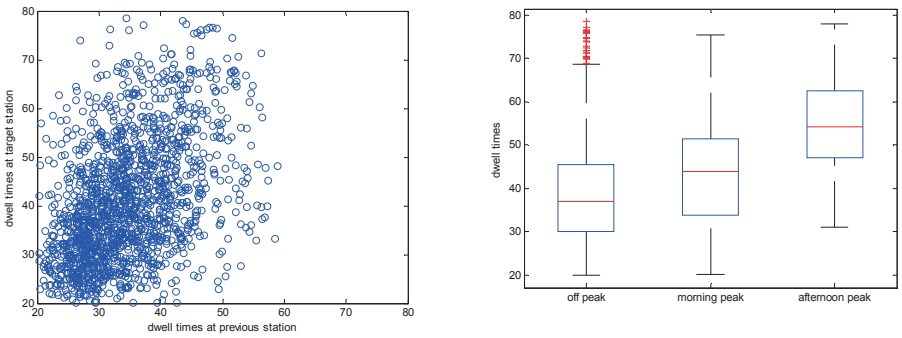
The correlation coefficients for all possible predictors and dependent variables are obtained from the data and shown in table3. It shows that all the predictors are statistically significantly different from zero ($\alpha = 0.001$). The peak hour, length of the train, dwell times at previous station and the second previous trains, dwell times of preceding train has weak linear relationships with dwell time of target train. The best predictor of the dwell time may be the dwell time of the previous station with a correlation coefficient of 0.456. Other relatively high correlation coefficients include: dwell time of the second previous station (0.381), peak time (0.376), dwell time of previous train (0.376), train length (0.308).

Table3: Correlation coefficients of possible predictors and dependent variable(n=1940)

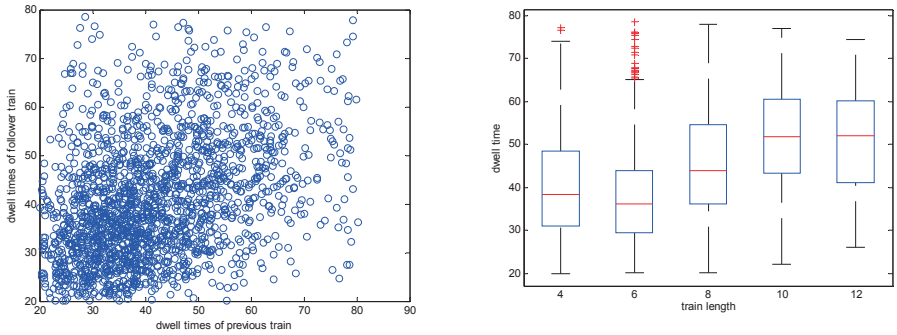
NO	Variables	Correlation	NO	Variables	Correlation
1	W_k	0.178	6	DT_k^{s-2}	0.381
2	P_k	0.376	7	DT_{k-1}^{s}	0.376
3	L_k^s	0.308	8	DT_{k-2}^{s}	0.305
4	D_k^{s-1}	0.224	9	L_{k-1}^{s}	0.101
5	DT_k^{s-1}	0.456	10	$DT_{k\text{-hist}}^{s}$	0.317

*p-value=0.000

The relationships between these four variables and the dependent variable are shown as figure 2



(a) Dwell times between consecutive stations (b) Dwell times at different periods of a day



(c) Dwell times between consecutive trains (d) Dwell times on different train lengths

Figure 3: Relationship between dwell time and the most significant variables

From figure 3, it can be seen that the dwell times are rather scattered, while the dwell time between two consecutive trains ranks the highest; dwell times at off-peak are significantly smaller than the morning peak and afternoon peak; there is a weak linear relationship between dwell times of the preceding train and following train; dwell times of different train length are significant. Longer trains lead to relatively larger dwell times. This is because for longer trains, conductors needs more times to confirm there is no passengers boarding before departure. It is also found that dwell times of train length 4 and 6 has a larger standard deviation than longer trains. This can be explained that shorter trains have a higher probability of deviating from their stop position, and the shorter trains may deviate more from their stop positions than longer trains.

By analysing the relationships (Appendix A) between the selected independent variables, the relationships between dwell times at the previous station and the second previous station, peak hour and train length, last week dwell time, and previous train dwell time are stronger than others. To avoid overfitting, these variables are tested separately to get the best fitting result.

4.2 Parametric Regression results

Models in section 3.2.1 are estimated using linear regression in different week days, peak or off-peak and different lengths of trains as well as a mixed lengths of trains. The results are compared by using the indicators adjusted R^2 and RMSE, which are shown at Appendix B-E. The following summary can be made:

(1) The estimation results at peak hours are better than off-peak hours. It is also found that the R^2 during peak hours are larger than the same model in off peak hours. This is because at off-peak hours, the dwell time variation is larger than peak hours..

(2) At off peak hours, the R^2 of longer trains dwell times are much higher than of shorter trains, which means that correlation between the dwell time of longer trains and with the preceding trains and previous stations are higher than shorter trains. This can be explained by the fact that longer trains have more “rigid” stop positions, so that the distribution of alighting and boarding passengers would not change from train to train.

(3) The delay at peak hour can increase the number of passengers on the platform and cause an increase of dwell time. This effect can be much stronger for shorter trains than for longer trains. This is consistent with the result in table 6. For shorter trains with 4 cars, the R^2 increased significantly, from 0.245 to 0.722 when delay is introduced.

(4) Non-linear items do not improve the result significantly except the dwell time of

eight car, train and ten cars, train during the afternoon peak hour, and ten cars train in the weekend.

In summary, at peak hours, the parametric regression model is better than off-peak hours, during which there are more uncertainties, the parametric regression model should not be used directly. Based on the most powerful model in the above mentioned models, low significance variables are removed using t-test. The final model of dwell time estimation during peak hours is shown as equation 2:

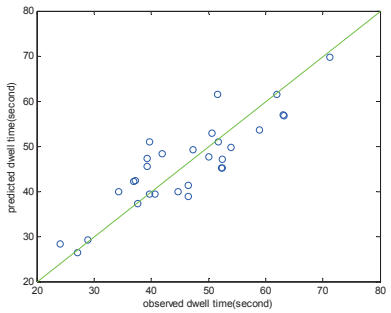
$$DT_k^s = c + \beta_1 L_k^s + \beta_2 L_{k-1}^s + \beta_3 DT_{k-1}^s + \beta_4 \sqrt{DT_k^{s-1} * D_k^{s-2}} \qquad (2)$$

In order to validate the model, the dwell time data are split into two parts based on the train running date with equal sample size. The first part is used for model parameter estimation. The remaining part is used to validate the model. For all trains at peak hours, the regression model is implemented. The estimated parameters and performance under each train length are shown in table4. The comparison between estimation results and observations are shown in figure 4. In case of perfect estimations, the observations would be on the line y = x (shown in green). Data points that are far away from this line represent situations with bad estimation quality.

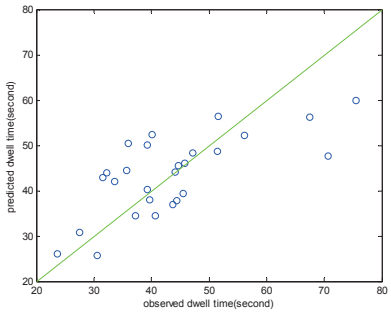
Table 4: Estimation of dwell time models at peak hours

Parameter\ L_k^s	**4**	**6**	**8**	**10**	**12**	**Mix**
Num of cases	33	79	11	56	64	243
constant	-27.46	-13.07	60.47	-3.8	-15.62	-14.50
	(0.16)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)
β_1	-	-	-	-	-	0.60
						(0.00)
β_2	0.47	1.60	-7.9	1.11	1.89	0.77
	(0.00)	(0.00)	(0.04)	(0.09)	(0.01)	(0.01)
β_3	0.00	0.03	0.00	0.00	0.34	0.18
					(0.00)	(0.00)
β_4	1.19	1.11	0.85	1.12	0.83	1.09
	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)
Performance						
Adjust R^2	0.708	0.577	0.742	0.428	0.653	0.574
RMSE(s)	6.96	6.22	6.44	8.69	6.78	7.95
MAPE	12.65%	13.55%	6.53%	12.98%	11.55%	13.9%

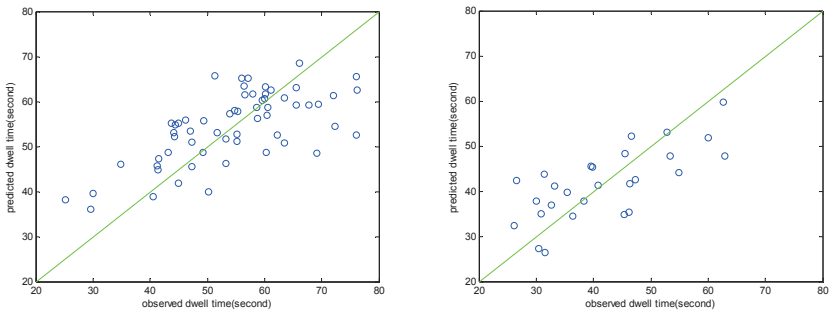
Note: p-values are shown in brackets



(a) L_k=4



(b) L_k=6



(c) L_k=10 (d) L_k=12

Figure 4: Estimation results of dwell time at peak time

From table 4, it can be summarized that the expected dwell times of a train can be represented by the dwell time of preceding train and the dwell time at previous stations. If the dwell times of previous stations and preceding train are large, it is likely the dwell time of the target train at the target station is also large and vice versa.

4.3 Non-parametric regression results

A non parametric regression model is introduced in order to predict dwell times at off-peak hours. Two types of variables are selected. Weekday, peak hour and train length are three variables to get the selection of the historical data. It means when predicting DT_i, the historical data set is chosen based on the same weekday, peak hour properties and same train length. $D_k^{s-1}, DT_k^{s-1}, DT_k^{s-2}, DT_{k-1}^{s}$ are selected to calculate the distance between the historical data and the observations.

In total 1560 records are identified without outliers. The data set is then split into two parts, the first parts contain 900 records is used as learning samples. The second part contains 660 records, which are used to predict. In order to avoid the arbitrary, the distance function is selected by using sum of differences and non weighed Euclidean distance. Because of the limited size of the learning samples, the value of k could only be selected from one to nine. The predicted result is shown as figure 5.

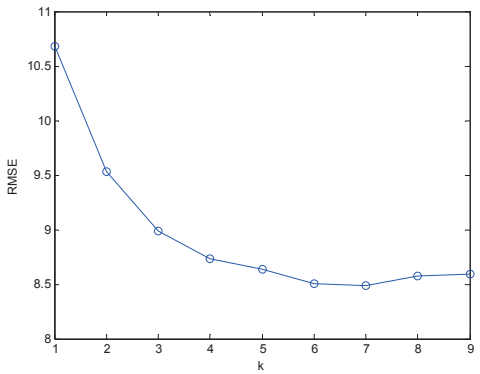


Figure 5: The relationship between k and RMSE in k-nearest neighbor method

From the figure it can be seen that the RMSE get the minimum value of 8.49 when k equals seven, which is higher than the parametric regression model. The MAPE is 19.95% which is within the acceptable estimation error and better than use the simplest 20% percentile value (RMSE=16.6084) which was used in the reference (Hansen 2010).

5 Conclusion

Although a lot of dwell time estimation models exists based on the number of passengers, they can rarely be used in rescheduling practice, because of a lack of real time passenger information. This paper tried to develop a more generalized and more practical estimation model based on train detection data. This paper proposed both parametric regression model and non parametric regression model for real time scheduling. Most importantly, the proposed model does not rely on passenger data, thus it is more practical in real time rescheduling when the number of passengers could hardly be obtained.

The proposed model also shows some potential for development of a more general estimation model despite of different type of rolling stock and stations. We conclude this would be very important for broad applications. The estimation error of dwell times at peak hour is 6.2 -8.8 seconds. The corresponding percentage accuracy is from 85.8% - 88.5%. Since trains are scheduled in minutes, this accuracy is promising. In some cases, especially for short trains in off-peak hours, the accuracy of the proposed estimation model still needs to be improved. Very recently, passenger check in and check out data is becoming available in Dutch network. We believe putting this data into the model can improve the accuracy of the estimation significantly. This work will be done in further research.

Appendix

A: Covariance of independent variables

	W_k	P_k	L_k	D_k^{s-1}	D_{k-1}^{s}	DT_k^{s-1}	DT_k^{s-2}	DT_{k-1}^{s}	L_{k-1}^{s}	DT_{k-2}^{s}	DT_{k-hist}^{s}
W_k	1	0.225	0.175	0.042	0.012	0.065	0.160	0.239	0.268	0.113	0.157
P_k	0.225	1	0.390	0.236	-0.001	0.259	0.264	0.409	0.244	0.211	0.309
L_k	0.175	0.390	1	0.190	-0.016	0.277	0.195	0.228	-0.015	0.176	0.209
D_k^{s-1}	0.042	0.236	0.190	1	0.006	0.239	0.058	0.178	.0528	0.121	0.123
D_{k-1}^{s}	0.012	-0.001	-0.016	0.006	1	0.001	0.000	-0.04	-0.066	0.035	0.035
DT_k^{s-1}	0.065	0.259	0.277	0.239	0.001	1	0.355	0.242	0.055	0.126	0.160
DT_k^{s-2}	0.160	0.264	0.195	0.058	0.000	0.355	1	0.228	0.0364	0.111	0.207
DT_{k-1}^{s}	0.239	0.409	0.228	0.178	-0.04	0.241	0.228	1	0.217	0.251	0.269
L_{k-1}^{s}	0.244	-0.015	0.053	-0.066	-0.066	0.055	0.0364	0.217	1	0.142	0.090
DT_{k-2}^{s}	0.113	0.211	0.176	0.121	-0.035	0.126	0.111	0.251	0.142	1	0.126
DT_{k-hist}^{s}	0.1570	0.309	0.209	0.123	0.045	0.160	0.207	0.269	0.090	0.126	1

B: Model test at morning-peak periods

Model	Train Length									
	4		6		10		12		Mix	
Cases	20		27		16		27		91	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.193	9.212	0.232	10.953	-	12.083	-	11.122	0.080	11.235
2	0.519	7.112	0.452	9.257	-	12.365	-	11.165	0.164	10.708
3	0.619	6.333	0.439	9.359	-	12.755	0.015	10.824	0.231	10.267
4	0.604	6.454	0.452	9.250	-	13.096	0.010	10.855	0.225	10.310
5	0.643	6.124	0.434	9.407	-	13.275	-	11.067	0.220	10.344
6	0.624	6.288	0.426	9.473	-	13.989	0.162	9.988	0.216	10.369
7	0.557	6.826	0.367	9.947	0.0860	11.460	0.253	9.4292	0.246	10.167
8	0.539	6.964	0.367	9.948	-	12.237	0.259	9.390	0.246	10.176
9	-	0.656	-	0.871	-	6.865	-	0.908	-	0.299
10	0.510	7.176	0.413	9.579	0.3595	9.593	0.258	9.393	0.250	10.140

C: Model test at afternoon peak periods

Model	Train Length											
	4		6		8		10		12		Mix	
Cases	13		52		11		40		37		153	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.054	8.786	0.207	7.547	0.300	8.598	0.189	8.538	0.071	8.332	0.244	8.782
2	0.223	7.966	0.309	7.041	0.225	9.050	0.201	8.472	0.317	7.146	0.391	7.878
3	0.163	8.264	0.295	7.114	0.255	8.875	0.255	8.182	0.395	6.723	0.389	7.896
4	0.152	8.322	0.339	6.887	0.193	9.235	0.237	8.281	0.377	6.821	0.392	7.875
5	0.245	7.849	0.325	6.960	0.067	9.928	0.285	8.014	0.358	6.926	0.388	7.900
6	0.722	4.762	0.380	6.669	-	10.823	0.329	7.764	0.345	6.998	0.386	7.913
7	0.611	5.638	0.356	6.802	0.831	4.228	0.292	7.977	0.344	7.001	0.386	7.914
8	0.487	6.470	0.398	6.572	0.663	5.965	0.405	7.310	0.372	6.850	0.392	7.874
9	-	20.978	-	0.225	-	19.574	-	0.964	-	0.359	0.345	0.153
10	0.897	2.899	0.407	6.524	-	100.000	0.400	7.346	0.334	7.056	0.385	7.921

D: Model test at off-peak periods in workday

Model	Train Length											
	4		6		8		10		12		Mix	
Cases	182		499		110		68		9		868	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.113	10.323	0.073	9.425	0.181	10.262	0.118	10.682	0.208	6.994	0.149	10.371
2	0.119	10.289	0.081	9.382	0.186	10.235	0.259	9.793	0.452	5.813	0.172	10.231
3	0.126	10.246	0.090	9.335	0.254	9.793	0.320	9.377	0.529	5.390	0.211	9.985
4	0.121	10.274	0.090	9.340	0.293	9.537	0.359	9.103	0.448	5.835	0.221	9.921
5	0.120	10.284	0.092	9.326	0.300	9.492	0.405	8.775	0.566	5.174	0.234	9.841
6	0.123	10.261	0.096	9.305	0.294	9.532	0.400	8.812	0.907	2.395	0.246	9.762
7	0.120	10.282	0.116	9.203	0.293	9.537	0.417	8.687	-	-	0.256	9.695
8	0.115	10.307	0.115	9.211	0.288	9.570	0.415	8.699	-	-	0.259	9.677
9	-	0.345	0.132	0.240	0.134	0.246	0.006	0.234	-	-	0.259	9.677
10	0.11	10.341	0.114	9.211	0.274	9.664	0.398	8.826	-	-	0.266	9.632

E: Model test in weekend

Model	Train Length							
	4		6		10		Mix	
Cases	81		195		23		299	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.162	11.438	0.107	9.679	0.313	12.228	0.160	10.922
2	0.152	11.509	0.133	9.539	0.404	11.387	0.185	10.756
3	0.203	11.158	0.136	9.521	0.384	11.577	0.197	10.679
4	0.194	11.220	0.133	9.536	0.364	11.761	0.195	10.695
5	0.187	11.271	0.134	9.536	0.374	11.666	0.204	10.632
6	0.176	11.346	0.135	9.526	0.337	12.008	0.205	10.627
7	0.241	10.885	0.135	9.528	0.502	10.411	0.222	10.511
8	0.235	10.932	0.145	9.472	0.508	10.347	0.235	10.426
9	-	0.509	-	-	-	0.871	0.235	10.426
10	0.215	11.069	0.167	9.352	0.433	11.107	0.255	10.289

Acknowledgements

This research is supported by the State Key Laboratory of Rail Traffic Control & Safety（Contract No. RCS2014ZTY1）, China Scholarship Council (201308110079), Beijing Higher Education Young Elite Teacher Project (YETP0555), the Fundamental Research Funds for the Central Universities (2014JBM058). The author would also thank Prof Ingo Hansen for his advices on the paper.

References

Buchmuller, S., Weidmann, U., Nash, A.(2008). "*Development of a dwell time calculation model for timetable planning*". Institute for Transport Planning and Systems, Comprail XI 525. Switzerland.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., & Wagenaar, J. (2014). "*An overview of recovery models and algorithms for real-time railway rescheduling*". Transportation Research Part B: Methodological, 63, 15-37.

Caimi, G., Fuchsberger, M., Laumanns, M., & Lüthi, M. (2012). "*A model predictive control approach for discrete-time rescheduling in complex central railway station areas*". Computers & Operations Research, 39(11), 2578-2593.

Daamen, W., Lee, Y., Wiggenraad P.(2008). "*Boarding and alighting experiments: Overview of setup and performance and some preliminary results*". Transportation Research Record: Journal of the Transportation Research Board, 2042(1): 71-81.

Davis, G.A., Nihan, N.L.(1991). "*Nonparametric regression and short-term freeway traffic forecasting*". Journal of Transportation Engineering, 117 (2): 178-188.

Devijver P. (1982). "*Statistical pattern recognition*". Applications of pattern recognition, K.S. Fu, ed., CRC press, Boca Raton, Fla., 15-36.

Harris, N. G. (2006). "*Train boarding and alighting rates at high passenger loads*". Journal of advanced Transportation 40 (3) : 249-263.

Heinz, W. (2003). "*Passenger Service Times on Trains—Theory, Measurements and Models*". Licentiate thesis. Royal Institute of Technology, Stockholm.

Hansen, I. A., Goverde, R. M. P., van der Meer, D. J. (2010). "*Online Train Delay Recognition and Running Time Prediction*". 13th International IEEE Annual Conference on Intelligent Transportation Systems: 1783–1788.

Jone, J.(2011). "*Investigation and Estimation of Train Dwell Time for Timetable Planning*", Proceedings of 9th World Congress on Railway Research, May 22-26.

Kecman, P., Goverde, R. M. P. (2013). "*An online railway traffic prediction model*". Proceedings of the 5th International Seminar on Railway Operations Modelling and Analyisis, Copenhagen.

Karlsson, M., Yakowitz, S. (1987). "*Rainfall-runoff forecasting methods, old and new*". Stochastic Hydrology and Hydraulics, 1(4): 303-318.

Kecman, P. (2014). "*Model for Predictive Railway Traffic Management*". PhD Dissertation, Department of Traffic and Planning, Delft University of Technology, The Netherlands, 2014

Lam, W. H. K., Cheung, C.Y., Poon, Y. F. (1998). "*A study of train dwelling time at the Hong Kong mass transit railway system*". Journal of Advanced Transportation 32 (3): 285-295.

Li D., Goverde R.M.P., Daamen W., He H.(2014). "*Train Dwell Time Distributions at Short Stop Stations*". Proceedings of 17th International IEEE Conference on Intelligent Transportation Systems. October 8-11, Qingdao, China,

Lin, T., Wilson, N. H. M. (1992). "*Dwell Time Relationships for Light Rail Systems*". Transportation Research Record 1361, TRB, National Research Council, Washington, D.C., pp. 287–295.

NS Group. (2013) *NS Annual Report 2012*. Utrecht.

Parkinson, T. ,Fisher, I. (1996). TCRP Report 13: Rail Transit Capacity. TRB, National Research Council, Washington, D.C..

Puong, A. (2000). "*Dwell time model and analysis for the MBTA red line*". Massachusetts Institute of Technology Research Memo.

Quaglietta, E. , Corman, F. , Goverde, R. M. P. (2013). "*Analysis of a closed-loop control framework in a realistic railway traffic environment*". Proceedings of the 3rd conference on Models and technologies for intelligent transportation systems ,1-10.

Weston, J. G.(1989). "*Train service model – technical guide*". London Underground operational research note 89/18.

Smith, B.L., Williams, B.M., and Oswald R. K.(2002). "*Comparison of parametric and nonparametric models for traffic flow forecasting*". Transportation Research Part C: Emerging Technologies, 10(4): 303-321.

Wiggenraad, P.B.L.(2001). "*Alighting and boarding times of passengers at Dutch railway stations analysis of data collected at 7 stations in October 2000*". TRAIL Research School: Delft University of Technology, Delft.

Wirasinghe, S. C., Szplett, D. (1984). "*An Investigation of Passenger Interchange and Train Scheduling Time at LRT Stations: (ii) Estimation of Standing Time*". Journal of Advanced Transportation, Vol. 18, No. 1, pp. 13–24.

Yamamura, A., Koresawa, M., Inagi, T., Tomii, N. (2013). "*Dwell time analysis in Railway Lines using Multi Agent Simulation*". 13th World Conference on Transportation Research (WCTR) , July 15-18, Rio de Janeiro, Brazil.

Zhang Q., Han B., Li, D. (2008). "*Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations*". Transportation Research Part C: Emerging Technologies, 16 (5): 635-649.