Public Transp (2015) 7:295–319
DOI 10.1007/s12469-015-0106-7

ORIGINAL PAPER

Predictive modelling of running and dwell times in railway traffic

Pavle Kecman[1] · **Rob M. P. Goverde**[2]

Published online: 4 June 2015

Abstract Accurate estimation of running and dwell times is important for all levels of planning and control of railway traffic. The availability of historical track occupation data with a high degree of granularity inspired a data-driven approach for estimating these process times. In this paper we present and compare the accuracy of several approaches to model running and dwell times in railway traffic. Three global predictive model approaches are presented based on advanced statistical learning techniques: LTS robust linear regression, regression trees and random forests. Also local models are presented for a particular train line, station or block section, based on LTS robust linear regression with some refinements. The models are validated and compared using a test set independent from the training set. The applicability of the proposed data-driven approach for real-time applications is proved by the accuracy of the obtained estimates and the low computation times. Overall, the local models perform best both in accuracy and computation time.

Keywords Predictive modelling · Robust regression · Running and dwell times · Tree-based models

✉ Pavle Kecman
pavle.kecman@liu.se

Rob M. P. Goverde
r.m.p.goverde@tudelft.nl

[1] Department of Science and Technology, Linköping University, Linköping, Sweden

[2] Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands

1 Introduction

Accurate estimation of running and dwell times is important for all levels of planning and control of railway traffic: strategic, tactical and operational. The validity of capacity analysis and line planning on the strategic level depends to a great extent on the accuracy of process time estimation (Abril et al. 2008; Schöbel and Schwarze 2013). Similarly, on the tactical level, the accurate estimation of running and dwell times is important for creating feasible and realizable timetables (Goverde 2007; Büker and Seybold 2012; Medeossi et al. 2011). Finally, on the operational control level, process times need to be estimated for real-time traffic prediction and conflict detection (Dolder et al. 2009; D'Ariano et al. 2007; Kecman and Goverde 2014), as well as to provide reliable passenger information (Berger et al. 2011).

The essential drawback of the existing approaches to estimating running times (Brünger and Dahlhaus 2014; Wende 2003) and dwell times (Stam-Van den Berg and Weeda 2007; Buchmueller et al. 2008) in real time is that they do not consider actual traffic conditions on the network. In other words, process time estimates do not differ depending on the time of the day, train line or delays. The initial work in overcoming this problem was presented by Van der Meer et al. (2010) using a data-driven approach based on current monitoring data and historical data. However, the macroscopic character of that model prevents estimation of train runs on the level of block sections, which is essential for prediction of route conflicts.

This paper presents new data-driven approaches for deriving running and dwell times. A global approach consists of a generic statistical model, applied on an aggregated set of historical data. Data about all running times on the level of block sections and all dwell times are aggregated and used to train and validate statistical models. A set of predictor variables is identified for the purpose of building the global model. Three advanced supervised learning methods are used for computing accurate process time estimates. Furthermore, local models are developed that estimate process times for particular blocks, stations and train lines. Each local model represents an independent statistical model. The global and local approaches are compared by their accuracy and applicability for calibrating a real-time railway traffic prediction model described by Kecman and Goverde (2014).

The following section covers the relevant contributions in the literature. The methodology used to build the statistical learning models is described in Sect. 3. Section 4 describes briefly the advanced supervised learning methods for estimation of conflict-free running and dwell times followed by their implementation in the global (Sect. 5) and local model (Sect. 6). Model validation on a test set is presented in Sect. 7. Finally, the main conclusions and recommendations for further research are given in Sect. 8.

2 Literature review

2.1 Running time estimation

Train running times are usually computed by means of train motion equations (Brünger and Dahlhaus 2014; Wende 2003). This approach considers the characteristics of rolling-stock and infrastructure that are represented in parameters of train dynamics models. The empirical parameters for running time computation are typically given for a particular line or rolling-stock type and train composition. The parameters are usually determined and tuned by experts for a particular railway company. This method is often used in the process of timetable construction (Hooghiemstra 1996) and microscopic traffic simulation (Nash and Huerlimann 2004).

However, greater precision can be achieved by calibrating the parameters of the train motion equations for different traffic conditions using realised running time data derived from track occupation or train event recorders data. Longo et al. (2012) define a single parameter for each dynamic motion phase. Bešinović et al. (2013) extend this approach by calibrating each tractive effort and resistance parameter separately. These parameters are optimised by a procedure that minimises the gap between the simulated and actual train positions and speed profiles. The computational requirements for solving the train motion equation prevent the application of this method for running time estimation of a large number of trains in busy networks. Moreover, such approach is static in the sense that it does not consider the current traffic state and potential impact on running times.

These drawbacks can be overcome by data-driven approaches that compute robust estimates of process times based on historical data. The basic idea is to determine the explanatory variables and build statistical models that quantify their impact on process times using a training data set. Given the values of explanatory variables, the models provide the estimates of process times in real time. Van der Meer et al. (2010) presented an approach based on robust regression analysis to investigate the correlation between process times and delays. The results showed a strong correlation between arrival delays and dwell times, whereas the correlation between running times and departure delay was found to be much weaker. Similar results were obtained from a set of track occupation data from Switzerland by Lüthi (2009). Both approaches for modelling running times rely on data aggregated over open track sections. As a result, neither train separation principles on open track nor route setting and release principles in station areas can be included.

2.2 Dwell time estimation

Current approaches for the estimation of dwell times used in timetable construction and rescheduling rely on the measurements of realised dwell times. Wiggenraad (2001) performed a detailed analysis of dwell times, and passenger boarding and alighting processes using a set of manually collected data from seven busy stations in the Netherlands. The impact of platform and vehicle characteristics, delays,

station types and peak-hours was analysed with the purpose of a detailed analysis of dwell times. The analysis determined the average boarding and alighting time per individual passenger as well as per passenger within a cluster. Lee et al. (2007) performed a similar study using manually collected data from two busy stations in the Netherlands. They focused on the factors that determine passenger behaviour and their influence on dwell times. The limited scope of the studies due to the manual data collection makes it hard to derive general conclusions.

The availability of data from on-board event recorders inspired a stream of research on dwell time modelling. More precise and larger data sources were analysed with the purpose of deriving general conclusions about dwelling processes of trains in stations. Buchmueller et al. (2008) collected and analysed data from door sensors, passenger counters and train event recorders. They analysed the duration of each subprocess separately with respect to vehicle and platform design, and passenger demand. The impact of uncertainty of dwell times was studied by Longo and Medeossi (2013) who presented a complex model for dwell time estimation that separates dwell time into deterministic and stochastic subprocesses. They focused on the detailed modelling of stochastic processes such as boarding and alighting time, waiting time and departure imprecision time. The application of a detailed data-based approach that relies on sensor and train event recorders data strongly depends on data availability which in particular restricts online use.

For real-time prediction of traffic evolution it is more convenient to use track occupation data, which is received in real time in traffic control centres from all running trains. The main challenge then is to estimate the exact arrival and departure times on the platform tracks. Stam-Van den Berg and Weeda (2007) presented an approach that relies on knowing the exact location of access points to the platform, the actual stopping point of each train and assuming the constant acceleration (deceleration) of all trains. Even though this approach reduces the estimation error based solely on track occupation data it relies on the knowledge of platform design and track layout. In this paper we used a generic procedure to estimate exact arrival departure time that relies not only on section occupation times but also on section release times (Kecman and Goverde 2012). This requires only track occupation data as input.

3 Methodological framework for statistical analysis

3.1 Main methodology

The global and local approaches in this paper for estimating running and dwell times bridge the gaps identified in the literature review of Sect. 2. Both approaches rely on advanced statistical learning methods that are able to quickly produce accurate estimates of process times. The estimates are derived based on historical track occupation data thus overcoming the limitations of manual data collection for dwell time estimation. Moreover, the models are created without the need of relying on a detailed description of rolling-stock and platform design nor passenger data which are difficult to obtain.

A set of relevant predictor variables is identified for each process type. The principle idea is to derive process time estimates depending on the values of explanatory variables that reflect the current traffic conditions in the network, i.e., the actual train positions, delays and period of the day. Moreover, running times are estimated on the level of block sections. The estimates can be used for calibrating microscopic railway traffic models that can capture route conflicts. Finally, the resulting predictive models are applicable in a real-time environment. The estimates can be produced quickly depending on the traffic condition parameters obtained from the live stream of track occupation data.

3.2 Description of the data set

For this study, track occupation data archives for three months (March–May, 2010) from the areas Rotterdam and The Hague in the Netherlands were made available by the Dutch infrastructure manager ProRail. The experimental setup was built for the busy corridor Leiden–The Hague–Rotterdam–Dordrecht in the Netherlands. The 60 km long corridor is (partially) traversed daily by approximately 300 trains per direction. Figure 1 shows the schematic representation of the observed network along with the train lines and corresponding stopping pattern for the 2010 timetable. The thin line illustrates the train line that runs once per hour, whereas the other lines operate twice per hour.

The raw data archives were processed to extract and filter the process times of all trains (Kecman and Goverde 2012). The resulting data format, which classifies conflict-free running and dwell times per block section, station and train line, is an essential requirement for applying the models described in this paper to different networks and case studies. Data archives from 82 days were used as a *training set* to

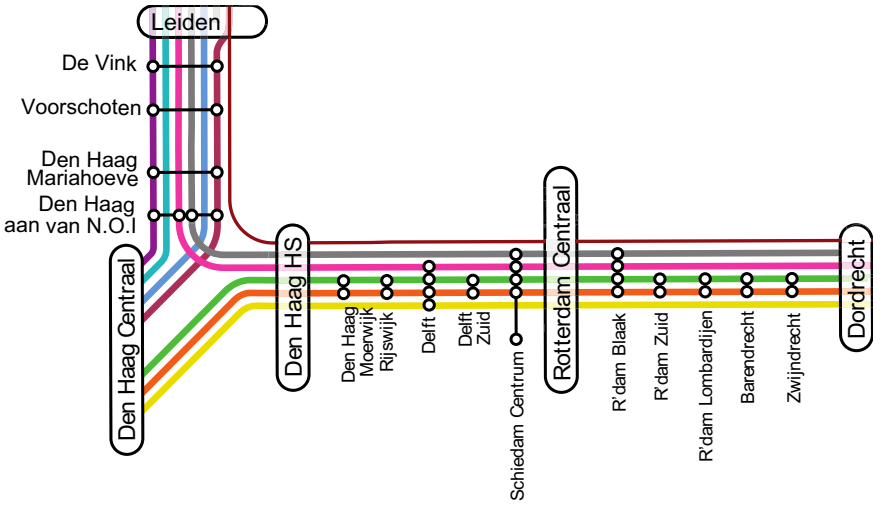

Fig. 1 Corridor and train lines for the case study

train and calibrate the statistical models. The archives of the remaining 10 days were used in a *test set* to test and compare the prediction accuracy of the models.

The dwell times at each station were computed using the occupation and release times of the platform sections, which may result in an estimation error (Stam-Van den Berg and Weeda 2007). For arrival times the error equals the time lag between the last section message before the train stop and the actual standstill. Similarly, for departure events, the estimation error is equal to the time lag between the actual departure and the first section message.

3.3 Global model

The global model for process time estimation aggregates the process times of all trains that were recovered from the raw track occupation data archives into a set of running times and a set of dwell times. A separate model is created for each process type.

3.3.1 Global model for running time estimation

Predictor variables used to estimate train running time over a block were determined. Table 1 summarises the training set used to build the model. The response variable 'running_time' represents the running time of a train over a block. The training data set contained 101,481 data points describing the running times of nine train lines over 143 blocks in 82 days. An obvious indicator of train running time is the block length ('block_length'), which can be derived from track occupation data. Moreover, we consider the distance of the block from the last scheduled stop ('distance_from'), as well as the distance to the next scheduled stop ('distance_to') in order to include the effects of extended running times due to braking and acceleration. The distances were computed between the middle of the platform and the middle of the block.

Furthermore, we consider the impact of peak-hours on train running times. A binary variable is created that indicates whether the observed process takes place during a peak-hour (27 % of data relate to peak-hours). The difficulty in separating the data set to peak and off-peak events stems from the fact that the limits of peak hours can be fuzzy, as well as train line and station dependent. The exact limits of peak periods are difficult to obtain without the additional data sets that reflect the

Table 1 Summary of the training set for running time estimation

	Mean	Median	St. dev.	Min	Max
running_time (s)	43.16	41.57	19.10	10.01	179.35
departure_delay (s)	100.29	53.97	148.90	−147.73	1199.13
block_length (m)	1137.82	1185.00	384.70	255.00	1915.00
distance_from (m)	5300.93	3685.00	5291.24	131.00	23,765.00
distance_to (m)	6986.98	5140.00	5500.12	1190.00	24,440.00
headway (s)	691.58	610.58	556.48	93.66	21,349.03

demand such as passenger counts, ticket sales information or smart card data. Therefore, in the global model, we use the definition of peak-hours on work days from the Dutch national train operator. Morning peak is the period between 6.30–9.00 and the afternoon peak is between 16.00–18.30. Note that the drawback in accuracy of using the predefined limits for peak periods is overcome in the local model (Sect. 6.2).

A categorical variable that indicates the train type is also considered as a predictor. In order to create a generic model, this variable has only two levels: intercity trains with scheduled stops in large stations and local trains that stop in every station along their route (68 % of data points are related to intercity trains). Freight trains are not included in the data set due to a small corresponding sample in the considered area. Moreover, even though hindered train runs are excluded from the data set, the headway time between successive trains is included as a predictor ('headway') that may explain the impact of the preceding train on train running time. Headway time is in this context defined as the time between successive occupations of the same block.

Finally, we test the validity of the assumption that the running time of a train depends on the delay at the previous departure ('departure_delay'). It is assumed that delayed trains may run with maximum performance in order to reduce the delay by using the running time supplements included in the timetable. On the other hand, trains running on time or ahead of their schedule are assumed to run at a lower speed, thus avoiding early arrivals and achieving energy efficient driving. This assumption was not validated in earlier approaches (Lüthi 2009; Van der Meer et al. 2010) on the macroscopic level.

3.3.2 Global model for dwell time estimation

This section presents dwell time predictors obtainable from track occupation data. Continuous variables for dwell time estimation are shown in Table 2. The training set contains 145,807 points describing the dwell times of 9 train lines in 19 stations (5 of which are large) over 82 days. The scheduled dwell time ('scheduled_time') for each scheduled stop is an obvious choice for predictor variable. Furthermore, the fact that trains do not depart before their scheduled departure time indicates that arrival delay may have a major impact on train dwell time. Early trains have longer dwell times than scheduled in order to avoid early departures. On the other hand, trains with a positive arrival delay that is larger than the dwell buffer time spend a minimum dwell time in order to minimize the departure delay.

Table 2 Summary of the training set for dwell time estimation

	Mean	Median	St. dev.	Min	Max
dwell_time (s)	136.60	114.72	80.17	10.01	599.66
arrival_delay (s)	8.48	−16.60	131.05	−299.94	1199.04
scheduled_time (s)	70.20	69.60	65.40	30.00	360.00

The impact of train and station type is examined by including the corresponding two-level categorical variables. Stations are separated into small (60 % of data points) and large stations, and trains into intercity (47 % of data points) and local trains. A station where only local trains stop (it is skipped by intercity trains) is considered small. On the other hand, a station where both local and intercity trains stop is considered large. Finally, the impact of peak-hours on dwell times due to the increased number of alighting and boarding passengers is included in the same manner as in the model for running time estimation (24 % of data points from peak-hours).

3.4 Local model

The processed track occupation data enable to derive a separate process time estimation model for each block, station and train line. The goal of the local model is to explain the variation of running times of trains from the same line over a particular block. For that reason, many of the predictors used for running time estimation in a global model, such as block length, distance to and from the last scheduled stop, train type and headway become redundant. In order to estimate the process times for a particular instance (train line, block section or station), we investigate the impact of departure delay and attempt to verify the assumption that delayed trains run faster to reduce their delay. Similarly, the local model for dwell time estimation derived for each station and train line considers only the impact of arrival delay and peak-hours.

The applicability of local models is limited by data availability. For example a local model for estimating the running time of trains from a certain train line over a particular block cannot be generalised to other block sections or train lines. Therefore, a sufficient amount of data is required to build each local model. It is important to have this in mind because train lines operate with different frequencies and some parts of the network may be utilised less than the busy main lines or station routes.

4 Statistical learning methods

This section introduces briefly the statistical learning methods used for developing the process time estimation models. The criteria used to select these methods are: prediction accuracy, the simplicity of implementation, computational requirements and the interpretability of results. Moreover, an important aspect is the trade-off between bias and variance, i.e, between underfitting and overfitting the models (Hastie et al. 2009). Finally, due to the envisaged real-time application of the obtained process time estimates, it is essential that they are robust against outliers, noisy and missing values in the data. We first apply linear models due to their simplicity. Second, the accuracy of predictions might be improved by regression tree based method, which can capture nonlinear relations between the predictors and the response. Finally, random forests are applied to overcome the high variance of regression trees.

4.1 Robust linear regression

Robust linear regression (Rousseeuw 2005) represents a modification to the ordinary least-squares linear regression method with the intention to identify and exclude outliers in the data set. Estimates that are resistant to outliers can be obtained by fitting the regression curve to the majority of data and subsequently identifying outliers as data points with large residuals from the robust solution. Rousseeuw and Driessen (2006) presented an efficient algorithm for computing robust linear regression coefficients using the least-trimmed squares (LTS) method. The objective is to find a subset of h from n data points and minimise the ordered squared residuals for each observation y_i and the corresponding estimate $\hat{y}_i$

$$\sum_{i=1}^{h}(y_i - \hat{y}_i)^2_{i:n} \tag{1}$$

where $(y_1 - \hat{y}_1)^2_{1:n} \leq \cdots \leq (y_n - \hat{y}_n)^2_{n:n}$ are ordered squared residuals and h is a point that reflects the percentage α of resisted outliers, $h = \lceil n(1-\alpha) \rceil$.

The simplicity of the linear model comes with a price of inaccuracy and inability to model interactions between predictors and their non-linear impact on the output variable. An example of the non-linear relationship between a predictor and the output variable are discrete categorical variables. Interactions between predictors indicate that they are correlated and the impact of one predictor is dependent on the value of another. It is therefore difficult to distinguish the impact of correlated predictors separately. An example of interacting predictors may be that delayed trains run faster in the acceleration and braking phase. Thus the impact of block position with respect to previous and next scheduled stop could indicate how important a departure delay is on running time estimation. The correlation between distance from the previous and distance to the next scheduled stop is clear.

4.2 Tree-based non-linear methods

4.2.1 Regression trees

A way to overcome the drawbacks of the linear method emerged with the development of tree-based methods (Breiman et al. 1984). The basic concept of these methods and their application in regression is to segment the predictor space into simple regions. The output variable is predicted in each region separately. The predictor space represents the set of values for the predictors $X_1, X_2, \ldots, X_p$ which is divided into J distinct and non-overlapping regions $R_1, R_2, \ldots, R_J$. For every observation of p predictors an appropriate region R_j (terminal node in the tree) exists. The terminal node for each observation is reached by applying splitting rules, i.e., binary decisions at internal nodes that direct the observation towards its corresponding region. The estimated value of the output variable is computed as the mean of the response values for the training observations in R_j.

Regression trees are obtained from an optimisation procedure that minimises the residual sum of squares $\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$, where $\hat{y}_{R_j}$ is the mean of all output values from training observations in R_j. Because of the high computational complexity of this problem, Therneau et al. (2014) developed an algorithm that recursively partitions the predictor space in a greedy manner. The tree is built by the following procedure: first a single variable X_j and point s is found that splits the data into two groups $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$. The procedure is recursively repeated in each partition until a threshold is reached in terms of the number of training observations in the region. The algorithm always selects the splitting variable that contributes the most to minimising the RSS. Certain variables with a low contribution to the main objective may thus be left out completely from the model and not be chosen as splitting variables. An indicator of importance is obtained for each variable used for growing the tree. It is computed as the sum of improvements of the objective function for each split for which the variable was used as the splitting variable. Furthermore, in order to increase the interpretability of a tree and avoid overfitting the model to the training set, the tree can be pruned. The resulting tree has fewer regions and performs better on the test set. More details on pruning can be found in Breiman et al. (1984).

The major advantage of regression trees is that they are transparent and easy to interpret and validate by experts. They are able to handle non-linear dependencies, interaction between predictors, and categorical variables. However, the prediction accuracy is often unsatisfactory when applied on a test set. Even after pruning the trees, the prediction in terminal nodes may be significantly affected by outliers.

4.2.2 Random forests

The drawbacks of regression trees can be overcome by generating a large number of trees on the training set and using the average values of all responses to estimate an instance from the test set. This concept is called bagging and relies on the repeated sampling of the training set and obtaining B different training sets (Breiman 1996). Each sample S_b, $b \in \{1, \ldots, B\}$, of $\lceil 2B/3 \rceil$ randomly drawn points is used to build a regression tree. The predicted response of the model to a test observation is computed as the average over all trees

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_{S_b}. \tag{2}$$

In each sample, the data that is left out is used to estimate the so-called 'out-of-bag' (OOB) error. The response for the bth observation is predicted using each regression tree for which this observation was left out from the training set ($B/3$ observations on average). All errors are averaged to obtain the OOB error as a cross-validated indicator of model accuracy.

Random forests have been introduced by Breiman (2001) in order to further improve the accuracy of bagging models. They rely on randomisation of the previously described recursive algorithm for construction of each tree. The major

modification is that not all predictors are considered for choosing the best split of the predictor space but only m randomly chosen variables. The prediction is again performed by averaging the response of each of the B trees thus further reducing the response error of regression trees.

5 Process time estimates by global models

This section presents the results of applying the statistical learning methods described in the previous section on the available data set (Sect. 3.2). The algorithms for creating the statistical models have been implemented in the R programming language for statistical computing (R Core Team 2013). The packages for robust linear regression (Rousseeuw et al. 2014), regression trees (Therneau et al. 2014) and random forests (Liaw and Wiener 2002) were used to build the models. All models were fitted on a computer equipped with an Intel Core i5-520M/2.4 GHz processor and 4 GB memory. The computation times are highest for the random forest models which take approximately 5 min to build. The other models are computationally less demanding and produce solutions in several seconds.

5.1 Running time estimates derived from the global model

5.1.1 Robust linear model for running time estimation

The results of applying LTS robust multiple linear regression to fit the data are given in Table 3. The coefficient is given for each variable. Moreover, we give an indicator of importance of a particular variable for the overall model, which is a representation of the corresponding p value. Note that the categorical variables are represented by only one level since the value for the second level is equal to zero.

Table 3 Summary of the LTS model for running time prediction

	Dependent variable: running_time	
	Coefficient	p value
peak_hour = 1	−0.1608	***
departure_delay	−0.0019	***
headway	−0.0013	***
distance_to	−0.0002	***
distance_from	−0.0002	***
block_length	0.0239	***
train_type = 'local'	0.6803	***
Intercept	13.0600	***
R^2	0.6514	
Residual std. error	6.5810	
F statistic	19,320.0800	***

*** $p < 0.01$

The results indicate that all considered variables have a significant impact on running time. The running times during peak-hours are slightly shorter than in the off-peak. A negative correlation with departure delay is determined, which indicates that delayed trains run slightly faster to recover from the delay. Furthermore, the negative correlation between running time and headway may indicate that in case of a short headway with the preceding train, trains tend to run slower to reduce the possibility of running into a route conflict. The position of the block with respect to the station of the previous and next scheduled stop reflects the train motion regime, i.e., acceleration, cruising, coasting or braking. The negative coefficients of the corresponding variables indicate shorter running times with increased distance from/to the scheduled stop. Furthermore, block length has an expected positive impact on train running times. Finally, local trains are estimated to have slightly longer running times than intercity trains.

The lower part of Table 3 presents the predictive quality of the model. The R^2 value indicates that 65 % of the variation of running times can be explained by the presented model. Having in mind that the variation within the training set is relatively low (Table 1), this implies that the presented model is useful for estimating running times. This is also demonstrated by the low residual square error (RSE) of less than 7 s. The large F statistic and low p value indicate a strong correlation between response and explanatory variables.

5.1.2 Regression tree model for running time estimation

The non-linear relationship between predictors and response, as well as interactions among predictors can be resolved using regression trees. Figure 2 presents the tree obtained after applying the recursive partitioning algorithm (Therneau et al. 2014) on the training set.

A complex regression tree with 16 internal nodes (splits), indicated by ovals, and 17 terminal nodes (rectangles) was generated. Each node contains the mean value of the response (running time) and the number of data points n in the corresponding region. The tree indicates the relative importance of the variables: 'length' (41 %), 'distance_from' (21 %), 'distance_to' (18 %), 'train_type' (11 %), 'headway' (5 %) and 'delay' (4 %). The data is split throughout the tree in accordance with the interpretation of the results from the linear regression for important variables. However, the final terminal nodes that are split according to the 'length' variable show inconsistencies with the assumption that running time is positively correlated with block length. In case of short headways ($<$214.4 s) and blocks close to the scheduled stop ($<$1970 m) the running time over short blocks tends to be longer. However, the large mean squared error (MSE) obtained after tenfold cross-validation of the tree indicates that these regions may be affected by outliers and therefore produce inaccurate estimates.

The overall quality of the model is presented in Fig. 3 which shows the improvement of error (left) and R^2 (right) after performing each split. The prediction error is presented relative to the initial estimation error which equals the mean of all observed running times in the training set. Each split contributes to a reduction of

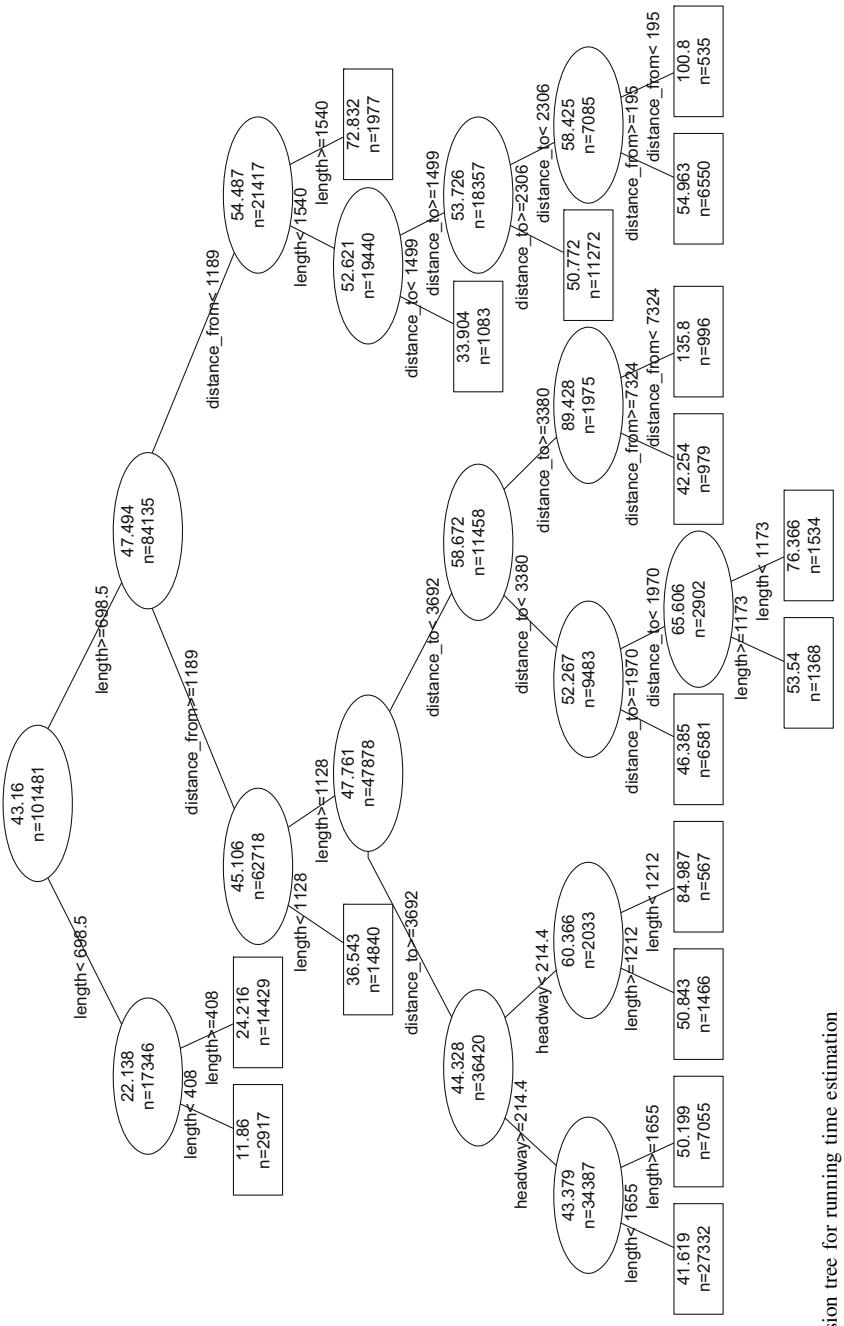

Fig. 2 Regression tree for running time estimation

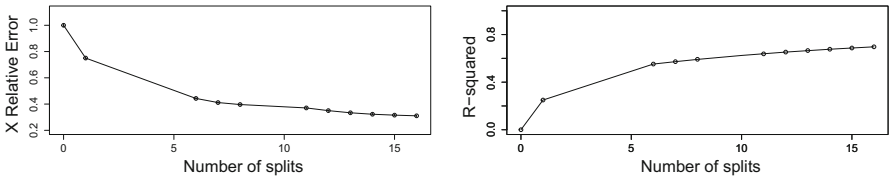

Fig. 3 Relative prediction error (*left*) and R^2 (*right*) of the regression tree running time model

the prediction error. The maximum value of $R^2 = 0.697$ is obtained for 16 splits. Even though the predictive quality of the linear model is improved by using a regression tree, the effects of outliers still may present a source of inaccuracy for estimating running times.

5.1.3 Random forest model for running time estimation

Random forests provide a further increase of prediction accuracy and improve the resistance of regression trees against outliers. The training set (Table 1) is used to create a random forest model with 300 trees (Liaw and Wiener 2002). Each split is performed using the best of three randomly chosen variables from the full set of predictors.

The MSE and R^2 depending on the number of trees in the forest are presented in Fig. 4. A significant error decrease of 30 % is visible for increasing the forest size up to 100 trees. Further increase of the number of trees has a limited contribution to error reduction. A significant improvement of R^2 is achieved compared to the approach with a single regression tree. The effects of increasing the number of trees above 100 are small. The value of $R^2 = 0.780$ indicates that 78 % of running time variation can be explained by the predictor values. The relative variable importance is obtained after computing the OOB error and does not differ from the single tree case.

5.2 Dwell time estimates derived from the global model

5.2.1 Robust linear model for dwell time estimation

LTS robust multiple linear regression is used to fit the data from the training set. The results show that all predictors have a strong impact on the response 'dwell_time' (Table 4). The relatively large intercept can be explained by the unavoidable error of dwell time estimation (Sect. 3.2). The realised dwell times are clearly closely

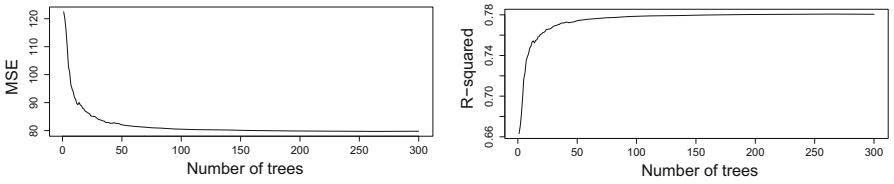

Fig. 4 MSE (*left*) and R^2 (*right*) of the random forest running time model

correlated with the scheduled dwell times. Dwell times in small stations, as well as the dwell times of local trains are estimated to be slightly larger than scheduled. Moreover, arrival delay is negatively correlated with dwell times. That finding reflects the fact that early trains have to wait for the scheduled departure time and late trains tend to depart as soon as possible to reduce their delay. Finally, dwell times during peak-hours are estimated to be longer than in off-peak periods.

The indicators of the predictive quality of the model (lower part of Table 4) show a high predictive power of the model with 73 % of response variation explained by the explanatory variables. The high value of the F statistic also indicates the relevance of selected predictors on the response value. However, the possible interactions between explanatory variables cannot be determined using the linear model which is why non-linear predictive models are also tested.

5.2.2 Regression tree model for dwell time estimation

The recursive partitioning algorithm (Therneau et al. 2014) is used to optimise regions in the prediction data space with respect to prediction error. The resulting regression tree containing eight splits and nine terminal nodes is presented in Fig. 5. The relative variable importance is also determined: 'scheduled_time' (56 %), 'station_type' (24 %), 'arrival_delay' (19 %) and 'train_type' (1 %).

The correlation between scheduled dwell times, arrival delays and the response, determined using robust linear regression are visible from the internal nodes of the tree. The interpretation of the splits is therefore consistent with the interpretation of correlation coefficients. However, terminal nodes and splits on the lower level of the tree did not manage to capture dwell time dependence on peak-hours. Moreover, train type does not influence any split of the tree. This can be explained with the correlation between station type and train type. In particular, data points for small stations imply the local train type.

The overall quality of the regression tree model is determined by a tenfold cross-validation. Figure 6 shows the decrease of prediction error (left) and the increase of R^2 (right) with increasing number of splits in the tree. For the optimal number of

Table 4 Summary of the LTS model for dwell time prediction

	Dependent variable: dwell_time	
	Coefficient	*p* value
peak_hour = 1	5.2223	***
arrival_delay	−0.1163	***
train_type = 'local'	2.2810	***
station_type = 'small'	7.4580	***
scheduled_time	1.0070	***
Intercept	38.7810	***
R^2	0.7281	
Residual Std. Error	31.5000	
F Statistic	73,700.2481	***

*** $p < 0.01$

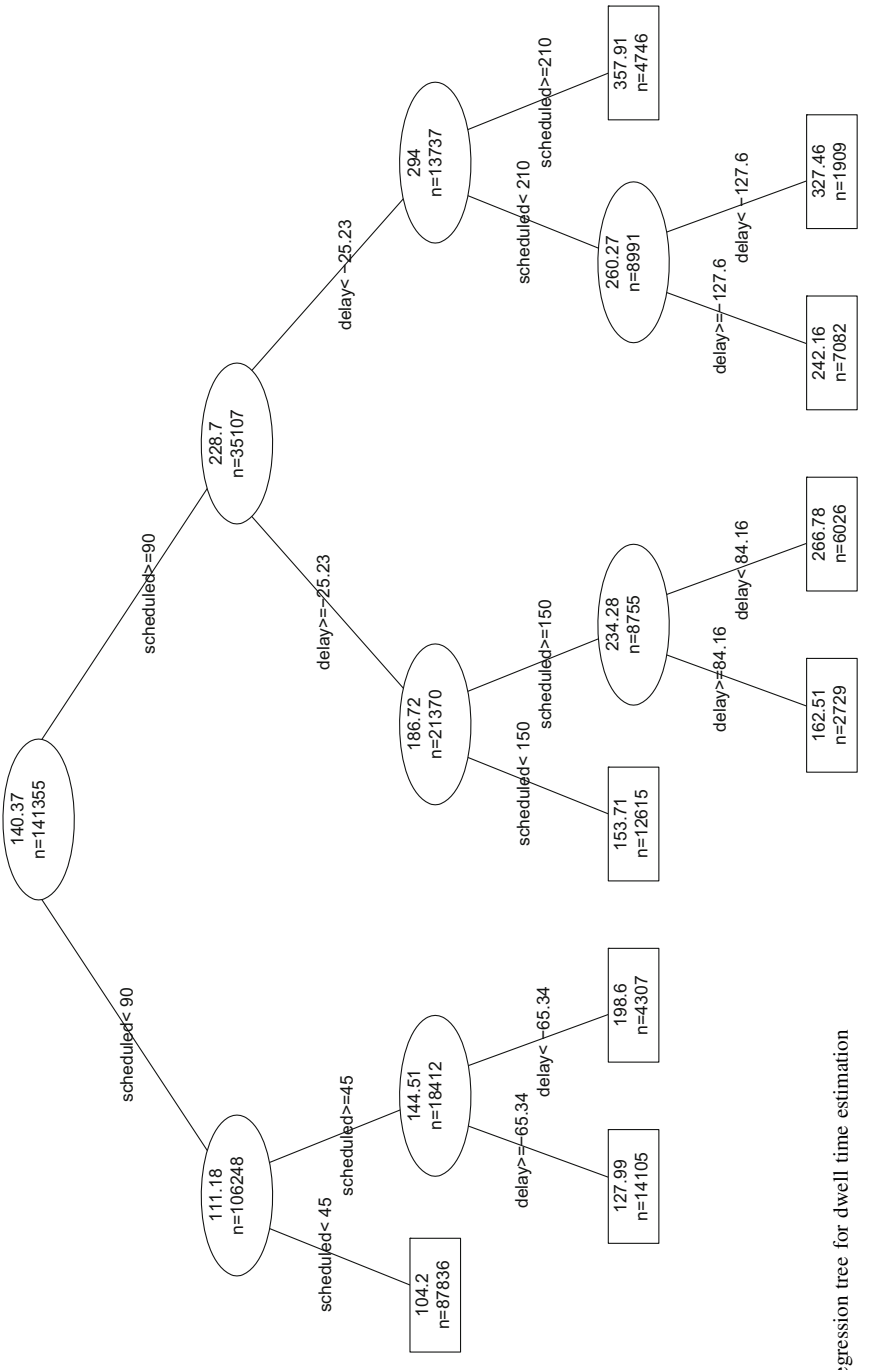

Fig. 5 Regression tree for dwell time estimation

splits 70 % of variation of dwell time from the training set can be explained using the presented regression tree model. The application of the regression tree method for prediction of dwell times resolved the issue of mutually correlated and interacting predictors, and non-linear impact on the response variable. The resulting tree is easy to interpret and the relative importance of each considered variable is given. However, the predictive power of the global model did not improve. Sensitivity to outliers, especially in lower internal and terminal nodes, may cause inaccuracy of prediction.

5.2.3 Random forest model for dwell time estimation

We attempt to improve robustness against outliers and prediction accuracy of the global model by applying the random forest method on the training set (Liaw and Wiener 2002). The resulting random forest contains 300 trees. Each split in each tree is created by choosing the best out of three randomly selected predictors.

The indicators used to examine the quality of the model are MSE and R^2. Figure 7 shows the reduction of MSE (left) and the increase of R^2 (right) with increasing number of trees in the forest. No significant improvement is achieved for forests larger than 100 trees. The final value of R^2 shows that by using the random forest algorithm, 76 % of dwell time variability can be explained and predicted. Thus, the predictive power improved compared to the regression tree model.

6 Process time estimates by local models

A running process in a local model is defined by the train line and block section, and a dwell process by the train line and station of the scheduled stop. LTS robust linear regression is used to predict running and dwell times depending on departure and

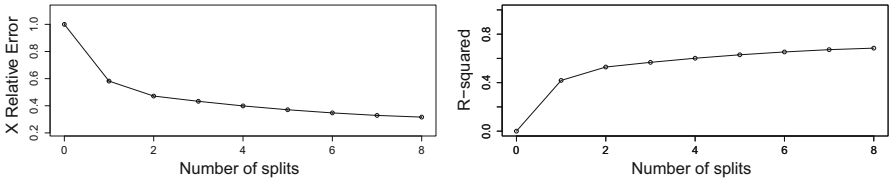

Fig. 6 Relative prediction error (*left*) and R^2 (*right*) for regression tree dwell time model

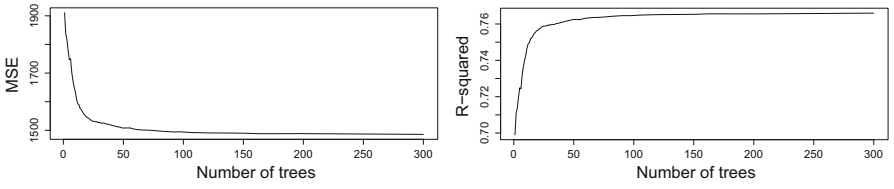

Fig. 7 MSE (*left*) and R^2 (*right*) of the random forest dwell time model

arrival delay, respectively. The assumption about different behaviour of delayed (delay larger than 60 s) and punctual or early trains is tested by separating the set of observed running and dwell times into corresponding sets of delayed and punctual trains and applying the Wilcoxon rank-sum test at 5 % significance level, with the null hypothesis that samples have distributions with equal medians.

6.1 Estimation of running times over a particular block

In Fig. 8, the realised running times are presented relative to the departure delay. The left part of the figure shows the dependence of running time over the last block before the scheduled stop in Delft station of train line 2200. The solid red line in the left part of the figure represents the robust fit. The black dashed line represents the tenth percentile of running times used as a robust estimator of minimum running times in order to avoid unrealistically low values for large delays (Van der Meer et al. 2010). A weak correlation between running times and departure delays was found on the level of block sections. Similar results were obtained for other blocks and train lines. Hence, most drivers did not adjust their train speed with respect to delay. The driver manual prescribes that in case of delays drivers must run at a higher speed where possible, however no online delay information or speed advice is available and moreover the traffic intensity of the considered railway corridors is amongst the highest in the Netherlands, so that train drivers may also try to avoid running in on a preceding train and thus keep their normal speed behaviour.

The Wilcoxon rank-sum test rejected the null hypothesis with $p \approx 0$ thus indicating different distributions of running times of delayed trains. The box-plots in Fig. 8 (right) show small differences in distributions of six data samples specified based on the value of departure delay. The box-plots used in this paper indicate the median (line in the middle of the box), the first and the third quartiles (upper and lower bound of the box) and data maximum and minimum (ends of the upper and lower whisker). Note that the outliers are excluded from the plots, for the sake of clarity of the figures, but not from the data set used to compute the quartiles. Outliers are detected in a conventional procedure by adding (subtracting) the

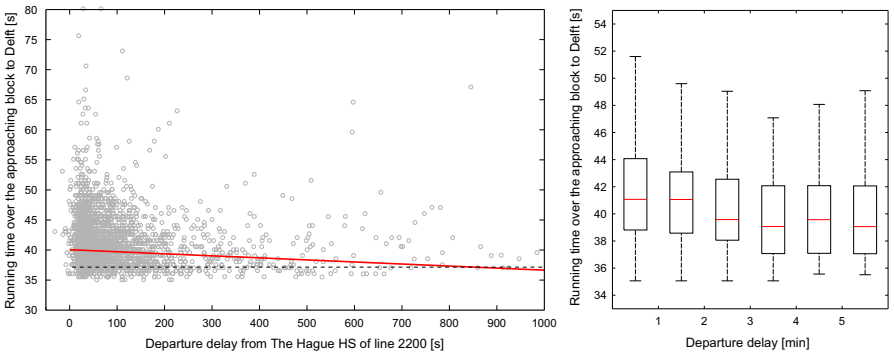

Fig. 8 Dependence of running time on delay (*left*) and box-plots of running times for punctual and delayed trains (*right*)

interquartile difference multiplied by 1.5 to (from) the upper (lower) quartile. All values outside the obtained range are considered as outliers.

6.2 Estimation of dwell times for a particular station

The dependence of dwell times on arrival delays was examined. The correlation is particularly strong for large stations. In smaller stations where only local trains are scheduled to stop, no significant correlation between dwell times and arrival delays was established. That can be explained by the fact that these stops are scheduled as short stops (Sect. 3.3.2). The trains only stop for boarding and alighting and depart as soon as possible.

Figure 9 (left) shows the dependence of dwell times on arrival delays for the train line 2200 in station Delft. The horizontal black dashed line represents the tenth percentile of all dwell times, whereas the red line represents the robust linear fit for punctual trains. The scheduled dwell time is 60 s. A strong correlation ($R^2 = 0.8704$) was captured for early and punctual trains. The Wilcoxon rank-sum test rejected the null hypothesis of equal dwell time for early and late trains ($p \approx 0$) and indeed the box-plots in Fig. 9 (right) show clearly different distributions of dwell times for punctual and late trains. Moreover, the variation of dwell times for delayed trains needs to be explained by other factors, and therefore the data set is divided into a set of punctual and delayed trains at the threshold of 60 s.

The variability of dwell times of delayed trains is explained by modelling dwell time as a time series to determine the impact of peak-hours. The dwell times of delayed trains normally equal the minimum dwell time required for passenger operations and route setting if the delay exceeds the dwell buffer time. We assumed that passenger volumes and consequently the time needed for alighting and boarding increases during peak-hours. Figure 10 shows dwell times (weekends and holidays were not considered) relative to the scheduled arrival times of train line 2200 in Delft. The increase in dwell times during peak-hours is clearly visible. The red line indicates the median dwell time.

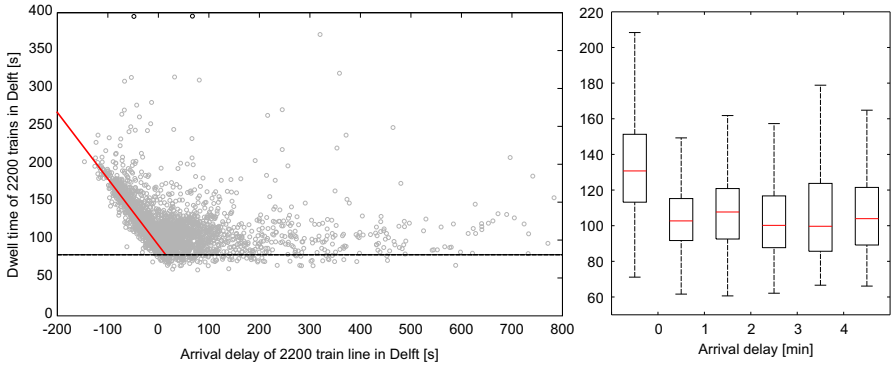

Fig. 9 Dependence of dwell time on delay (*left*) and box-plots of dwell time (*right*)

This clear distinction between causes of variability of dwell time for punctual and delayed trains requires a separate approach to prediction of dwell times. Therefore, dwell time for a delayed train is estimated from historical data based on dwell times of the same train number and adjacent train numbers of the same series (e.g. if train 2245 arrived with a delay, the dwell time will be predicted as the average dwell time of trains 2243, 2245 and 2247 obtained from the data set of delayed trains). The reason for including the data from the adjacent train numbers is to ensure the sufficient sample size and robustness of the moving average estimate (Van der Meer et al. 2010).

Figure 11 shows the effect of using the described moving average approach for predicting the dwell times of delayed trains on a test set. The prediction accuracy is compared to the approach based on LTS robust regression. The prediction error is computed by subtracting the estimate from the realised dwell time. The positive bias of the LTS estimate error indicates that dwell times of delayed trains are underestimated. This approach assumes a minimum dwell time for delayed trains thus disregarding the effects of peak-hours. The moving average approach significantly improves the prediction accuracy.

7 Comparison of statistical models

7.1 Validation of running and dwell time estimation models

Figure 12 shows the distribution of prediction error for each running time prediction model. The test set for running time estimation consists of the processed data for 10 days of traffic in areas Rotterdam and The Hague and contains 18,684 data points. All presented approaches quickly produce estimates for each instance from the test set. The random forests clearly give the most accurate estimates of running

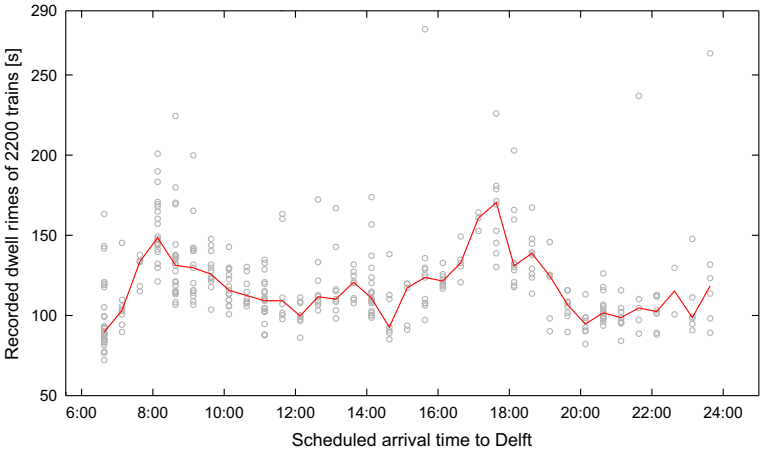

Fig. 10 Dependence of dwell time on scheduled departure time

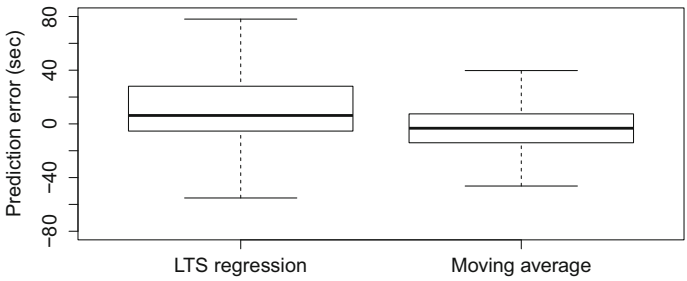

Fig. 11 Prediction error for dwell times of delayed trains

times with respect to other global models. The performance of random forests is comparable to the performance of local models that give the most accurate predictions of running times. Note the very small prediction errors within ± 10 s for the local LTS and random forest estimates.

Similar results are obtained for dwell time estimation (Fig. 13). The size of the test set for dwell times is 12,225. Random forests are the best performing global model. However, local models, consisting of an LTS robust linear regression model for punctual trains and a time series (TS) model for delayed trains, give more accurate estimates for the dwell times. The high standard deviation of the prediction error even for the most precise model indicates that relying on track occupation data as the sole data source for dwell time prediction may not result in sufficient accuracy.

7.2 Comparison of prediction accuracy for scheduled processes

To offer a fair comparison of the presented models, the accuracy of the estimates relative to the scheduled processes needs to be considered. Recall that the running time analysis presented in this paper relates to running times over block sections. The running time between two scheduled stops can be computed as the sum of the running times over blocks in the route of the train including the outbound route from the station of departure and the inbound route at the arrival station. In order to exclude the impact that other trains may have had on the running times of trains in the test set, all hindered train runs were excluded from the analysis. Figure 14 shows

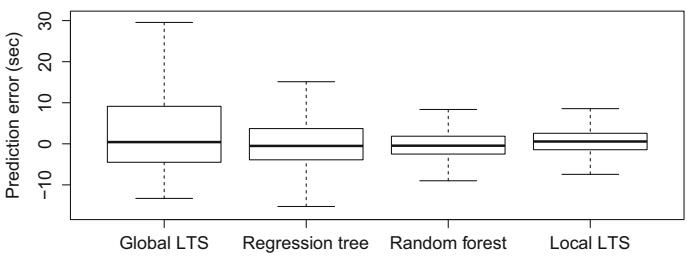

Fig. 12 Prediction error of running time estimation models

Fig. 13 Prediction error of dwell time estimation models

a comparison of the absolute prediction errors for dwell time and running time estimates (left) and the errors relative to the scheduled process times (right). The results are obtained using the local LTS models. The running times between two scheduled stops are clearly predicted more precisely than dwell times. This is also demonstrated by comparing the relative errors, obtained with respect to the scheduled time of the corresponding process. The errors of running time estimates are within 10 % of the corresponding scheduled running times, whereas the error of dwell time estimates may be even larger than the corresponding scheduled dwell times.

The methodology presented in this paper was used for the calibration of a real-time prediction model described in Kecman and Goverde (2014). Having in mind the real-time character of the tool, the main criteria for comparing the data-driven approach with other relevant approaches are prediction accuracy and computational requirements. Our approach quickly generates robust predictions of future process times for multiple trains. It is therefore appropriate for use in an online environment with frequent updates of train positions. Dolder et al. (2009) presented a comparable online prediction tool where running times are calculated using train motion equations. Due to a computationally demanding procedure for computing train trajectories, their approach uses deterministic offline computed estimates. Regarding the accuracy of prediction, the data-driven approach produces more flexible and accurate predictions for all prediction horizons.

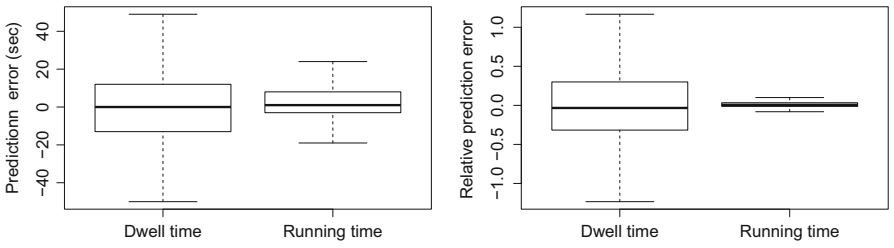

Fig. 14 Precision of dwell time and running time estimates (*left*), relative to scheduled time (*right*)

8 Conclusions

This paper presented four data driven approaches for estimation of conflict-free running times and dwell times. Three global models were developed by collecting all running time and dwell time data from the training set and creating a separate predictive model for estimation of each type of process time. Advanced supervised learning methods were tested and compared on predictive power, interpretability of results, and accuracy. On the other hand, local running time and dwell time models for a particular block and station were developed based on the data structure of the training set. Both approaches were validated on an independent test set. The local models provided more accurate predictions for both running and dwell times.

The running times showed a small variation which was to a great extent explained by predictors in both models. A weak dependence on delays was established. The data analysis showed that the majority of trains run with maximum performance regardless of departure delays. Furthermore, running times seem to be weakly affected by peak hours and do not have a significant daily variation. An interesting observation is that even for conflict-free train runs a short minimum headway after the preceding train may cause an extended running time in order to prevent a route conflict. The high accuracy of running time estimates on the level of block sections and between scheduled stops indicates the applicability of the proposed data-driven approach for calibration of running times in railway traffic models with different degrees of granularity.

Dwell times of punctual trains show a strong correlation with arrival delays, in particular in large stations. On the other hand, the dwell times of delayed trains are more sensitive to the impact of passenger volume variability in peak and off-peak periods. Despite the strong predictive power of the presented dwell time models, the validation on an independent test set showed that the variability of dwell times cannot be fully explained by the selected predictor variables. Hence, dwell times need to be modelled by more factors since the variation of prediction error is significantly larger than for running times. One way to do this is to include other data sources such as platform design and rolling-stock properties to improve the predictions. This might enable computation of more accurate estimates of arrival and departure events. Moreover, data related to behavioural properties of passengers and train drivers can be used to derive more accurate estimates of dwell times.

The major advantage of the global model is that the results can be generalised and applied to other parts of the network and different train lines that are not included in the training data set. However, calibration of the global model as well as the application in real-time is computationally more demanding than creating the multiple local models and using them for prediction. Therefore, the local models were used for real-time calibration of the railway traffic prediction model described in Kecman and Goverde (2014).

Acknowledgments This work is partially funded by the Dutch Technology Foundation STW, research project: Model-Predictive Railway Traffic Management (Project No. 11025).

References

Abril M, Barber F, Ingolotti L, Salido M, Tormos P, Lova A (2008) An assessment of railway capacity. Transp Res Part E Logist Transp Rev 44(5):774–806

Berger A, Gebhardt A, Müller-Hannemann M, Ostrowski M (2011) Stochastic delay prediction in large train networks. In: Caprara A, Kontogiannis S (eds) 11th Workshop on algorithmic approaches for transportation modelling, optimization, and systems, Dagstuhl, pp 100–111

Bešinović N, Quaglietta E, Goverde RMP (2013) A simulation-based optimization approach for the calibration of dynamic train speed profiles. J Rail Transp Plan Manag 3(4):126–136

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman J, Ohlsen R, Stone C (1984) Classification and regression trees. Wadsworth, New York

Brünger O, Dahlhaus E (2014) Running time estimation. In: Hansen IA, Pachl J (eds) Railway timetable and operations—analysis. Modelling, optimisation, simulation, performance evaluation. Eurailpress, Hamburg, pp 65–89

Buchmueller S, Weidmann U, Nash A (2008) Development of a dwell time calculation model for timetable planning. In: Allan J, Brebbia CA, Rumsey AF, Sciutto G, Sone S, Goodman CJ (eds) Computers in railways XI. WIT Press, Southampton, pp 525–534

Büker T, Seybold B (2012) Stochastic modelling of delay propagation in large networks. J Rail Transp Plan Manag 2(1–2):34–50

D'Ariano A, Pranzo M, Hansen IA (2007) Conflict resolution and train speed coordination for solving real-time timetable perturbations. IEEE Trans Intell Transp Syst 8(2):208–222

Dolder U, Krista M, Voelcker M (2009) RCS—rail control system—realtime train run simulation and conflict detection on a net wide scale based on updated train positions. In: Proceedings of the 3rd international seminar on railway operations modelling and analysis (RailZurich2009), Zurich, pp 1–15

Goverde RMP (2007) Railway timetable stability analysis using max-plus system theory. Transp Res Part B Methodol 41(2):179–201

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer Science+Business Media, New York

Hooghiemstra JS (1996) Design of regular interval timetables for strategic and tactical railway planning. In: Allan J, Brebbia CA, Hill RJ, Sciutto G, S S (eds) Computers in railways V, Computational Mechanics Publications. WIT Press, Southempton, pp 393–402

Kecman P, Goverde RMP (2012) Process mining of train describer event data and automatic conflict identification. In: Brebbia CA, Tomii N, Mera JM (eds) Computers in railways XIII, WIT transactions on the built environment, vol 127. WIT Press, Southampton, pp 227–238

Kecman P, Goverde RMP (2014) Online data-driven adaptive prediction of train event times. IEEE Trans Intell Transp Syst 16(1):465–474

Lee YC, Daamen W, Wiggenraad PBL (2007) Dwell times of public transport vehicles: a state-of-the-art report. In: Transportation Research Board 86th Annual Meeting, Washington, pp 1–14

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22

Longo G, Medeossi G (2013) An approach for calibrating and validating the simulation of complex rail networks. In: Transportation Research Board 92nd Annual Meeting, Washington, pp 1–19

Longo G, Medeossi G, Nash A (2012) Estimating train motion using detailed sensor data. In: Transportation Research Board 91st Annual Meeting, Washington, pp 1–15

Lüthi M (2009) Improving the efficiency of heavily used railway networks through integrated real-time rescheduling. Ph.D. thesis, ETH Zurich, Zurich

Medeossi G, Longo G, de Fabris S (2011) A method for using stochastic blocking times to improve timetable planning. J Rail Transp Plan Manag 1(1):1–13

Nash A, Huerlimann D (2004) Railroad simulation using OpenTrack. In: Allan J, Brebbia CA, Hill RJ, Sciutto G, Sone S (eds) Computers in railways IX. WIT Press, Southampton, pp 45–54

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Rousseeuw PJ (2005) Robust regression and outlier detection. Wiley, New York

Rousseeuw PJ, Driessen K (2006) Computing LTS regression for large data sets. Data Min Knowl Discov 12(1):29–45

Rousseeuw PJ, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Maechler M (2014) robustbase: basic robust statistics. R package version 0.90-2

Schöbel A, Schwarze S (2013) Finding delay-resistant line concepts using a game-theoretic approach. Netnomics 14(3):95–117

Stam-Van den Berg BWV, Weeda VA (2007) VTL-tool: detailed analysis of dutch railway traffic. In: Proceedings of the 3rd international seminar on railway operations modelling and analysis (RailHanover2007), Hanover, pp 1–10

Therneau T, Atkinson B, Ripley B (2014) rpart: recursive partitioning and regression trees. R package version 4.1-5

Van der Meer DJ, Goverde RMP, Hansen IA (2010) Prediction of train running times and conflicts using track occupation data. In: Proceedings of the 12th world conference on transport research (WCTR 2013), Lisbon

Wende D (ed) (2003) Fahrdynamik des Schienenverkehrs. Teubner Verlag, Wiesbaden, B.G **(in German)**

Wiggenraad PBL (2001) Alighting and boarding times of passengers at Dutch railway stations—analysis of data collected at 7 stations in October 2000. In: Papers of the TRAIL workshop train delay at stations and network stability, TRAIL Research Scool, Delft