Statistical dwell time model for metro lines

I. Martínez, B. Vitoriano, A. Fernández & A. P. Cucala
Escuela Técnica Superior de Ingeniería ICAI,
Universidad Pontificia Comillas de Madrid, Spain

Abstract

Traffic simulation models in metro lines are widely used in predictive control algorithms for traffic regulation and robustness analysis of timetables. The simulation results are highly dependent on the uncertainty modelling. The two main parameters in the simulation models are running times and dwell times. In lines operated with ATO systems (Automatic Train Operation), running times are more deterministic than random (consequence of control actions), while dwell times show a higher stochastic behaviour due to the influence of passengers and drivers. Typically, simulation models do not include a realistic modelling of dwell time uncertainty, and the confidence on results is affected. This paper is focused on the stochastic component of dwell times in order to obtain a realistic model suitable for traffic simulation of metro lines. For this purpose, several statistical studies have been developed considering peak and off-peak hours, incidences, relations with other operation variables, etc. Models have been obtained and validated using data of different days and Metro de Madrid lines.
Keywords: traffic simulation models, dwell time, stochastic component.

1 Introduction

Simulation models of metro lines are used in predictive control algorithms for the real-time regulation of trains, and in the robustness analysis of train timetables. The two main variables in these models are the running time between stations and the dwell time. In metro lines equipped with ATO (automatic train operation) the running times present, in general, less variability than the dwell times, and they depend mainly on the control commands and technical restrictions. On the other hand, the dwell times are random processes that depend on the number and behaviour of the passengers as well as the action of closing

WIT Transactions on The Built Environment, Vol 96, © 2007 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT070221

door and departure/take-off carried out manually by the drivers. In addition, these dwell times depend on the physical design of the station (width and length of the station, accesses to the platforms, etc.), since the passengers can concentrate in certain points of the station making access to the train difficult [1].

The passengers' arrival to the metro stations is a random process, and the number of passengers waiting in the station for a train increases with the time interval between consecutive trains. If no control action is applied, when a delayed train arrives the number of passengers waiting at the station is increased, thus dwell time could be greater than nominal and the train delay increases also [2]. This accumulated effect makes the system unstable, provoking, in turn, delays in the following trains and disruptions in the functioning of the line. These imbalances have to be corrected by means of control actions.

Typically, the existing simulation models do not take into account the uncertainty in dwell times and thus the confidence in the obtained results is affected. The bibliography on the topic of modelling dwell times on metro lines is very limited, and such models have not been validated with experimental measurement. The most frequently model used for dwell times in predictive control assumes that dwell time increases linearly with the interval between the passing trains, since at a higher interval between trains the number of passengers on the platform increases [2, 3]:

$$s_{k+1}^{i} = S + c_{k+1}\left(t_{k+1}^{i} - t_{k+1}^{i-1}\right) + w2_{k+1}^{i}$$

where s_{k+1}^{i} is the dwell time of the train ith in the station k+1, S is the minimum time the train remains in the station (if there are no passengers), c_{k+1} represents the linear effect the accumulation of passengers has on stoppage during the interval between the departure of two consecutive trains from the same station $\left(t_{k+1}^{i} - t_{k+1}^{i-1}\right)$ and $w2_{k+1}^{i}$ is a disturbance term.

In the previous model, the parameters of model S and c_{k+1} could be adjusted for a set of measurements through a linear regression, although, in the control method proposed in [2], their adjustment to real time will be through adaptive control techniques. In [4] values for c_{k+1} between 0.01and 0.05 are proposed.

Both Breusegem et al. [2], and Silvino and Milani [5] take into account the uncertainty that exists in dwell time, but finally the models developed are deterministic ones.

The main objective of this paper is to analyse the uncertainty in order to obtain a realistic model of dwell times that will improve the simulation models of metro lines.

For this purpose, traffic measures of Metro de Madrid (dwell times, running times, intervals, delays, commands of control, etc.) corresponding to peak hours and off-peak hours have been analysed. Those dwell time measures affected by signalling or by control actions have been filtered and a distribution function has been adjusted to fit the remaining data. As result, a dwell times model and an adjustment method is proposed, which has been validated in different metro lines.

Furthermore, the possible relation between the dwell times and the time intervals between consecutive trains has been analysed.

In section 2 the registered traffic information is described, as well as the filtering process applied to obtain just those measures that depend on the passengers' getting on and off. In addition, a method is described to obtain the distribution function that better fits the dwell times, and results are validated using suitable statistical tools. In section 3 the possible relation between the dwell times and the time interval is analysed. This relation is typically modelled as linear by other authors and this paper tries to establish the existence or not of this relation according to the measures obtained from Metro de Madrid. In section 4 the analysis is focused on the occurrence of severe delays (incidences) during the stops, for example due to breakdowns in doors. The aim is to identify and to model the dwell times that are abnormally long, and these will be modelled independently as incidences. Section 5 describes the proposed model for the dwell times of the trains in the stations, and in section 6 the model is validated. Finally, in section 7 the conclusions of the research are presented.

2 Model of dwell times without incidences

Real data on train circulation times for the Circular Line (L6) of Metro de Madrid have been used to carry out this study. The Circular Line has 27 stations traversed counter-clockwise. Two timetable modalities were studied, peak time from 8h 8min to 9h 18min (21 trains) and off-peak time from 10h 42min to 13h 38min (11 trains), both during workdays.

The measurements of dwell times are first filtered, discarding those for trains that arrive in the station with the exit signal in red, as only station dwell times corresponding to trains affected solely by the number of passengers or incidences (not by signals) were of interest.

For statistical analysis of the dwell times in each station in a homogeneous way it is necessary to consider times relative to the minimum dwell time in a station, that is, the minimum dwell time is subtracted to all the measurements T_e. In this way, a set of dwell time data T_L is obtained for each station, and histograms and Box and Whisker Plot diagrams are plotted for these data collections.

Abnormally long dwell times occur due, in some cases, to breakdowns. These dwell times must be identified and filtered as they do not represent a normal stop in a station to load and unload passengers. These longer than normal dwell times will be called incidences and will be modelled separately later.

The first hypothesis to be tested is that the dwell times follow a normal distribution. To check the hypothesis, goodness of fit tests are stated where the null hypothesis is that the dwell times follow a normal distribution. The analyses compute the p-values in each station, which is used to measure the goodness of fitness to the distribution proposed as well as to reject the null hypothesis if the p-value is equal or less than the level of significance adopted (in this case 0.10).

After performing the hypothesis tests with the dwell times in peak and off-peak hours, following the previous methodology p-values obtained allow discard the possible normality of these data.

From the Box and Whisker Plot of fig. 1, a log-normal distribution is proposed. This model assumes that natural logarithms of the dwell times follow a normal distribution. To the set of data Ln[T_L] of a station a Box and Whisker Plot analysis was carried out to identify and eliminate the outlier and extreme outlier times. On these filtered-out times hypothesis test as the previous ones were performed.

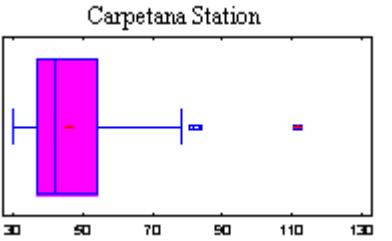

Figure 1: Box and Whisker Plot.

Table 1: p-values off-peak hour and peak hour to log-normal distribution.

	p-values	p-values	Stations	Off-peak	Peak
Stations	**Off-peak**	**Peak**	Nuevos Ministerios	0.21	0.91
Laguna	0.58	0.99	Cuatro Caminos	0.55	0.88
Carpetana	0.15	0.30	Guzmán el Bueno	0.95	0.81
Oporto	0.04	0.11	Metropolitano	0.86	0.29
Opañel	0.12	0.19	Ciudad Universitaria	0.42	0.90
Plaza Elíptica	0.48	0.70	Moncloa	0.86	0.93
Usera	0.10	0.46	Argüelles	0.39	0.13
Legazpi	0.13	0.98	Príncipe Pío	0.43	0.52
O´Donnell	0.26	0.94	Puerta del Ángel	0.74	0.74
Manuel Becerra	0.45	0.38	Alto de Extremadura	0.72	0.62
Diego de León	0.37	0.74	Lucero	0.78	0.98
Avenida de América	0.39	0.91			
República Argentina	0.83	0.89			
	p-values	**p-values**			

p-values obtained allow accept that the data can be modelled as a log-normal distribution in peak hours with a reasonable level of confidence, see table 1. However, for off-peak hours there is a case, Oporto, whose p-value is under the bound 0.1. Because, the hypothesis can not be rejected for the rest of cases and, moreover, it can be accepted with a high level of confidence, the hypothesis also will be accepted for Oporto, as a possible case of Type I error. It is known that a 0.1 significance level assumes that it is possible with a probability of 0.1to obtain as a result of a hypothesis test to reject the null hypothesis being a mistake. So, the log-normal distribution is accepted for every station in both peak and off-peak hours.

It is advisable that the dwell time data have a single log-normal distribution, not only in their form, but also in their parameters. For this reason a variance analysis has been carried out to compare the stations data and to see if this factor is relevant. This study checks if all the dwell times can be considered as arising from a single log-normal distribution; in such a case, all the data could be studied together. Otherwise, each station would have to be studied and parameterized separately. Through Analysis of Variance (ANOVA) it is concluded that the dwell times of all the stations can not be considered as coming from a single distribution, which involves that the dwell times have to be parameterised for each station separately (both for peak hour and off-peak hour). This conclusion confirms that each station has its specific characteristics of length, passengers' affluence, etc.

Technical results of the performed analyses are not shown in order to clarify and emphasise methodologies and results.

3 Dwell time interval between trains relationship

The aim is to find the relationship between dwell time and the interval of time between two consecutive trains in a single station. First a linear type relationship was sought:

$$T_{Lk} = S_k + c_k\left(t_k^i - t_k^{i-1}\right) + \varepsilon_k$$

T_{Lk} is the dwell time of a train in a station *k*-th. S_k is the minimum dwell time in station *k*-th. c_k is the coefficient of linear regression and $\left(t_k^i - t_k^{i-1}\right)$ the time between the departure of two consecutive trains from station *k*-th. ε_k is a random variable of station *k*-th associated to other non-controlled and randomness factors.

To measure the relationship the coefficient of determination. coefficient R^2 associated to linear regression will be used. This coefficient measures the percentage of total variability explained by the factor. It is observed that the values of R^2 are very low, see table 2, both for off-peak time as well as peak time, concluding that the proposed linear relationship does not occur.

From the following logarithmic regression:

$$LnT_{Lk} = c_k Ln\left(t_k^i - t_k^{i-1}\right) + \varepsilon_k$$

low values of R^2 are also obtained. In both cases the values of R^2 obtained are close to 0. So the relation is considered not useful to forecast the dwell time behaviour known the interval between trains.

Table 2: Maximum values of R^2 for linear and logarithmic regression.

	Linear regression	Logarithmic regression
R^2 peak	13.5 %	15.23%
R^2 off-peak	14.25%	16.60%

4 Model of incidences

Incidences will be considered those dwell times when the train remains in the station for longer than a specific period of time and which are not due to the normal process of passengers getting on and off trains. The total number of incidences in the line are low, but should be considered separately for each station, because these incidences are obtained with different parameters of the log-normal distributions.

The Box and Whisker Plot of fig. 1, reveals an extremely long tail to the right with more probability accumulated than that one shown by a log-normal distribution. Moreover, from the analysis of other cases, it seems to indicate the mixture of two populations.

As the distribution assumed for the dwell times is the log-normal, it seems logical to consider that an abnormal dwell time, or incidence, is the time that appears as an outlier in the natural logarithm. Therefore, the study of incidences will be performed with outlier dwell times data (data greater than third quartile plus 1.5 times the interquartile range).

Because the absence of enough data, the study of the incidences will be carried out in all the stations being considered as a single set. To do it, data were typified with the mean and standard deviation of the station, in such a way that they are non-dimensional and can be treated all together. Therefore, the incidence data will be outliers of variables Z_e. with distribution $N(0,1)$, through the change $Z_e = \frac{X - \mu_e}{\sigma_e}$. where μ_e is the mean and σ_e the standard deviation of the dwell times of station e.

The percentage of incidences is computed as:

$$\%Incidences = \frac{\sum_{stations} \frac{Incidences}{n}}{N} \times 100$$

where $\sum_{stations} \frac{Incidences}{n}$ represents the number of dwell times that are incidences belonging to a specific station among the total number of dwell times from that station. N is the total number of stations, in this case 23 stations.

The histogram of the typified incidences for peak hour, fig. 2, shows a longer tail at the right, because they would be censored data at the left side (a minimum value is imposed in order to select data considered incidences). Also, more probability is accumulated at the left side that can be explained as an overlapping area of the right tail of the usual dwell times distribution and the left tail of the incidences distribution.

The starting point for the analysis is the hypothesis that the logarithms of the incidences follow a normal distribution and normality tests are used to check the distribution. The results of these tests show that the lowest p-value is 0.45 and given the p-values are all greater than 0.10, it can not be discarded that the normalized incidences for peak time come form a normal distribution, with a level of confidence of at least 90%.

The off-peak hour incidences were handled in an analogous way. The normality tests were applied to these incidences and as the p-values obtained were greater than 0.10, it could not be discarded that normalized incidences for off-peak time come form a normal distribution, with a level of confidence of at least 90%.

These results suggest that the natural logarithms of the incidences of peak and off-peak hour can be modelled according to a normal distribution; and thus the original times follow a log-normal distribution, which was the model proposed initially.

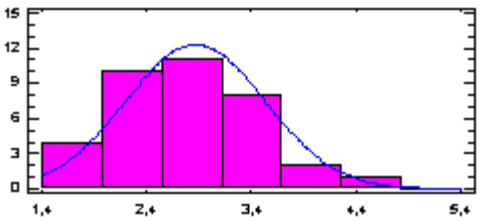

Figure 2: Histogram of typified incidences at peak hour.

5 A complete model of dwell times

A model for train dwell times in each station, based on previous studies, is proposed as a mixture of two distributions: a log-normal distribution for the dwell time without incidences and another log-normal distribution for the incidences:

$$\text{Dwell time} = T_{mín} + \beta_1 X + \beta_2 Y$$

where:

$T_{mín}$ ≡ the minimum dwell time of the train in each station.

β_1 ≡ parameter defining the percentage of dwell time without incidences. This parameter is different for peak hours and off-peak hours and is equal for all stations. It is estimated through the percentage of incidences minus the probability given by the fitted log-normal distribution of incidences in the censored area (0.34%).

$\beta_2 \equiv$ parameter defining the percentage of dwell times that are incidences, $1-\beta_1$, is different for peak and off-peak time and is equal for all stations.

$X \equiv X \overset{d}{=} LN(\mu_e, \sigma_e)$: random variable modelling the dwell times in each station that are not incidences. μ_e is the mean of the dwell times logarithms, different for each station and does not include incidences. σ_e is the standard deviation of the dwell times for each station, also different for each station and not including incidences.

$Y \equiv Y \overset{d}{=} e^{\mu_{e_1} + \sigma_{e_1} T} \quad T \overset{d}{=} N(\upsilon, \rho) \rightarrow Y \overset{d}{=} LN(\mu_{e_1} + \sigma_{e_1}\upsilon, \sigma_{e_1}\rho)$: random variable modelling the train dwell times for each station which are clearly incidences (outlier dwell times). μ_{e_1} is the mean of the dwell times logarithms in each station, including outliers; in the same way σ_{e_1} is the standard deviation. υ is the mean of the outlier times typified (aggregated for all the stations because the absence of enough data to carry out the study separately) and ρ is the standard deviation of these data; so, υ and ρ are constants varying only between peak and off-peak hours.

6 Validating the model

In this section, results of the validation process of the proposed model are presented. To achieve it, a comparison between observed data and simulated data from the model is proposed. The observed data are the original dwell times in each station, filtered but without subtracting the minimum dwell times and without applying logarithms or separating out the outliers. Simulated data are obtained through the following simulation process:

a) Simulation of u values uniformly distributed in (0,1)
b) If $u \leq \beta_1$ (estimated probability of non-incidence) $\Rightarrow$ X is generated (model for non-incidence dwell time), otherwise, Y is generated (model for incidences).

Homogeneity tests will be carried out to compare both samples. To perform these tests, previously randomness tests were performed over the observed data in order to discard temporal relations. Results confirmed this time independence.

The homogeneity tests were carried out and results allow us to conclude that both samples could be considered from a same population. During peak hours, p-values vary between 0.2 and 0.9; in non-peak hours results are not so clear, they vary between 0.12 and 0.89, but also could be considered into the limit to do not reject the hypothesis.

To validate the model, the model has been applied to another line of Metro de Madrid. Line 1. During peak hours, 38 trains circulate along Line 1 and during non-peak hours, 24.

Firstly, incidences on Line 1 were studied. They were obtained in an analogous way to those on the Circular Line. Differentiating between peak and off-peak hours. The percentage of incidences in the peak hours on Line 1 is

0.67%. The normality tests applied to the incidences logarithms in peak hours show that incidences have a log-normal distribution with a confidence level of at least 90%.

The percentage of incidences in the off-peak hours on Line 1 is 0.32%. Given this low percentage, it could be considered that during off-peak hours dwell times of trains in stations can be modelled with a log-normal single distribution (tail probability for this distribution). So, in this case, the value of the parameter of probability of incidence would be 0, which is not contradicting the model proposed.

Parameters of the distributions proposed for non-incidences and for incidences in the peak hour are estimated from the data as in the previous case.

Following the methodology applied for the Circle Line, the model assuming the mixture of two log-normal distributions is validated comparing the observed data with simulated data of the model through a homogeneity test. Based on the results obtained, above 0.10, the model cannot be rejected, and in some cases, it is clearly accepted.

Finally, the model obtained from this first collection of data of Line 1 is checked with real dwell times for this line, but corresponding to different days of study. The result obtained from the comparison indicates that the model reflects real data in a satisfactory way as p-values are determined to be over 0.11. From the results obtained, the proposed model for dwell times in stations is considered valid.

7 Conclusions

In this paper a statistical model of the dwell times in metro lines has been proposed. It is a realistic model of the dwell times due to the passengers getting on and off, distinguishing peak and off-peak hours.

Traffic measures from Metro de Madrid have been filtered and analysed. It has been observed that the minimal dwell time is different for every station, and also that the average dwell time in peak hours is greater than the average in off-peak hours. This confirms the assumption that the dwell time increases with the number of passengers.

Next, the statistical distribution of the dwell times was analysed, concluding that the dwell times can be modelled according to a mixture of two lognormal distributions (one of them to model severest delays called incidences and another one for non-incidences) both in peak and off peak hours.

As result of the ANOVA analysis applied to the registered measures in Metro de Madrid. It has been concluded that the model is valid for every station, but distribution parameters must be estimated separately for each station due to the specific characteristics of each one.

In this paper, traffic measures have been analysed to establish a direct relation between the dwell times and the time intervals between consecutive trains. However, according to the statistical models obtained for Metro de Madrid lines, such a relation could not be verified, which contradicts the basic hypotheses of the existing models proposed by other authors.

Also, this work has focused on identifying and isolating the severest delays (incidences). Incidences are the dwell times identified as outliers in the statistical analysis. Incidences in peak and off-peak hours have to be normalised and studied as a whole because the absence of data to perform analyses for each station.

Finally, the proposed model of dwell times is a mixture of two log-normal distributions. The first one models the typical dwell times and the second one the dwell times corresponding to severe delays. A methodology to estimate the distributions parameters from real traffic measures has been proposed. This methodology has been applied to the Circular Line and to Line 1. The fitted models are validated comparing simulated times obtained from the proposed models and the registered measures, obtaining good values for validation.

Furthermore, in Line 1 the model (with parameters, not only the methodology) was compared with real data from other days, obtaining satisfactory results.

According to the results obtained it can be concluded that the model and the methodology proposed to estimate the distribution parameters of the dwell times can be accepted.

References

[1] Welding. P.I. & Day. D.J. Simulation of underground railway operation. *The railway Gazette*. pp. 438-441. June 4. 1965.

[2] Breusegem. V. V. Campion. G. and Bastin. G. *Traffic modelling and state feedback control for metro lines.* IEEE Trans. Automatic Control. pp. 770-784. 1991.

[3] Fernández. A..Cucala. A.P.. Vitoriano. B.. Cuadra de. F. *Predictive traffic regulation for metro loop lines based on quadratic programming*. Proc. ImechE Vol.220 Part F: J. Rail and Rapid Transit. 2006.

[4] Campion. G. Breusegem. V.V.. Pinson. P. Bastin. G. *Traffic regulation of an underground railway transportation system by state feedback*. Optimal control applications & methods. Vol 6. pp. 385-402. 1985.

[5] Silvino C. and Milani B. *Regulaçao Robusta de Tráfego em Linhas de Metro*. Revista Controle & Automaçao. Vol. 12 no.2. pp. 118-130. 2001.