A systematic literature review of machine learning models for train delay prediction

D. White, N. Al-Moubayed

November 25, 2019

Abstract

Context: train delay prediction (TDP)

To identify a suitable paper to use as a basis for the author's own work.

Train delays impose a huge cost on both train operators and passengers. A preliminary analysis of historical delay attribution data released by Network Rail (NR) shows that 35 million minutes of delay were experienced by passengers in the 2018 - 2019 financial year. The prediction of train delays allows the rescheduling or re-routing of crews and rolling-stock, the reduction in the the amount of delay, and improvement of information available to passengers, and subsequent better decision-making.

Objective: the goal of this work is to synthesise available research results to inform evidence-based selection of a machine learning (ML) model for TPD suitable for replication by the author.

Method: relevant studies about ML techniques were gathered via a systematic literature review. Supporting contextual studies were also gathered.

Results: 19 studies were selected. 13 distinct models are used: Bayesian networks, support vector regression, random forests, neural networks, fuzzy Petri nets, extreme learning machines, various forms of regression () The data used, and results obtained, are compared.

To identify a replicable study using UK daata.

Conclusion: to do.

Contents

1 Introduction

1.1 Delays

A *delay* is a "positive deviation between the realized time and scheduled times of [an] activity" [5]. In this case, the activity is either the departure or arrival of a train. Although precise terminology differs, the literature agrees that there are two principal classes of delay [18]: primary and secondary.

1.1.1 Primary delays

A *primary* or *exogenous* delay is "caused by external stochastic disturbances" [20]. The causes of primary delays are varied and numerous [2][5][15][25] and many different classifications exist. For the purposes of this dissertation, the following classification is proposed:

- *Weather*: severe heat, flooding, landslips, leaves, snow, and ice
- *Passenger*: prolonged alighting and boarding times
- *Maintenance*: construction work, repair work
- *Other*: accidents, vandalism, trespassing, fatalities, strikes, holidays

A large component of this dissertation is the effect of the inclusion of exogenous data on the predictive capability of various machine learning techniques: this classification broadly follows the datasets that will be considered.

1.1.2 Secondary delays

A *secondary* (*knock-on*, *consecutive*) delay is "generated by operations conflicts" [5], i.e. primary delays. Secondary delays often affect both the route on which the primary delay occurred and any connecting routes; delays 'cascade' as "trains, drivers, and crews aren't in the right place at the right time to run other services" [26], or as trains are held according to waiting policies [2]. Secondary delays cannot be exactly forecast [2][15] because they are influenced by multiple interacting factors: the severity of the primary delay, the timetable of the train, infrastructure, and even the behaviour of the driver.

1.2 Timescales

TPD models are either *online*, in which they are dynamically updated with new data as it becomes available, or *offline*. Models can be further subdivided into whether they are used to predict train delays in real-time, or in the future, though this is not a perfect delineation: [17] merges these categories in an ensemble by utilising different underlying models. Real-time models are by necessity online, and make up the majority of TDP systems.

Future models can be further broken down into short-term and long-term. Long-term "tactical" [14] models are used for timetabling and resource planning. Short-term models are relatively rare, and are typically constrained by features which can only be forecast up to a limit. In TDP, this is weather, as in [17][32].

1.3 Metrics

Two key metrics are used to measure delays, and more broadly, performance: punctuality and reliability.

1.3.1 Punctuality

Punctuality is "a feature consisting in that a predefined vehicle arrives, departs, or passes at a predefined point at a predefined time" [28]. This definition has the interesting effect that trains that arrive *early* cannot be considered punctual. However, the use of punctuality hides a lot of information [30]; reliability and variability are better metrics [18]. As variability is infrequently used, it is disregarded here. The Office of Road and Rail (ORR), the UK's railway regulator, uses Public Performance Measure (PPM) to assess punctuality.

1.3.2 Reliability

Reliability has several measures [27]: typically, it is taken it taken to be the probability that a train arrives x minutes late (i.e. is punctual). ORR uses Cancellations and Significant Lateness (CaSL) to assess reliability.

2 Overview of systematic literature review method

A systematic literature review (SLR) is "a means of evaluating and interpreting all available research relevant to a particular research question or topic area or phenomenon of interest" [34]. The research questions that this SLR are intended to answer are:

- RQ1: What ML models are commonly used for TDP?
- RQ2: What exogenous data is used to improve the performance of those models?
- RQ3: What are the current research areas of TDP?

The SLR was based on a recent literature review [8], the methodology of which is discussed thoroughly below. The studies identified in [8] are used to select key search terms for further database queries. Additionally, the cited references of those studies were used to find other relevant papers. These studies are then selected for inclusion in this review by title, abstract, and finally by content on their relevance to the application of ML to TDP.

2.1 Recent applications of big data analytics in railway transportation systems: A survey [8]

Only an overview of content relevant to this literature review is presented here. The authors identified the following data-related keywords and railway transportation system (RTS) -related keywords. Of particular significance here is "Railway Operations", which covers the actual *running* of trains on a RTS, and therefore delays.

The authors limited the scope of their search to papers in scientific journals, conferences, and dissertations in English from the last 15 years (i.e. 2003 - 2017). They specifically included only papers with quantitative results. They searched ScienceDirect, Emeralds, Scopus, EBSCO, and IEEE Xplore, and also used cited references of studied papers as a source. 115 papers were found and were classified by a four-layer structure:

1. Area of RTS: Maintenance, Operations, Safety
2. Analytic category: descriptive, predictive, prescriptive
3. BDA model: clustering, numeric prediction, association, statistical analysis, image processing, and so on.
4. Implementation technique: Bayesian network, SVM, SVR, Decision Tree, ANN, Regression

We are interested in Operations; papers in the other areas are disregarded. Within Operations, the authors discuss the applications of BDA to RTS, data collection and sources in RTS, and finally the studies themselves. Only those that focus on TPD are considered. In total, 19 papers were selected by title for inclusion in this literature review; they provided the foundation for a subsequent database search.

2.2 Database search

The papers selected from [8] were used to identify the following key search terms: "train", "delay" and "prediction". Alas, "train" is a common word in scientific literature, and this confounded initial results somewhat. Appropriate synonyms were identified for each term. The following databases, based on those used by [10], were searched:

- ACM Digital Library
- IEEE Xplore
- ScienceDirect
- SpringerLink

Where possible, the discipline was restricted to Computer Science. Approximately 3000 studies were identified; some post-processing was necessary to reduce the number of studies retrieved. In total, 69 studies were selected.

2.3 Study selection

Study selection was a three-stage process:

1. Initial selection by title
2. Selection by abstract
3. Further selection by content

Of the 88 total studies found, 5 were duplicates. 64 studies were excluded by extract, and 4 by content. 4 studies were discovered through other means - either from a preliminary, less structured, search, or from cited references. In total, 17 relevant papers were identified.

3 Overview of studies

We identified 19 studies in the literature that focus on the application of ML to TPD. An preliminary analysis shows that all work occurred in the past decade (i.e. during, or after, 2010), with most studies (47%) published in 2017 and 2019.

Year	#
2010	1
2011	1
2012	1
2014	1
2015	2
2016	2
2017	4
2018	3
2019	5
Total	19

13 distinct ML techniques were identified. 29 were applied in total, as several papers compared and contrasted different techniques, or variations of the same technique (as in [12][20][15][14]), or even combined multiple models (emsemble model of [17]; in these cases, each distinct 'usage' is counted separately.

The most popular techniques are random forests (20%) and extreme learning machines (17%)., although these figures are skewed by the inclusion of four papers CITE HERE which are closely related.

Some studies use techniques that may be more accurately classified as 'statistics' rather than ML (i.e. simple variants of regression, as in [24][33]) or which are closer to *algorithmic* models than ML ([9]); however, as these lines is blurred, and for completeness' sake, all were selected for inclusion in this review.

Many defy easy classification. RFR should be just RF. Dynamic interpretable should be under Ensemble. The focus of this review was very specific

ML model	Acronym	#
Bayesian network	BN	4
Kernel method	KN	1
Extreme learning machine	ELM	5
Random forest	RF	6
Fuzzy Petri net	FPN	1
Adaptive neural fuzzy inference system	ANFIS	1
Gradient-boosted regression trees	GBRT	1
k-nearest neighbour	k-NN	1
Artificial neural network	ANN	2
Support vector regression	SVR	2
Kernel regression	KR	2
Markov		2
Decision tree	DT	1
Total		29

4 ML models

In this section, each of the distinct ML models identified previously is discussed. For studies that compare multiple models, the model the authors found to be superior is used.

4.1 Bayesian networks

A Bayesian network (BN) is a "probabilistic graphical model that uses Bayesian inference for probability computations" [29]. Each directed edge models a conditional independence, allowing "the incorporation of massive historical data" [12]. BNs allow the updating of probability distributions and reduce the uncertainty of future train delays in real-time as more data continuously comes available from the monitoring system [6].

4.1.1 A hybrid Bayesian network model for predicting delays in train operations [12]

Objectives: To identify which BN architectures are most valid / reliable for predicting train delays for a particular network structure.

Timescale: offline; short-term; long-term

Methodology: Compared three different BN schemes: heuristic hill-climbing, primitive linear, and hybrid. The data is used to rationalise the dependency graph of the BNs. Each is then trained with k-fold cross validation to avoid over-fitting and to evaluate performance. The hybrid BN is based on the structure from the former two schemes, and subsequently refined using domain knowledge and expert judgements about the sequences of stations and the relationships between consecutive train operations. This structure is intended to differentiate between the delay propagated from upstream operations and that due to the most recently performed operation in the prediction process. Continuous variables were discretised; specifically, 3 minutes was used as a width for prediction intervals (as late arrivals of less than 90s are not considered delays).

Data: Data comes from train operations on the Wuhan-Guangzhou (WH-GZ) high-speed rail (HSR) line in China. The line is 1096km long, with 18 stations. Only 15 stations and 14 sections were used due to jurisdiction. Date was collected between February 2015 to November 2015, comprising approximately 380,000 arrival and departure events between stations on the specified line, excluding early arrivals and departures.

Results: The first hybrid BN-based prediction model in the literature. The reconstructed hybrid heuristic BN was found to be perform the best out of the three, with an 80% prediction accuracy within a 60-minute horizon, a MAE of 30s, a RMSE of less than two minutes (suggesting the presence of outlier prediction errors). No-information rate of 58%. Sensitivity was $> 60\%$. Model is simple, interpretable, and computationally efficient. They also note that prediction errors soon accumulate, and that this could be addressed using an online model, as in [6].

4.1.2 Stochastic prediction of train delays in real-time using Bayesian networks [6]

Objectives: To examine the effect that the prediction horizon and incoming information about a running train may have on the predictability of subsequent arrival and departure times of all trains. To answer the question: how does the prediction of a single delay event change over time as the same event approaches the time *now*?

Timescale: online; real-time

Methodology: Developed a stochastic model for predicting the propagation of train delays based on BNs. Extended by modelling the interdependence between trains that share the same infrastructure or a scheduled passenger transfer (i.e. a connection).

Data: The model was tested on a 180km double-track mixed traffic line between Stockholm and Norrkoping, in Sweden. 90% of traffic is passenger trains. The line comprises 27 stations and junctions; 10 of these accommodate scheduled stops of passenger and freight trains. Approximately 300 trains per day traverse the corridor. Two months' worth of data, 1 January - 28 February 2015, was gathered from the Swedish infrastructure manager Trafikverket. The punctuality performance of the dataset is within 0.4% of the average performance for the whole of 2015. No strong seasonaility effect is present in the performance of the Swedish network. All event times are rounded to full minutes. Their database comprises the scheduled and realised times for departures, arrivals, and through runs for all trains and stations.

Results: accuracy of predictions is significantly increased within the 30-minute prediction horizon. Median error increases very slowly. S.D of the probability of an occurrence of an event, as predicted half an hour ahead of time, is 2 minutes. Major contribution is the representation of delays due to interactions between trains.

4.2 Neural networks

Neural networks are models based on human brains. The oldest application of ML to the general problem of predicting delays in public transport that the authors are aware of used a NN [23]. This study was not selected for inclusion in this review, however. The popularity of neural network has declined in recent years as more sophisticated models have been developed; they are primarily used as a benchmark in the studies selected for this review.

4.2.1 Railway passenger train delay prediction via neural network model [35]

Objectives: To develop a high-precision neural network model to predict the late arrival of passenger trains in Iran.

Timescale: Offline; real-time; short-term; long-term

Data: Data from 2005 to the end of 2009 is used. This comprises approximately 180,000 passenger trains, with a total delay of approximately 5.5 million minutes. Stopping time at interval stations (i.e. dwell time) for praying, boarding, and alighting are excluded from delay time.

Methodology: Compares three different input methods (normalised real number, binary coding, and binary set encoding) for a neural network model. Used decision trees and multiple logistic regression to evaluate the quality of results. Also

compared three different architectures: quick, dynamic, and multiple. Quick trains a single network; dynamic adds nodes until a specified level of performance is reached; multiple trains networks in parallel and selects the most accurate.

Results: The accuracy level of the binary quick model was found to be superior to other statistical models such as decision tree and multinomial logistic regression models, though with significantly longer training time.

4.3 Random forests

Random forests were first introduced by [11]. A random forest is a collection of individual decision trees. Simply put, each individual tree predicts the class of an input and the class with the most votes is the output of the model. This improves accuracy and reduces overfitting. Decision trees in which the target variable can take a discrete set of values are called classification trees; those that take continuous variables, regression trees. Leaves represent class labels and branches conjunctions of feature that lead to that class. Accuracy depends on the number of trees composing the forest, the accuracy of each tree, and the correlation between them [4]. Accuracy converges to a limit as the number of trees increases, and rises as the accuracy of each tree increases and the correlation between them decreases.

4.3.1 Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data [20]

Objectives: To build a data-driven train prediction system that exploits the most recent analytics tools.

Timescale: online; real-time

Methodology: The authors compared kernel methods, extreme learning machines, and random forests. The random forest comprised 500 trees. For each train and algorithm, a model is build. The hyperparameters of each model are tuned. The models are then applied to the current state of the trains, and finally they are validated (i.e. predictions are compared to actual future events). Notably, the authors derive a set of novel Key Performance Indicators (KPIs) to evaluate performance. Approximately 600,000 models would have to be trained daily across the whole network.

Data: This studie worked closely with RFI, the Italian railway authority. Used more than 6 months of data relating to two main areas in Italy, including more than 1000 trains and several checkpoints. The study also used weather data, discussed later.

Results: The RF method performed up to twice as well as the current RFI system. Included weather data increased accuracy by approximately 10%. Both ELM and KM also improve over RFI, though not to the same extent.

4.3.2 Predicting near-Term Train Schedule Performance and Delay Using Bi-Level Random Forests [16]

Objectives: To predict passenger train delays in the Netherlands.

Timescale: online, real-time

Methodology: Develops a novel bi-level RF to solve a coupled classification-regression task: identifying *whether* delays will increase, decrease, or remain constant in a fixed timeframe (20 minutes), and then predicting by how much. Assumes that delays have already occurred. Compared to linear regression, multinomial logistic regression, decision trees, k-NN, and SVM / SVRs. Assumes that all trains rode on a single track. Each leaf in the primary classification level is a random forest for regression for that classification.

Data: Data was provided from the Netherlands' railway infrastructure manager, ProRail, between September 4 2017 and December 9 2017. Includes the planned timetable, actual historical train performance, crew schedules, rolling stock circulation, (limited) infrastructure data: the distance between consecutive stations. 10 million data points over a 13 week period. Excluded Wednesdays (which are apparently particularly busy), and delays longer than 15 minutes.

Results: Found that the proposed model provides the best prediction accuracy. Constructing the model is computationally cheap.

4.3.3 Train delay analysis and prediction based on big data fusion [32]

Objectives: To develop a machine-learning model to predict

Timeframe: online; short-term

Methodology: unusual in that it focused on short-term delay prediction. Used gradient-boosted regression trees (GBRT). Model predicts delays up to 10 days in advance using weather forecasts. Used density-based clustering algorithm (DBSCAN) to identify a time interval threshold $t_epsilon$, which determines whether the delay of a train at a given station propagates to the following train.

Data: used a three-month dataset of weather, train delay, and train schedule records. The weather data is thoroughly discussed later. Data collected between 1st January to 31st March 2018. Schedule data for 7172 railway trains was obtained. GIS for 2761 railway stations was obtained from Tencent Maps (including name, longitude, and latitude. Observed nearly 2.7 million delays over the course of three months, of which 37.4% came from high-speed trains.

Results: In severe weather, train delays are determined mainly by the type of bad weather, but that in ordinary weather train delays are determined mainly by historical delay time and the delay frequency of trains. Proposes concepts of key train delay stations and the time interval threshold. Relatively poor results due to only 3 months' worth of data.

4.4 Regression

Regression is used to construct a relationship between two or more explanatory variables (independent) and a response (dependent) variable by fitting a linear equation to the data.

4.4.1 Data driven approaches for passenger train delay estimation [33]

Objectives: To develop a historical regression model to estimate future trains delays using only the past performance of a train along a given route.

Timescale: offline; online; short-term; real-time.

Methodology: developed four regression models, one offline, three online, to predict train delays. Assumes that delays from one trip to the next follow a vector autoregressive process. Assumption is valid because passenger trains operate on a fixed frequency (daily) and schedule, and so prior delays on previous trips hold information to estimate the train delay at each station for the current trip. The vector autoregressive process predicts train delays at each station along the route simultaneously. Determines parameters of regression model through least squares estimation on the training dataset.

Data: Data is from 282 Amtrak trains from 2006 – 2013, more than 100,000 train trips. Amtrak trains have operational priority yet the on-time rate of Amtrak is less than 50%. Average delay for several trains can reach as high as 50 minutes. Coarse; lots of missing records.

Results: Historical regression models improves the RMSE estimate of delay by 12%. The online proposed model improves the RMSE estimate of delay by 60%.

4.5 Support vector regression

Support vector regression (SVR) is a popular ML model that uses support vector machines (SVMs) to predict continuous values.

4.5.1 Prediction of arrival times of freight traffic on US railroads using support vector regression [1]

Objectives: To develop a data-driven approach to predict estimated arrival times (ETAs) of individual freights trains.

Timescale: Online; real-time

Methodology: Construct distinct regression model for each origin-destination pair for which predictions are required. All of the same form; differ only in feature weights and hyper-parameters. Test four different SVR-based algorithms: three linear, with varying input features, and one RBF (radial basis function) kernel SVR. RBF kernel offers no improvement over fully-featured linear kernel. Only needed to train 140 models. Estimate 10,000 models for all ETA predictions.

Data: Used freight train movement, train car operations, crew, and locomotive data. Network data is extracted from dispatching, operations, and signalling data. Final features: train length, train tonnage, train horsepower per ton, train priority, crew time remaining, on duty time to departure, full traffic count, directional traffic count, available sidings.

Results: Large gains from including track segment occupancy features. Results compare closer to a deep NN trained on the same dataset (average improvement of 16%, max of 25%) RBF kernel SVR offers a mean 14.3% improvement, a max 21.8%. Don't report on prediction error. Notes the difficulty of using features that change such the amount of traffic on the line of the road, the number of available sidings (as trains enter and leave).

4.5.2 Analyzing passenger train arrival delays with support vector regression [14]

Objectives: Develop a model to capture the relation between passenger train arrival delays and various characteristics of a railway system, specifically infrastructure.

Timescale: offline; short-term; long-term

Methodology: First application of an SVR to TDP. Compared an ANN and CVR. Categorical variables were converted to binary variables, trained using Levenberg-Marquardt backprop. 100 independent ANNs were trained and the outputs averaged. Considers seven variables: passenger train category, scheduled time of arrival at station, infrastructure influence defined by expert opinions (discussed later), percent of journey completed distance-wise, distance travelled, time travelled, and headway.

Data: 727 passenger trains (99 long-distance, 321 regional, 307 suburban, northbound towards Belgrade. Delays recorded on a minute scale.

Results: Found that scheduled time of arrival and headway are not strongly correlated with any other co-variates. SVR performed better than the ANN.

4.6 Fuzzy Petri nets

Fuzzy Petri nets (FPNs) are "modifications of classical Petri nets for dealing with imprecise, vague, or fuzzy information in knowledge based systems" [13]. A Petri net is simply a directed bipartite graph, in which nodes represents transitions such as events (as bars) and places (as circles) and edges represent which places are pre- or post-conditions for transitions. PNs can be easily understood by graphical presentation.

4.6.1 A fuzzy Petri net model to estimate train delays [15]

Objectives: To develop a FPN model for estimating train delays.

Timescale: Offline; short-term; long-term

Methodology: construct an FPN using expert knowledge to define fuzzy sets and rules. Use historical data to train an Adaptive Network Fuzzy Inference System (ANFIS). An ANFIS combines the learning ability of an ANN and the capacity of fuzzy logic to interpret imprecise data. Import this ANFIS into an FPN and test on part of the Belgrade railway network.

Data: Used data on train delays for July 2010 from part of the Belgrade railway network. 3710 trains: international passenger, domestic passenger, suburban and regional passenger, international freight, and so on. Features are train category, time of arrival at the station, the distance travelled, and the infrastructure influence (explored later).

Results: Sub-5% accuracy for two stations, but over 10% for two others, due to the lower number of trains and many delay outliers. Anticipates the applicability of the model to infrastructure investment decisions.

4.7 Ensemble methods

Ensembles use multiple models to generate predictions. The rationale behind such ensembles is simple: using a diverse set of models reduces bias and error rates.

4.7.1 An ensemble prediction model for train delays [17]

Objectives: To develop a large-scale, data-driven ensemble forecasting system for train delays

Timescale: Online; real-time; short-term; long-term

Data: Used 3.25 years of data collected from Deutsche Bahn. Incorporated a wide range of exogenous data. Roughly 350 features for operational (i.e. running) trains, and 70 for non-operational trains, including train properties, real-time train state, network-related such as track and platform occupation conflicts, connections, and external features.

Methodology: began with extensive consultations with SMEs to arrive at a set of factors with explanatory factors. Started on 2 performance metrics; ended up with 80. Evaluated SVRs as a potential candidate for inclusion in their EM. However, they were found not to provide the best accuracy, quadratic in training data volume, and very sensitive to hyper-parameters, so selected RFs instead for the accuracy of forecasts and the possibility of incremental (updating model parameters as fresh data becomes available) and parallel training.

Used 3 models for operational trains (those currently running): a RF (n-stop ahead), mesoscopic simulation and kernel regression, and a static RF for non-operational trains, and mesoscopic simulation. For kernel regression, a reference catalog of movements for each train is stored. A forecast is then generated by weighted sum, with weights computed by measuring the similarity between the train of interest and the weighted set. Only used for operational trains. A threshold-based heuristic was used to identify trains that ought to be stored. The simulation model was based on [31]. The model was run every minute, initialised using the most recent train status message.

Results: Found greatest potential for improvements is in shorter-term forecasts of operational trains. Initially scoped to do two-day ahead forecasts, but found that predictions beyond 24 hours were only marginally better than the schedule. Found an interesting edge case where trains with long delays did not report positions.

4.7.2 A dynamic, interpretable, and robust hybrid data analytics system for train movements in large-scale railway networks [19]

Objectives: To develop the first hybrid model (HM) in the literature. Such a model is dynamic, interpretable, and robust.

Timescale: Online; real-time

Metholodogy: combined a experience-based model (EBM) and a RF into a hybrid model (HM), to overcome the limitations of both. The EBM is based on knowledge of the network and the experience of the operators. Introduces the idea of a penalty cost, whereby the cost of delay is variable dependent on train type, location, type of railway section, amount of delay. Predict running time, dwell time, penalty cost, and train overtaking, and construct delay from these variables. The HM is a DT

where each leaf is a RF. Trains are directed to the appropriate RF by similarity. A new leaf is added each time a new train movement is added that belongs to a previously unexplored branch of the decision tree. The RF regressor in the leaf is trained based on all the past train movements that fall in that leaf. The HM forget movements older than 3 months, based on experience of operators and different window size. Prediction is just consulting the appropriate leaf. The whole HM is constructed and updated incrementally as new train movements become available.

Data: This study worked closely with RFI, the Italian railway authority. Used more than 6 months of data relating to two main areas in Italy, including more than 1000 trains and several checkpoints. The study also used weather data, discussed later.

Results: Found the HM to provide the best trade-off between accuracy and computational requirements. HM clearly outperforms the EBM and the DDM, both of which are state-of-the-art. Most evident for freight and regional trains. Constantly better across the whole year. Reaches optimal accuracy after 10 days of results. Handles infrequent events well.

4.8 Extreme learning machines

ELMs are feedforward neural networks. They were introduced to overcome problems posed by backprop, namely "slow convergence rates, critical tuning of optimisation parameters, and presence of local minima necessitating multi-start and re-training strategies" [22].

4.8.1 Train Delay Prediction Systems: A Big Data Analytics Perspective [22]

Objectives: to build a data-driven TDP system for large-scale railway networks that exploits the most recent big data technologies, learning algorithms, and statistical tools.

Timescale: online; real-time

Methodology: Used Apache Spark in-memory technology. Explores the application of a both a shallow ELM (SELM) and a deep ELM (DELM). Uses 10-Fold Cross Validation to tune hyperparameters. For every train, DELM and SELM models were built, tuned. applied, and finally validated based on actual future data.

Data: 6 months, from January 2016 to June 2016, of train movements records from the entire Italian railway network.

Results: DELM is up to 200% more performant than the current system in use. SELM all shows improvements. but not to the same extent. For DELM, a deep architecture with a small number of neurons in each layer is preferred.

4.8.2 Dynamic Delay Predictions for Large-Scale Railway Networks: Deep and Shallow Extreme Learning Machines Tuned via Thresholdout [21]

Objectives: To develop a dynamic data-driven TDP system that allows exploiting both historical data about train movements and exogenous data about the weather.

Timescale: Online; real-time

Methodology: Used Apache Spark in-memory technology. Explores the application of a both a shallow ELM (SELM) and a deep ELM (DELM). For every train, DELM and SELM models were built, tuned. applied, and finally validated based on actual future data. Uses the thresholdout procedure to optimise the hyperparameters of SELM and DELM.

Data: 6 months, from January 2016 to June 2016, of train movements records from the entire Italian railway network, and weather data covering the region.

Results: DELM tuned via thresholdout is the best performing model, with up to 2x better accuracy that then current RFI model, and superior performance to DELM tuned via the standard hold out method (cross validation).

4.9 Markov models

A Markov chain is a stochastic model that describes a sequence of possible events. The probability of each event depends only on the state attained in the previous event. Such models, then, are likely too simplistic for predicting knock-on delays

4.9.1 Estimating Train Delays in a Large Rail Network using a Zero Shot Markov Model [7]

Objectives: To develop a scalable, train-agnostic, and Zero-Shot competent framework for predicting train arrival delays

Timescale: Offline; short-term; long-term

Methodology: studies systemic delays in train arrivals using n-order Markov frameworks. Used RF regressors and ridge regressors. Avoided building train-specific models for real-time deployment and scalability. RF had 231 trees.

Data: Uses train operations data for the past two years. Only a small dataset: 135 trains between March 2016 and February 2018, at single station in India.

Results: RF regression was found to be superior than MLR: approximately 80% of instances had a absolute error or < 1 minute. Using just 1.2% of trains in India, was able to covert more than 11.3% of stations

5 Exogenous data

It is widely accepted in ML that the greater the quantity of information available for the creation of a model, the greater the performance of that model will be. Features can either be *engineered* from existing data or incorporated from *exogenous* data. Data is exogenous if it is independent of other input data but the output data is dependent on it. The scope for inclusion is essentially limitless: any source of data which may affect railway operations is a viable candidate. This section explores the use of exogenous data in the selected studies according to the classification defined previously: weather, infrastructure, maintenance, and "other".

5.1 Weather

Weather data is the most popular candidate for inclusion. It was first used in a TDP model, to the best of the authors' knowledge, in [20]. It is expected that weather-induced delays are seasonal, with heat causing delays in summer, rain and leaves causing delays in autumn, and snow and ice causing delays in winter: [3] found that most weather-caused delays occurred in the last third of the year, with a peak in November. It is therefore important that models be trained with at least a year's worth of data: [32] partly attribute the poor performance of their model to the short duration (1 January 2019 - 31 March 2019) of data collection.

There is also, naturally, a geographical element to the influence of weather on train delays: more temperate climes are likely to have less extreme weather, and so fewer delays due to that weather. [32] observed that in locations less prepared for specific types of severe weather - such as snowy weather in southern cities - delays were greater. They found that in severe weather train delays are determined mainly by the type of bad weather. [20] note that weather conditions can additionally influence passenger flow and consequently dwell times, which have already been described as a key influence on delays. Typically, each location from which train data is reported is mapped to a geographically close weather observation station.

Resolution is also important. Hourly data, as might be expected, improves the performance of a model [16][32].

The fields used are fairly heterogeneous: maximum, minimum, and average temperature; some weather type category (e.g. cloudy; overcast; thunderstorm); wind speed; wind direction; and precipitation. More unusual fields include pressure [20][21][19], air quality [32].

[20] found that the inclusion of weather data improved the accuracy of their RF model by approximately 10%, with the caveat that the further ahead in the future the forecast is (and thus the less accurate), the smaller this increase was. However, [17] found that weather has only a small impact on delays; an analysis of delay attribution data showed that less than 3% of delays were directly attributed to weather.

5.2 Infrastructure

Infrastructure is the second most popular candidate for inclusion. It is a thoroughly modelled in the analytical models for TDP.

Of the four papers to include infrastructure data, two use expert opinions to assign a score to track sections [15] [14] and two use the data directly [1] [17].

In the former two, opinions are collected from traffic dispatchers, operators, and subject matter experts. Each line or section considered is assigned a score based on characteristics such as the number of tracks, the percentage of rail on which trains must travel at reduced speeds, the number of stations, stops, loops, level crossings, junctions, length, block section, track clear section (station distance, braking distance, automatic block system, centralised traffic control, axle counters, and so on.

The Delphi method is used to produce a single output score. The obvious limitation of this method is that each section must be independently human-assessed; in rail networks thousands of kilometres in length, this is clearly impractical. In [14], only 39 lines were evaluated. A score of 1 in denotes a route with the highest number of infrastructural factors that could cause unplanned delays. The study found a strong correlation between expert opinions and train delays.

[1] takes into account individual tracks, switches, mileposts, and terminals, in order to build a network graph describing single-track edges, multi-track edges, and single-track edges with sidings.

single-tracking, reduced speeds, characteristics of block and interlocking systems, number of stations, stops, loops, road-rail level crossings, and junctions

[15] groups infrastructure opinions. Collected opinions from traffic dispatchers, operators, and experts famiilar with the functioning of the system. Was used more broadly to define input variables, and the primary causes of delay (not the causes of primary delay). Defined three input parameters: the train category, timetable influence, and the distance travelled by the train. Timetable influence was used as a catch-all of sorts; the study is vague on specifics. It included the influence of infrastructure parameters, timetable characteristics, operation time, the type of locomotive, local conditions, technological solutions, principles for safety and signalling, and weather conditions. This is for the FPN!

For the ANFIS, which used real-life data (go into detail here), an 'infrastructure influence', which included the percentage of restricted speed sections, the number of junctions, and the number of stations). Included section length, section plans, restricted speed, and track routes.

[17] take exactly this approach. The authors reconstruct the network and estimate capacity directly from passing messages. Furthermore, the method used generated train-class specific networks. The inferred method is employed for various downstream tasks: inferring train paths, conflict status estimation, typical travel time estimation. "Passing" messages are sorted by date, time, and train. If there are sufficient observations, the control point and track stretch is recorded as an edge. The frequency of transitions from each outgoing edge, the mean and standard deviation of travel times are also recorded for each edge. A feasibility matrix for each outgoing edge is recorded at each vertex, which records pairwise edge feasible flows at each section by identifying movements by two trains in a short time window; this is used to identify potential conflicts between trains when there are deviations from the schedule. Reconstructed networks around several major hubs were inspected by hand and found to be accurate.

Station attributes used included the designated platform, station attributes, historical mean delay at tracks, platforms, actual platform, track allocation, and track / platform change status.

5.3 Maintenance

[17] used work zone information, indicating location, duration, and the likely impact on different train categories.

5.4 Other

No papers were found to incorporate accidents, vandalism, trespassing, fatalities, or strikes. Holidays are, however, included in [17][21].

6 Evaluation

6.1 RQ1: What ML models are commonly used for TDP?

A wide variety of ML models have been applied for TDP. Interest in the area has only really developed in the past decade, and as ML theory has developed so quickly in this time, many such models have become outdated even in this short timeframe. Neural networks make only a brief appearance in [35], though they are often used as a benchmark to compare more sophisticated models against. Popular models such as support vector regression (SVR) and support vector machines (SVM) are applied in [14] [1]. Markov models are also used, but only in limited contexts, in [7].

More esoteric ML models, such as fuzzy Petri nets, are also used in [15], although without further development.

Currently at the forefront of research are three models: random forests [20][16][17], Bayesian networks [12][6] and extreme learning machines [22][21]. Increasingly, these models are combined into an ensemble or hybrid model, as in [17] (random forests, kernel regression, and mesoscopic simulation) [19] (random forests, experienced-based model).

For the authors' own work, the methodology and model structure described in [17] is deemed most suitable for replication, with scope limited to the real-time: the approach is state-of-the-art, thoroughly described, and uses data very similar to that available to that of the author. Additionally, their models are entirely data-driven, and so do not require a close working relationship with a railway manager to define the rules of an experience-based model as used in [19].

6.2 RQ2: What exogenous data is used to improve the performance of those models?

A four-way classification was defined early for exogenous data: weather, infrastructure, maintenance, and 'other'.

Weather was first used in [20], and has subsequently been investigated in [3] and applied in [32][17][16]. [20] found 10% performance increase when incorporating weather data into their RF model. However, [17] cast doubt on the effect of weather: only 3% of delays in their dataset were directly attributable to weather. The authors plan to do their own statistical analysis of delay attribution records.

Infrastructure is the next most popular. [15] and [14] both use the expert opinions of dispatchers to construct a variable describing the effect of various infrastructure features along a rail line (single-tracking, reduced speeds, characteristics of block and interlocking systems, number of stations, stops, loops, road-rail level crossings, and junctions, etc.) on the likelihood of the delays on that line. Both (unsurprisingly) find a strong correlation between this score and severity of delays.

[17] take a more programmatic approach and reverse engineer characteristics such as capacity and network structure directly from their dataset.

The authors think that infrastructure has been somewhat neglected. Perhaps, however, describing it as "exogenous" is disingenuous here: many models use some descriptor of the rail network as an input, or at least characteristic.

The authors will use the approach of [17] or an existing data-set (if they can finally fucking parse it)!

Maintenance is used only by [17], who quite rightfully claim to have constructed the most comprehensive dataset yet for TDP. The authors will also use historical data of maintenance work.

As we continue, the factors becomes less important, and perhaps less predictable. "Other" is very much a catch-all. It contains events that likely cannot be predicted (e.g. accidents, vandalism, trespassing, fatalities) due to insufficient data and fairly menial features: holidays and the like.

Of these categories, the authors will include holidays and strikes. NR categorises the day by weekday, Saturday, Sunday, Christmas, and Bank holiday, reflecting the different timetables used for each.

6.3 RQ3: What are the current research areas of TDP?

This has perhaps been the hardest RQ to answer. Only one data point is available. [19] note that, since their earlier paper [20], RFI, the Italian railway manager, has adopted their model and a experience-based models similar to that described in [9]. It seems clear the the future of TDP is ML, not analytical.

It is also noted in [20] that RFI use a system similar to [9], which lends credence to their suggestion that most extant TDP systems are analytical.

Incorporating delay attribution into the model as part of the issue. Currently there's an extensive and complicated model.

Noted also by [17]: that different categories of trains need different models. Incorporating delay attribution back into it.

[19], in their hybrid model, have essentially a RF for categories of trains.

7 Conclusion

We performed a systematic literature review for applications of machine learning in train delay prediction. A wide variety of studies and models have been discussed. We have thoroughly investigated exogenous data used to improve the performance of these models, and have identified a paper suitable for replication with the data available to the authors.

References

[1] William Barbour et al. "Prediction of arrival times of freight traffic on US railroads using support vector regression". In: *Transportation Research Part C: Emerging Technologies* 93 (2018), pp. 211–227. DOI: 10.1016/j.trc.2018.05.019.

[2] Annabell Berger et al. "Stochastic Delay Prediction in Large Train Networks". In: *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*. Ed. by Alberto Caprara and Spyros Kontogiannis. Vol. 20. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2011, pp. 100–111. ISBN: 978-3-939897-33-0. DOI: 10.4230/OASIcs.ATMOS.2011.100. URL: http://drops.dagstuhl.de/opus/volltexte/2011/3270.

[3] William Brazil et al. "Weather and rail delays: Analysis of metropolitan rail in Dublin". In: *Journal of Transport Geography* 59 (2017), pp. 69–76. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2017.01.008. URL: http://www.sciencedirect.com/science/article/pii/S0966692316304409.

[4] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324.

[5] Fabrizio Cerreto et al. "Causal Analysis of Railway Running Delays". In: *11th World Congress on Railway Research (WCRR 2016)*. 2016.

[6] Francesco Corman and Pavle Kecman. "Stochastic prediction of train delays in real-time using Bayesian networks". In: *Transportation Research Part C: Emerging Technologies* 95 (2018), pp. 599–615. DOI: 10.1016/j.trc.2018.08.003.

[7] Ramashish Gaurav and Biplav Srivastava. "Estimating Train Delays in a Large Rail Network Using a Zero Shot Markov Model". In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (2018). DOI: 10.1109/itsc.2018.8570014.

[8] Faeze Ghofrani et al. "Recent applications of big data analytics in railway transportation systems: A survey". In: *Transportation Research Part C: Emerging Technologies* 90 (2018), pp. 226–246. DOI: 10.1016/j.trc.2018.03.010.

[9] Ingo A. Hansen, Rob M.P. Goverde, and Dirk J. van der Meer. "Online train delay recognition and running time prediction". In: *13th International IEEE Conference on Intelligent Transportation Systems* (2010). DOI: 10.1109/itsc.2010.5625081.

[10] Sarah Heckman and Laurie Williams. "A systematic literature review of actionable alert identification techniques for automated static code analysis". In: *Information and Software Technology* 53.4 (2011), pp. 363–387. DOI: 10.1016/j.infsof.2010.12.007.

[11] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE, 1995.

[12] Javad Lessan, Liping Fu, and Chao Wen. "A hybrid Bayesian network model for predicting delays in train operations". In: *Computers & Industrial Engineering* 127 (2019), pp. 1214–1222. DOI: 10.1016/j.cie.2018.03.017.

[13] Hu-Chen Liu et al. "Fuzzy Petri nets for knowledge representation and reasoning: A literature review". In: *Engineering Applications of Artificial Intelligence* 60 (2017), pp. 45–56. DOI: 10.1016/j.engappai.2017.01.012.

[14] Nikola Marković et al. "Analyzing passenger train arrival delays with support vector regression". In: *Transportation Research Part C: Emerging Technologies* 56 (2015), pp. 251–262. DOI: 10.1016/j.trc.2015.04.004.

[15] Sanjin Milinković et al. "A fuzzy Petri net model to estimate train delays". In: *Simulation Modelling Practice and Theory* 33 (2013), pp. 144–157. DOI: 10.1016/j.simpat.2012.12.005.

[16] Mohammad Amin Nabian, Negin Alemazkoor, and Hadi Meidani. "Predicting Near-Term Train Schedule Performance and Delay Using Bi-Level Random Forests". In: *Transportation Research Record: Journal of the Transportation Research Board* 2673.5 (2019), pp. 564–573. DOI: 10.1177/0361198119840339. URL: https://journals.sagepub.com/doi/full/10.1177/0361198119840339?casa_token=DFAg4bqXSC0AAAAA%3AE_fjvjujqVan3ZdY5rmfeVAUl-q5Qsr7ExDIW3fhcEKTQqGg_qae5mredq8XJyeR5plFdL8WcwAXuw.

[17] Rahul Nair et al. "An ensemble prediction model for train delays". In: *Transportation Research Part C: Emerging Technologies* 104 (2019), pp. 196–209. DOI: 10.1016/j.trc.2019.04.026.

[18] Nils O.E. Olsson and Hans Haugland. "Influencing factors on train punctuality—results from some Norwegian studies". In: *Transport Policy* 11.4 (2004), pp. 387–397. DOI: 10.1016/j.tranpol.2004.07.001.

[19] Luca Oneto et al. "A dynamic, interpretable, and robust hybrid data analytics system for train movements in large-scale railway networks". In: *International Journal of Data Science and Analytics* (2019). DOI: 10.1007/s41060-018-00171-z.

[20] Luca Oneto et al. "Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data". In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2016). DOI: 10.1109/dsaa.2016.57.

[21] Luca Oneto et al. "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* 47.10 (2017), pp. 2754–2767. DOI: 10.1109/tsmc.2017.2693209.

[22] Luca Oneto et al. "Train Delay Prediction Systems: A Big Data Analytics Perspective". In: *Big Data Research* 11 (2018), pp. 54–64. DOI: 10.1016/j.bdr.2017.05.002.

[23] J. Peters et al. "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)* (2005). DOI: 10.1109/cimca.2005.1631451.

[24] Suporn Pongnumkul et al. "Improving arrival time prediction of Thailand's passenger trains using historical travel times". In: *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (2014). DOI: 10.1109/jcsse.2014.6841886.

[25] Network Rail. *Delays explained*. 2019. URL: https://www.networkrail.co.uk/running-the-railway/looking-after-the-railway/delays-explained/ (visited on 11/19/2019).

[26] Network Rail. *Knock-on delays*. 2019. URL: https://www.networkrail.co.uk/running-the-railway/looking-after-the-railway/delays-explained/knock-on-delays/ (visited on 11/19/2019).

[27] P Rietveld, F.R Bruinsma, and D.J van Vuuren. "Coping with unreliability in public transport chains: A case study for Netherlands". In: *Transportation Research Part A: Policy and Practice* 35.6 (2001), pp. 539–559. DOI: 10.1016/s0965-8564(00)00006-9.

[28] Andrzej Rudnicki. "Measures of Regularity and Punctuality in Public Transport Operation". In: *IFAC Proceedings Volumes* 30.8 (1997), pp. 661–666. DOI: 10.1016/s1474-6670(17)43896-1.

[29] Towards Data Sciencel. *Introduction to Bayesian networks*. 2019. URL: https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e (visited on 11/19/2019).

[30] R Skagestad. "Kritiske prestasjonsindikatorer i jernbanedrift". PhD thesis. Norwegian University of Science and Technology, 2004.

[31] Jácint Szabó, Sebastien Blandin, and Charles Brett. "Data-Driven Simulation and Optimization for Incident Response in Urban Railway Networks". In: *AAMAS '17 Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 819–827.

[32] Pu Wang and Qing-peng Zhang. "Train delay analysis and prediction based on big data fusion". In: *Transportation Safety and Environment* 1.1 (2019), pp. 79–88. DOI: 10.1093/tse/tdy001.

[33] Ren Wang and Daniel B. Work. "Data Driven Approaches for Passenger Train Delay Estimation". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (2015). DOI: 10.1109/itsc.2015.94.

[34] C.C. Williams and J.K. Hollingsworth. "Automatic mining of source code repositories to improve bug finding techniques". In: *IEEE Transactions on Software Engineering* 31.6 (2005), pp. 466–480. DOI: 10.1109/tse.2005.63.

[35] Masoud Yaghini, Mohammad M. Khoshraftar, and Masoud Seyedabadi. "Railway passenger train delay prediction via neural network model". In: *Journal of Advanced Transportation* 47.3 (2013), pp. 355–368. DOI: 10.1002/atr.193.