Approaches for real-time train delay prediction

Thomas Spanninger, Alessio Trivella, Francesco Corman
Institute for Transport Planning and Systems, ETH Zurich
thomas.spanninger@ivt.baug.ethz.ch

Predicting the delays of trains in real-time is an active area of research with a considerable amount of literature published in recent years. Moreover, increasing availability of historic and live train movement data creates new opportunities and challenges for future research. This paper examines and classifies approaches on train delay prediction in terms of prediction input, type of output, prediction horizon and dynamics of application. This should clear the jungle of the incredible diversity of approaches in this lively area of research and will be the initial point of a discussion of certain advantages and disadvantages of applied methods. Finally we discuss two research gaps and possible enrichments for future analysis and research.

1. Introduction

Delays of trains are unavoidable in complex railway networks, where different trains use a shared track infrastructure. Marković et al. (2015) state that train delays in the United Kingdom in 2006 and 2007 can be expressed as costs for passengers of 1 billion pounds. Despite highly optimised railway timetables, a lot of uncertainty about train punctuality remains due to stochastic factors influencing train movements. Moreover, a train running behind its schedule is likely to hinder and block other trains, which is called propagation of delay (knock-on delays).

The arrival or departure delay of a train is measured as a difference between the predefined scheduled time of the event and its realized time-stamp. Goverde (2007) shows the existence of an intuitive trade-off between infrastructure usage and vulnerability to delays of a time schedule: the more trains use the same infrastructure the more likely delays occur.

Railway operators benefit from accurate train delay predictions in many different ways. First of all, the prediction of future train delays can be communicated to passengers. Being informed about delays as early and as accurate as possible increases the service quality of railway operators for passengers significantly, even though delays will clearly not lead to passenger satisfaction. Secondly, accurate predictions of train delay development are a crucial decision support information for traffic controllers who try to minimize the propagation of delay in railway networks. Thirdly, delay predictions are a useful source of information for timetable optimization.

The prediction of train delays and punctuality in railways has been an active area of research throughout the last decades. Ghofrani et al. (2018) present a survey on big data analytic applications in railway transportation systems, including a taxonomy of train delay estimation. In contrast to their review, we focus on approaches for train delay predictions, that are applicable in real-time and make use of latest available information about live train delays.

In the following section we classify existing approaches for train delay prediction of the literature focusing on real-time applicable research. Section 3 discusses advantages and disadvantages of various approaches and propose some rules of thumbs about when and why to use which kind of train delay prediction approach. Finally, Section 4 gives and outlook for possible future research analysis in the field of real-time train delay prediction.

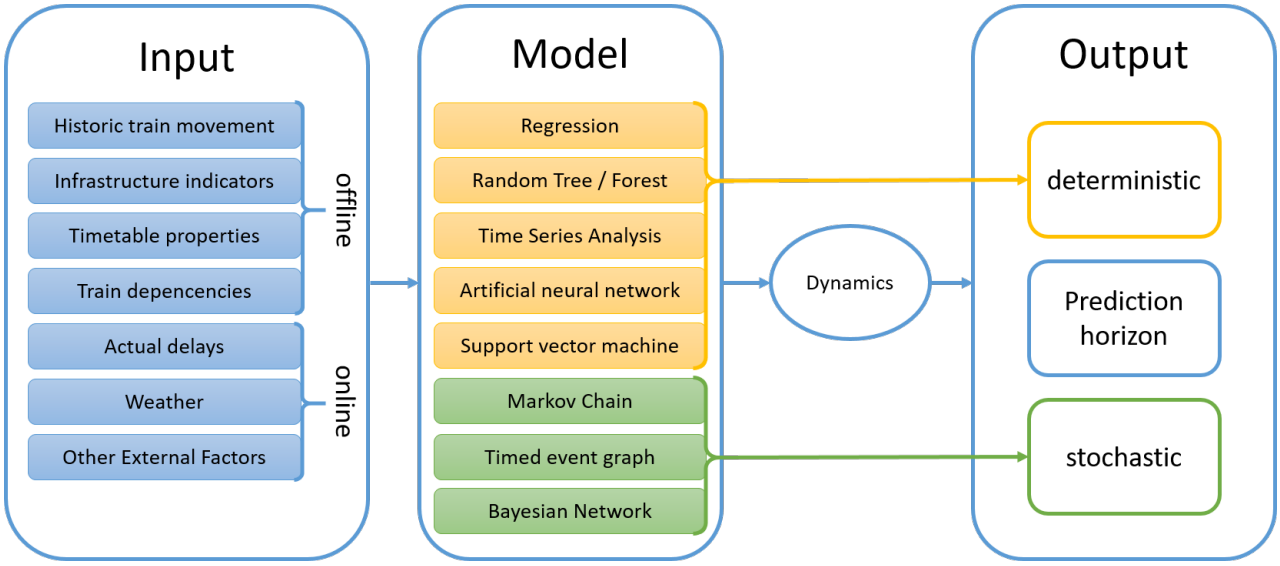
2. Classification of prediction approaches

Our classification of train delay prediction approaches shall contribute to a better understanding of challenges and complexities in the field of train delay prediction and shall serve as basis for the discussion about advantages and disadvantages of different approaches. As we focus on approaches being applicable in real-time, we do not distinguish between descriptive, prescriptive and predictive models (Karlaftis and Vlahogianni 2011). We classify approaches according to:

i. Mathematical model
ii. Input data
iii. Type of output
iv. Dynamics of application

Figure 1 visualises the criteria of classifications in context of train delay prediction proposed in this paper and lists attributes, that will be discussed in the following.

Figure 1: Illustration of the classification of train delay prediction approaches.



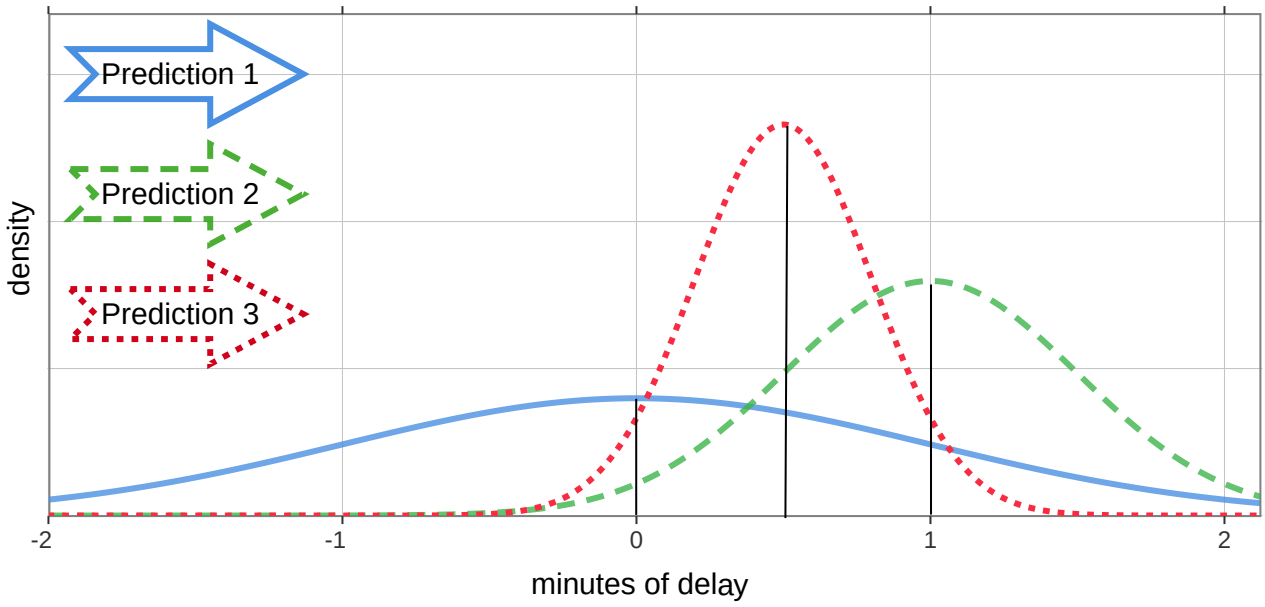
Many different mathematical models and approaches have been used in literature for the purpose of train delay prediction. (Robust) linear regression, k-nearest neighbours (k-NN) random trees (RT), random forest (RF), timed event graph (TEG), time series analysis (TSA), Markov chains (MC), artificial neural networks (ANN), support vector machines (SVM) or bayesian networks (BN) are among them.

Concerning input data for the prediction method, we want to distinguish between historic train movement (HTM), actual delays (AD), infrastructure indicators (II), timetable properties (TP) and external factors (EF). Almost all approaches take into account observations of realized historic train movements which include basic information about the rolling stock (train ID, service ID, train category) and the general setting (weekday, time of the day). As soon as approaches try to predict delay development in real-time, actual train delays (AD) of arrival, departure or passing events are an indispensable input in prediction models. Infrastructure indicators like section lengths are also commonly used inputs for predictions of train delays (Kecman and Goverde 2015a, Marković et al. 2015).

Some approaches use timetable properties like section specific catch up potential (buffer times), minimal headway times, planned connections (Goverde 2010) as explicit source of information. Weather information is also used as input factor for train delay predictions (Oneto et al. 2018).

Deterministic prediction approaches result in a single value best-estimate amount of delay for an event. *Stochastic* prediction models provide probability distributions for future events. Bayesian networks that graphically model conditional dependencies are a very demonstrative example of approaches of stochastic predictions (Corman and Kecman 2018). Figure 2 visualises the differences of dynamically updated stochastic and deterministic prediction outputs (2 updates of predictions for the same event). The deterministic predictions are represented by the vertical lines at $x = 0, 1$ and 0.5. Although the mean of the probability distribution have the same values as the deterministic predictions, the shape of the density functions provides additional information about the certainty of this future event as a result of a lower standard deviation.

Figure 2: Illustration of stochastic and deterministic predictions.



We classify train delay prediction approaches as *dynamic* if they explicitly include a foreseen process of prediction update. In contrast, we classify a prediction approach *static* if there does not exist a specific procedure of updates (one-shot predictions). Table 1 summarises papers published in the last 15 years focusing on train delay prediction based on the classifications made above.

3. Discussion

The most important difference between mathematical approaches for train delay prediction is whether they explicitly model the railway network structure. Approaches like TEGs, MCs or BNs explicitly model train dependencies whereas purely data driven approaches like linear regression, random trees and forests as well as SVMs and ANNs have their strengths in implicitly finding these dependencies in the used dataset. Marković et al. (2015) compare an ANN approach and a SVM approach and

Table 1: Summary of the literature focusing on train delay prediction approaches.

References	Mathematical Model	Used Data	Output	Dynamics
Peters et al. (2005)	ANN	HTM	deterministic	static
Meester and Muns (2007)	TEG	HTM	stochastic	static
Hansen et al. (2010)	TEG	HTM, AD	deterministic	dynamic
Goverde (2010)	TEG	AD, TP	deterministic	static
Murali et al. (2010)	Regression	HTM	deterministic	static
Berger et al. (2011)	TEG	HTM, AD, II	stochastic	dynamic
Keyhani et al. (2012)	TEG	HTM, AD	stochastic	dynamic
Büker and Seybold (2012)	TEG	HTM	stochastic	static
Yaghini et al. (2013)	ANN	HTM	deterministic	static
Milinković et al. (2013)	ANN	HTM	deterministic	static
Bauer and Schöbel (2014)	TEG	HTM	deterministic	static
Pongnumkul et al. (2014)	TSA, k-NN	HTM, AD	deterministic	dynamic
Lemnian et al. (2014)	TEG	HTM, AD	stochastic	dynamic
Kecman and Goverde (2015b)	TEG	HTM, AD	deterministic	dynamic
Kecman and Goverde (2015a)	RT, RF	HTM, II	deterministic	static
Marković et al. (2015)	SVM, ANN	HTM, II	deterministic	static
Wang and Work (2015)	TSA	HTM, AD	deterministic	dynamic
Martin (2016)	BN	HTM, AD	deterministic	dynamic
Oneto et al. (2016a)	ANN	HTM, Weather	deterministic	dynamic
Oneto et al. (2016b)	ANN	HTM, Weather	deterministic	dynamic
Oneto et al. (2017)	ANN	HTM, Weather	deterministic	dynamic
Sahin (2017)	MC	HTM	stochastic	static
Oneto et al. (2018)	ANN	HTM, Weather	deterministic	dynamic
Corman and Kecman (2018)	BN	HTM, AD	stochastic	dynamic
Lessan et al. (2019)	BN	HTM, AD	stochastic	dynamic

conclude that there exists the hazard of overfitting using purely date driven models.

Deterministic predictions clearly have the advantage of being easy to analyse a posteriori by calculating the difference of the predicted and the realised time-stamp of a certain event. On the contrary this difference cannot be calculated straight forward with stochastic predictions, since a probability distribution is assigned to the future event. Predicted probability distributions provide much deeper insights in the uncertainty of future events. Hence the key questions is whether one can make use of this additional information provided by stochastic predictions. The application of stochastic optimization for traffic control models is a perfect example of the usage of stochastic delay predictions.

If a prediction tool for train delays shall be applicable online in real-time, it needs to be as computational efficient as possible. Büker and Seybold (2012) point out that the advantage of relaxed analytical approaches is that they outperform simulation approaches when increasing the complexity of a network. Purely data driven approaches have the advantage that they can be trained offline using historical train movement data and be applied in real-time easily.

Only a few papers have analysed the prediction quality for different prediction horizons. Berger et al. (2011) provide the average difference in minutes for a prediction horizon of 30 to 240 minutes, which increases from 4 to 6.5 minutes depending on the specification of their approach. Kecman and Goverde (2015a) show that the mean absolute error (MAE) for a 0 to 20 minutes prediction horizon increases from 0 seconds to 40 seconds and stays at this level until a prediction horizon of 120 minutes. Additionally, they show that their real-time prediction tool outperforms a deterministically constant

propagation of delays (parallel shift) significantly for any prediction horizon between 0 and 8000 seconds. Oneto et al. (2018) differentiate the prediction horizon in terms of stations ahead and show that for stations 1 to 5 the average accuracy is 1.5, 1.6, 1.8, 2.1 and 2.2 minutes respectively using their ANN purely data driven dynamic prediction approach. Corman and Kecman (2018) are able to show that their BN approach provides stochastic predictions where the MAE measures on the expected value increases from 0.5 minutes to 1.4 minutes for a prediction horizon of 0 to 60 minutes.

As a result of the improvement of prediction methods in the processing of actual train delay data, the dynamics of prediction updates has become very important. Generally speaking, it is worth reproducing predictions whenever new information becomes available, which increases the quality of the prediction. As current delays are the biggest source of future delays, online prediction approaches should be repeated whenever information about a train delay/status is available. Nevertheless, also uncontrollable external factors like weather changes could result in a dynamic rule for a prediction approach.

4. Conclusion and future research possibilities

Analysing and classifying existing approaches of train delay prediction gives two interesting insights. First, there is only little research on the prediction quality along the prediction horizon. Only Kecman and Goverde (2015a) and Corman and Kecman (2018) include some analyses on that research question. There is the need to analyse which models give the best prediction quality in which prediction horizon. This analysis should take into account the computationally efficiency and lead times of possible railway traffic management actions.

There is a high amount of literature on the robustness and resilience of railway timetables. However, only few approaches of train delay prediction take into account parts of railway timetable quality measurements (e.g. buffer times allocated to running and dwell times, statistical probability of a connection to be broken and signals that significantly often imply slow driving or stopping). Goverde (2010) is one of few examples that explicitly take into account buffer times of scheduled running times to model the catch up potential of delayed trains explicitly. There is clearly the need to analyse deeper, whether explicitly taking into account different types of timetable quality measurements can further increase the prediction quality in railway systems.

References

Bauer, R. and Schöbel, A. (2014). Rules of thumb: Practical online-strategies for delay management. *Public Transport*, 6(1-2):85–105.

Berger, A., Gebhardt, A., Müller-Hannemann, M., and Ostrowski, M. (2011). Stochastic Delay Prediction in Large Train Networks. In *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems.*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Büker, T. (2010). *Ausgewählte Aspekte der Verspätungsfortpflanzung in Netzen.* PhD thesis, Rheinisch-Westfälischen Technischen Hochschule Aachen.

Büker, T. and Seybold, B. (2012). Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning & Management*, 2(1-2):34–50.

Corman, F. and Kecman, P. (2018). Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C: Emerging Technologies*, 95:599–615.

Ghofrani, F., He, Q., Goverde, R. M., and Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 90:226–246.

Goverde, R. M. (2007). Railway timetable stability analysis using max-plus system theory. *Transportation Research Part B: Methodological*, 41(2):179–201.

Goverde, R. M. (2010). A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research Part C: Emerging Technologies*, 18(3):269–287.

Greenberg, B. S., Leachman, R. C., and Wolff, R. W. (1988). Predicting Dispatching Delays on a Low Speed, Single Track Railroad. *Transportation Science*, 22(1):31–38.

Hansen, I. A., Goverde, R. M., and Van Der Meer, D. J. (2010). Online train delay recognition and running time prediction. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 1783–1788. IEEE.

Karlaftis, M. G. and Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399.

Kecman, P. and Goverde, R. M. (2015a). Online data-driven adaptive prediction of train event times. In *IEEE Transactions on Intelligent Transportation Systems*, volume 16, pages 465–474. IEEE.

Kecman, P. and Goverde, R. M. (2015b). Predictive modelling of running and dwell times in railway traffic. *Public Transport*, 7(3):295–319.

Keyhani, M. H., Schnee, M., Weihe, K., and Zorn, H. P. (2012). Reliability and delay distributions of train connections. In *12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lemnian, M., Rückert, R., Rechner, S., Blendinger, C., and Müller-Hannemann, M. (2014). Timing of train disposition: Towards early passenger rerouting in case of delays. In *14th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, Cham. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lessan, J., Fu, L., and Wen, C. (2019). A hybrid Bayesian network model for predicting delays in train operations. *Computers and Industrial Engineering*, 127:1214–1222.

Marković, N., Milinković, S., Tikhonov, K. S., and Schonfeld, P. (2015). Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C: Emerging Technologies*, 56:251–262.

Martin, L. J. (2016). Predictive Reasoning and Machine Learning for the Enhancement of Reliability in Railway Systems. In *International Conference on Reliability, Safety, and Security of Railway Systems*, pages 178–188. Springer.

Meester, L. E. and Muns, S. (2007). Stochastic delay propagation in railway networks and phase-type distributions. *Transportation Research Part B: Methodological*, 41(2):218–230.

Milinković, S., Marković, M., Vesković, S., Ivić, M., and Pavlović, N. (2013). A fuzzy Petri net model to estimate train delays. *Simulation Modelling Practice and Theory*, 33:144–157.

Murali, P., Dessouky, M., Ordóñez, F., and Palmer, K. (2010). A delay estimation technique for single and double-track railroads. *Transportation Research Part E: Logistics and Transportation Review*, 46(4):483–495.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., and Anguita, D. (2016a). Advanced analytics for train delay prediction systems by including exogenous weather data. In *IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016*, pages 458–467. IEEE.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., and Anguita, D. (2016b). Delay Prediction System for Large-ScaleRailway Networks Based on Big Data Analytics. In *Inns conference on big data*, pages 139–150. Springe, Cham.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., and Anguita, D. (2017). Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout. In *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, volume 47, pages 2754–2767.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., and Anguita, D. (2018). Train Delay Prediction Systems: A Big Data Analytics Perspective. *Big Data Research*, 11:54–64.

Peters, J., Emig, B., Jung, M., and Schmidt, S. (2005). Prediction of delays in public transportation using neural networks. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, volume 2, pages 92–97.

Pongnumkul, S., Pechprasarn, T., Kunaseth, N., and Chaipah, K. (2014). Improving arrival time prediction of Thailand's passenger trains using historical travel times. In *2014 11th Int. Joint Conf. on Computer Science and Software Engineering (JCSSE)*, pages 307–312. IEEE.

Sahin, I. (2017). Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances. *Journal of Rail Transport Planning and Management*, 7(3):101–113.

Sahin, I. (2018). A Markov Chain Model for Measuring Robustness of Train Schedules. Technical report, (No. 18-04843).

Wang, R. and Work, D. B. (2015). Data Driven Approaches for Passenger Train Delay Estimation. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 535–540. IEEE.

Yaghini, M., Khoshraftar, M. M., and Seyedabadi, M. (2013). Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation*, 47(3):667–678.