# Train Overtaking Prediction in Railway Networks: a Big Data Perspective

Luca Oneto[1], Irene Buselli[1], Alessandro Lulli[1],
Renzo Canepa[2], Simone Petralli[2], and Davide Anguita[1]

[1] DIBRIS, University of Genoa, Via Opera Pia 13, I-16145, Genoa, Italy,
(email: {luca.oneto, irene.buselli, alessandro.lulli, davide.anguita}@unige.it)
[2] Rete Ferroviaria Italiana S.p.A., Via Don Vincenzo Minetti 6/5, I-16126, Genoa,
Italy (email: {r.canepa, s.petralli}@rfi.it)

**Abstract.** Every time two or more trains are in the wrong relative position on the railway network because of maintenance, delays or other causes, it is required to decide if, where, and when to make them overtake. This is a quite complex problem that is tackled every day by the train operators exploiting their knowledge and experience since no effective automatic tools are available for large scale railway networks. In this work we propose a train overtaking hybrid prediction system. Our model is hybrid in the sense that it is able to both encapsulate the experience of the operators and integrate this knowledge with information coming from the historical data about the railway network using state-of-the-art data-driven techniques. Results on real world data coming from the Italian railway network will show that the proposed solution outperforms the fully data-driven approach and could help the operators in timely identify and schedule the best train overtaking solution.

**Keywords:** Railway Network, Train Overtaking, Big Data, Data-Driven Models, Hybrid Models.

## 1 Introduction

Railway Transportation Systems (RTSs) play a vital and crucial role in public mobility and goods delivery. In Europe the increasing volume of people and freight transported on railway is congesting the network [5]. The only fast and economically viable way to increase capacity is then to improve the efficiency of daily operations in order to be able to control a larger number of running trains without requiring massive public investments in new physical assets [23]. For this reason, in the last years, every actors of the RTSs has started extensive modernization programs that leverage on advanced information and communication solutions. The objectives are to improve system safety and service reliability, to enhance passenger experience, to provide higher transit capacity and to reduce operational costs.

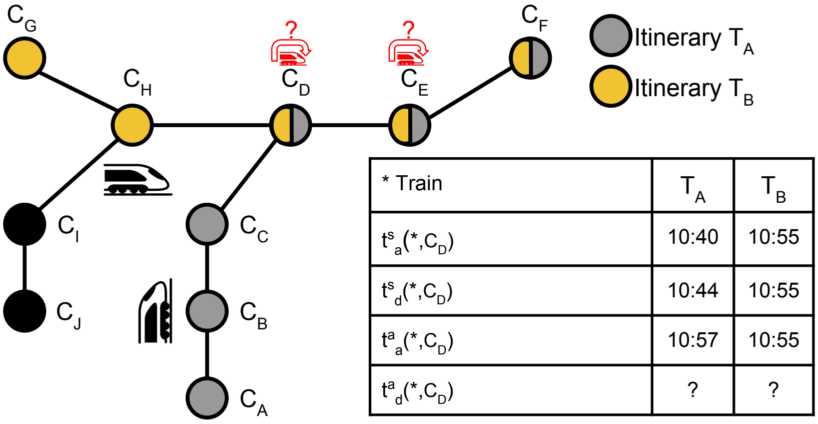
In this work we focus on the problem of analyzing the train movements in Large-Scale RTSs for the purpose of understanding and predicting their behaviour. In particular, we will study the problem of the train overtaking prediction exploiting data-driven solutions leveraging on the huge amount of data produced and stored by the new RTSs information systems. Train overtaking prediction is the problem of predicting when it is required or preferable to make a train perform an overtaking in order to minimize the train delays and the penalty costs associated with them. The study of this problem allows to improve the quality of service, the train circulation, and the Infrastructure Managers and Train Operators management costs.

A large literature covering the prediction problems related to the train circulation already exists [10]. However, in general, the majority of the works focus on different problems: the running time prediction, the dwell time prediction, and the train delay prediction. These focus on predicting, respectively, the amount of time needed to traverse a section of railway between two checkpoints [1, 6, 11, 13, 15, 17], the amount of time spent in a checkpoint and the difference between the actual arrival (or departing time) and the scheduled one in each of the stations composing the itinerary of a train [2,3,9,12,14,19,20,24].

The problem of train overtaking prediction, instead, has never been studied exploiting data-driven solutions. Current solutions model this problem as a complex optimization task [7,8,16] that is usually not easy to solve (or impossible to solve in large scale railway networks) and requires a lot of human effort during the modeling phase. For this reason, in practice, the current solution is to rely on the experience of the operators and on their knowledge of the network. We call this solution Experience-Based Model (EBM). A solution that we proposed here is to adopt a data-driven approach. In this framework, advanced analytic methods [10, 13] can be exploited to analyze the historical data and to build Data-Driven Models (DDMs) which automatically predicts when it is better to perform the train overtaking. Unfortunately, also DDMs have their drawbacks since they do not handle easily the fact that prior knowledge about the problem may be available, apart from the historical data. For this reason, in this work, we propose an hybrid approach to the train overtaking prediction problem mixing together the EBMs and DDMs taking inspiration from our previous works, where we employed a similar idea to deal with the running time, the dwell time and the train delay prediction problems [18]. The combination of the two approaches allows us to create a model that shows the strengths of both EBMs and DDMs while limiting their weaknesses. On one hand, encapsulating the experience of the operators enables the creation of an interpretable and robust model which can be better exploited in a human-oriented environment like the one of the train operators. On the other hand, the exploitation of data-driven techniques allows to build more accurate predictive models. Results on real world data about the Italian railway network provided by Rete Ferroviaria Italiana (RFI - the Italian Infrastructure Manager) will show the effectiveness of our proposal.

## 2 The Train Overtaking Prediction Problem

In this section we introduce the notation needed to formally describe the problem of train overtaking. A railway network can be easily described with a graph. Figure 1 depicts a simplified railway network where two trains follow their itineraries. Let us consider the train at the station $C_B$, characterized by its itinerary (origin at station $C_A$, destination at station $C_F$, some stops and some transits). In the following, we will call checkpoint a station without differentiating where the train stops or transits and between actual stations and points of measure. The railway sections are the pieces of the network between two consecutive checkpoints and have also an orientation (e.g. transit from $C_D$ to $C_E$ is different from transit from $C_E$ to $C_D$).



| * Train | $T_A$ | $T_B$ |
|---|---|---|
| $t^s_a(*,C_D)$ | 10:40 | 10:55 |
| $t^s_d(*,C_D)$ | 10:44 | 10:55 |
| $t^a_a(*,C_D)$ | 10:57 | 10:55 |
| $t^a_d(*,C_D)$ | ? | ? |

**Fig. 1.** Train *Overtaking.*

For any checkpoint $C$ in the itinerary, the train $T$ is scheduled to arrive and depart at different specified times, defined in the timetable, respectively, $t^s_a(T, C)$ and $t^s_d(T, C)$. The difference between the actual time (either for arrival $t^a_a(T, C)$ or for departure $t^a_d(T, C)$) and the scheduled time is defined as train delay. Each train and checkpoint have additional characteristics such as an unique identifier, the category of the train, and the category of the railway network. However, due to delays caused by many different reasons, it is common that two trains are in a wrong relative position along their itinerary.

Let us refer again to Figure 1 for a graphical description of the problem. In our scenario, we say that two trains are in a wrong relative position at a checkpoint $C_D$ when $t^s_a(T_A, C_D) < t^s_a(T_B, C_D)$ and $t^a_a(T_A, C_D) > t^a_a(T_B, C_D)$, i.e. $T_A$ is expected to arrive before $T_B$ in checkpoint $C_D$ but, for some reason, train $T_B$ arrives before train $T_A$ in checkpoint $C_D$.

When an event like this occurs it is required to predict and enforce as soon as possible, with an overtake, the correct relative position of the trains for the purpose of minimizing delays and deviations from the timetable. In order to enforce the overtake it is necessary to predict what is the best subsequent station in the itinerary to perform the overtake minimizing the deviation from the timetable. Note that not all the checkpoints allow the overtake (e.g. no additional track is available to perform the overtake in all the checkpoints).

In the example of Figure 1, the system detects the incorrect position in checkpoint $C_D$ because train $T_A$ is scheduled to arrive before train $T_B$ on checkpoint $C_D$, but train $T_A$ has a delay and, for this reason, train $T_B$ arrives before train $T_A$ in checkpoint $C_D$. After the detection of the incorrect relative position, the system starts evaluating if and when, on the current or subsequent checkpoints in common between the itineraries of train $T_A$ and $T_B$, it is preferable to make train $T_A$ overtake train $T_B$.

## 3 Our Proposal: an Hybrid Model

In this work, we propose an HM to tackle the train overtaking prediction problem. In particular, we mapped the problem of overtaking into a series of binary classification problems where the task is to predict if or not an overtake will be performed at a particular checkpoint. The idea is to leverage on both the experience of the operators (EBM) and historical data with the use of advanced data analytics methods (DDM). The goal is to build an accurate, dynamic, robust, and interpretable model able to support the decision of the operators.

For this purpose we propose a two level architecture. At the top level, we construct a tree following the suggestions of the operators, which captures the characteristics of the two trains under consideration for the overtake and additional information to better describe the scenario under examination. Such top level tree encapsulates the EBM developed by the operators during the years. At the bottom level, for each of the leaves composing the tree we have built a dataset with all the past occurrences of the overtake corresponding to that particular leaf. This dataset is richer, in terms of feature set, with respect to the top level tree and, leveraging on this we have built a DDM able to improve the accuracy of the top level tree.

More in detail, the top level decision tree encapsulates the experience of the operators in taking decisions when two trains are in the wrong relative position and one has to overtake the other (see Section 2). The proposed HM groups all the possible situations in subgroups based on a series of similarity variables (see Table 1), defined together with the RFI experts, which allow us to have, on one side, robust statistics, thanks to the possibility to learn from a reasonable group of similar overtake situations and, on the other side, a rich feature set, able to capture the variability of the phenomena. Then, in each leaf of this tree, we exploit a DDM able to learn from the historical data in that particular leaf, based on a super-set of features with respect to the one used in the tree (see Table 2). In particular, each leaf is a Random Forest (RF) classifier [4] (following the experience of the DDM developed in [21]), which predicts if the trains will perform the overtake in a particular checkpoint. The whole HM is built and updated incrementally as soon as new train movements are recorded. During the prediction phase, instead, we just visit the tree considering the particular overtake situation and we exploit the corresponding RF classifier to make the actual prediction.

**Table 1.** HM top level decision tree feature set (Cat. means Categorical)

| Feature Name | Cat. | Description |
|---|---|---|
| Railway Section | Yes | The considered railway section |
| Railway Checkpoint | Yes | The considered railway checkpoint |
| Train Type | Yes | The considered Train Type |
| Daytime | No | The time of the day with an hourly granularity |
| Weekday | Yes | The day of the week |
| Last Delay | No | The last known delay with the following granularity in minutes ($[0, 2]$, $(2, 5]$, $(5, 10]$, $(10, 20]$, $(20, 30]$, $(30, 60]$, $(60, 120]$, $(120, \infty)$); |
| Weather Conditions | Yes | The weather conditions (Sunny, Light Rain, Heavy Rain, Snow). |

**Table 2.** HM bottom level RF feature set (Cat. means Categorical).

| Feature Name | Cat. | Description |
|---|---|---|
| Weather Information | Yes | Weather conditions (Sunny, Light Rain, etc.) in all the checkpoints of the train itinerary (for the already traveled checkpoints we use the actual weather while for the future checkpoints we use the predicted weather conditions) |
| Past *Train Delays* | No | Average value of the past *Train Delays* in seconds & Last known *Train Delay* |
| Past *Dwell Times* | No | Average value of the past differences between actual and scheduled *Dwell Times* in seconds & Last known difference between actual and scheduled *Dwell Time* |
| Past *Running Times* | No | Average value of the past differences between actual and scheduled *Running Times* in seconds & Last known difference between actual and scheduled *Running Time* |
| Network Congestion | No | Number of trains traversing the checkpoints of the train itinerary in a slot of 20 minutes around the actual and scheduled times respectively for the past and future checkpoints |
| Network Congestion Delays | No | Average *Train Delay* of the trains traversing the checkpoints of the train itinerary in a slot of 20 minutes around the actual and scheduled times respectively for the past and future checkpoints |

## 4 Experimental Evaluation

In this section we will perform an extensive evaluation of the proposed HM to show its effectiveness based on real world data coming from the Italian railway network and provided by RFI. We will prove the effectiveness of our approach by using, as a baseline, a fully DDM based on the one derived in [18, 21] for predicting delays, transit time, and dwell time.

### 4.1 The Available Data

The experiments have been conducted exploiting the real data provided by RFI about the Italian railway network[3]. In particular, RFI has provided

- data about train movements which contains the following information: Date, Train ID, Checkpoint ID, Actual Arrival Time, Arrival Delay, Actual Departure Time, Departure Delay and Event Type. The Event Type field can assume different values: Origin (O), Destination (D), Stop (F), Transit (T);
- timetables, including planning of exceptional trains, and cancellations.

For the purpose of this work, RFI provided the access to the data of 12 months (the whole 2016 solar year) of train movements of one critical (in the sense that many overtakes need to be planned every day) Italian region. The data are relative to more than 3.000 trains and 200 checkpoints. The dataset contains 5.000.000 train movements.

For improving the quality of the predictors, we also exploited as exogenous information, with a Big Data oriented approach, the weather data coming weather stations in the area, freely available from the Italian weather services [22]. For each checkpoint we consider the closest weather station. Then we collected the historical data relative to the solar radiation and precipitations for the same time span. From this data it is possible to extract both the actual and the forecasted weather conditions (Sunny, Rain, Heavy Rain, and Snow).

### 4.2 Our baseline: a Data Driven Model

In order to better understand the potentiality and effectiveness of our HM we decided to use a purely DDM as a baseline. The DDM has been constructed removing the top level tree structure of HM described in Section 3, which models the experience of the operators, and by building a single DDM based on the whole set of available data and the features reported in Table 2. By removing the top level structure of the HM we basically remove from the HM the experience of the operators resulting in a fully DDM. Instead of creating sub-groups of overtake situations sharing similar characteristics like in the HM, we leave to the DDM the task to learn everything from the historical data and perform the predictions.

A comparison with a fully EBM could not be performed since the Italian RTS information system does not store the prediction made by the operators, who just rely on their intuition and experience.

### 4.3 Results

We start this analysis showing which checkpoints were involved in the most and the least number of overtakings in 2016, and the number of overtakings identified by the HM and the DDM. Table 3 depicts the 5 checkpoints in which the most and least number of overtakes happened.

From Table 3 we can observe that:

[3] We cannot report all the details because of confidentiality issues.

**Table 3.** The checkpoints having the most and least number of overtakes in 2016.

| Most Overtaking | | | |
|---|---|---|---|
| Checkpoint Name | Real | HM | DDM |
| Checkpoint *A* | 624 | 518 | 378 |
| Checkpoint *B* | 273 | 248 | 226 |
| Checkpoint *C* | 220 | 211 | 152 |
| Checkpoint *D* | 188 | 206 | 185 |
| Checkpoint *E* | 177 | 168 | 164 |

| Least Overtaking | | | |
|---|---|---|---|
| Checkpoint Name | Real | HM | DDM |
| Checkpoint *F* | 17 | 0 | 1 |
| Checkpoint *G* | 14 | 12 | 11 |
| Checkpoint *H* | 13 | 2 | 2 |
| Checkpoint *I* | 11 | 1 | 3 |
| Checkpoint *L* | 10 | 2 | 0 |

- the 3 checkpoints which had the highest number of overtakes account for around the 50% of the total overtaking happened in 2016 in the area under examination;
- HM predictions are close to the real one;
- in general HM underestimates the number of overtakings, which is expected because the model requires a certain number of historical data before starting to work correctly;
- The DDM underestimates even more the overtakes in each checkpoint, reflecting that the amount of data required to learn how to correctly predict an overtake is larger.

It is possible to make use of the information of Table 3 in order to identify in which checkpoints it is possible an overtake. In fact the DDM need to learn this information directly from the data while, in the HM, we can exploit the experience of the operators and plug this information directly in the model. In the Italian RTS, overtakes are not possible in all the checkpoints, and, in the area under examination, only in 48 checkpoints it is possible to perform an overtaking.

At this point we are ready to show the precision of the HM in identifying the overtakings. Tables 4 and 5 depict the confusion matrix of, respectively, the DDM and the HM. We reported also the confusion matrices relative to the different typologies of trains. From Tables 4 and 5 we can observe that:

- the HM clearly outperforms the DDM;
- the HM is highly accurate in predicting when two trains must not swap their positions;
- freight trains are the ones which perform the least number of overtakes and also the ones affected by the largest error;
- high speed trains are the ones which perform the highest number of overtakes and the ones in which the HM performs the least number of false-positive and false-negative predictions.

Finally, Figure 4.3 reports the accuracy of the HM and the DDM in detecting overtakings during the whole 2016. We report both the overall accuracy and the

**Table 4.** Confusion matrices for the overtakes predicted by the DDM (ALL: all trains. REG: regional trains. HS: high speed trains. FRE: freight trains).

| ALL | Yes | No |
|---|---|---|
| Yes | 2246 | 893 |
| No | 639 | 12657 |

| REG | Yes | No |
|---|---|---|
| Yes | 1082 | 476 |
| No | 347 | 7870 |

| HS | Yes | No |
|---|---|---|
| Yes | 1055 | 307 |
| No | 175 | 2599 |

| FRE | Yes | No |
|---|---|---|
| Yes | 109 | 110 |
| No | 117 | 2188 |

**Table 5.** Confusion matrices for the overtakes predicted by the HM (ALL: all trains. REG: regional trains. HS: high speed trains. FRE: freight trains).

| ALL | Yes | No |
|---|---|---|
| Yes | 2355 | 783 |
| No | 500 | 12827 |

| REG | Yes | No |
|---|---|---|
| Yes | 1121 | 436 |
| No | 318 | 7899 |

| HS | Yes | No |
|---|---|---|
| Yes | 1112 | 250 |
| No | 103 | 2671 |

| FRE | Yes | No |
|---|---|---|
| Yes | 122 | 97 |
| No | 79 | 2226 |

recall (actual overtakes in a checkpoint correctly predicted). From Figure 4.3 we can observe that:

- it is required 1 month of data to have a fully operational HM while the DDM requires more data.
- in general the HM shows higher accuracy with respect to the DDM and this advantage is costant over the whole year.

## 5 Conclusion

In this work we dealt with the problem of understanding and predicting the train overtakes. For this purpose, we exploited an hybrid approach which is able to encapsulate in one single model the knowledge about the network, the experience of the operators, the historical data, and other exogenous variables taking inspiration from the state-of-the-art approaches in this field of research. The result is a dynamic, interpretable and robust hybrid data analytics system able to handle non recurrent events, changes in the behaviour of the network,

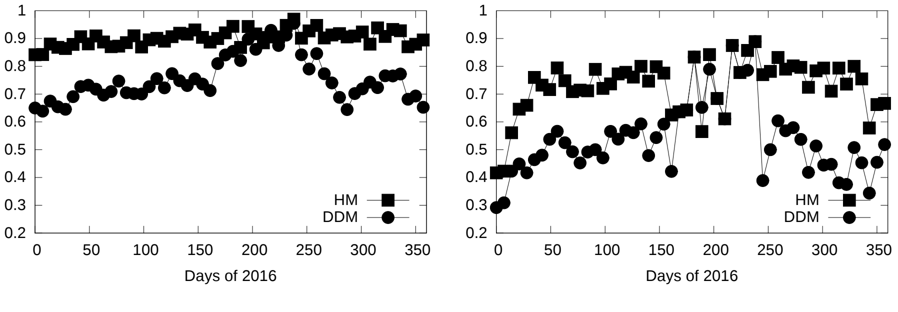


**Fig. 2.** Overtakes prediction accuracy in time. Overall accuracy (left). Recall (right).

and to consider complex and exogenous information like weather information. Basically, the proposed approach preserves the strengths of the experience based methods and the data-driven methods and limits their weaknesses. Results on real world data coming from the Italian railway network show that the proposed solution provides remarkable results in addressing the train overtakes prediction problem.

**Acknowledgments** This research has been supported by the European Union through the projects IN2DREAMS (European Union's Horizon 2020 research and innovation programme under grant agreement 777596).

## References

1. Albrecht, T.: Reducing power peaks and energy consumption in rail transit systems by simultaneous train running time control. WIT Transactions on State-of-the-art in Science and Engineering 39 (2010)
2. Barta, J., Rizzoli, A.E., Salani, M., Gambardella, L.M.: Statistical modelling of delays in a rail freight transportation network. In: Proceedings of the Winter Simulation Conference (2012)
3. Berger, A., Gebhardt, A., Müller-Hannemann, M., Ostrowski, M.: Stochastic delay prediction in large train networks. In: OASIcs-OpenAccess Series in Informatics. vol. 20 (2011)
4. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
5. Bryan, J., Weisbrod, G.E., Martland, C.D.: Rail freight solutions to roadway congestion: Final report and guidebook. Transportation Research Board (2007)
6. Daamen, W., Goverde, R.M.P., Hansen, I.A.: Non-discriminatory automatic registration of knock-on train delays. Networks and Spatial Economics 9(1), 47–61 (2009)
7. D'Ariano, A.: Improving real-time train dispatching: Models, algorithms and applications. TRAIL Research School (2008)
8. D'Ariano, A., Pranzo, M.: An advanced real-time train dispatching system for minimizing the propagation of delays in a dispatching area under severe disturbances. Networks and Spatial Economics 9(1), 63–84 (2009)
9. Fang, W., Yang, S., Yao, X.: A survey on problem models and solution approaches to rescheduling in railway networks. IEEE Transactions on Intelligent Transportation Systems 16(6), 2997–3016 (2015)
10. Ghofrani, F., He, Q., Goverde, R.M., Liu, X.: Recent applications of big data analytics in railway transportation systems: A survey. Transportation Research Part C: Emerging Technologies 90, 226–246 (2018)
11. Goverde, R.M.P., Meng, L.: Advanced monitoring and management information of railway operations. Journal of Rail Transport Planning & Management 1(2), 69–79 (2011)
12. Hansen, I.A., Goverde, R.M.P., Van Der Meer, D.J.: Online train delay recognition and running time prediction. In: Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on. pp. 1783–1788 (2010)
13. Kecman, P., Goverde, R.M.P.: Process mining of train describer event data and automatic conflict identification. Computers in Railways XIII: Computer System Design and Operation in the Railway and Other Transit Systems 127, 227 (2013)

14. Kecman, P., Goverde, R.M.P.: Online data-driven adaptive prediction of train event times. IEEE Transactions on Intelligent Transportation Systems 16(1), 465–474 (2015)
15. Ko, H., Koseki, T., Miyatake, M.: Application of dynamic programming to the optimization of the running profile of a train. WIT Transactions on The Built Environment 74 (2004)
16. Lamorgese, L., Mannino, C.: An exact decomposition approach for the real-time train dispatching problem. Operations Research 63(1), 48–64 (2015)
17. Lukaszewicz, P.: Energy consumption and running time for trains. Ph.D. thesis, Doctoral Thesis). Railway Technology, Department of Vehicle Engineering, Royal Institute of Technology, Stockholm (2001)
18. Lulli, A., Oneto, L., Canepa, R., Petralli, S., Anguita, D.: Large-scale railway networks train movements: a dynamic, interpretable, and robust hybrid data analytics system. In: IEEE International Conference on Data Science and Advanced Analytics (2018)
19. Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P.: Analyzing passenger train arrival delays with support vector regression. Transportation Research Part C: Emerging Technologies 56, 251–262 (2015)
20. Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N.: A fuzzy petri net model to estimate train delays. Simulation Modelling Practice and Theory 33, 144–157 (2013)
21. Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., Anguita, D.: Advanced analytics for train delay prediction systems by including exogenous weather data. In: IEEE International Conference on Data Science and Advanced Analytics (2016)
22. Regione Liguria: Weather Data of Regione Liguria. `http://www.cartografiarl.regione.liguria.it/SiraQualMeteo/script/PubAccessoDatiMeteo.asp` (2018)
23. Trabo, I., Landex, A., Nielsen, O.A., Schneider-Tilli, J.E.: Cost benchmarking of railway projects in europe-can it help to reduce costs? In: International Seminar on Railway Operations Modelling and Analysis-RailCopenhagen (2013)
24. Wang, R., Work, D.B.: Data driven approaches for passenger train delay estimation. In: Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on. pp. 535–540 (2015)