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

D. White, N. Al-Moubayed

November 21, 2019

Abstract

Context: train delay prediction (TDP)

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* may be defined as "the difference between the minimum, or unopposed, travel time, and the actual travel time or [as] the difference between the scheduled running time and actual running time" [6].

A *delay* is a "positive deviation between the realized time and scheduled times of [an] activity" [3].

They are reduced by the provision of *slack* in railway timetables [3]. Slack is the amount of time a train can delayed without the delay propagating.

Mention relationship between slack, capacity, and so on and so forth.

Although precise terminology differs, the literature agrees that there are two principal classes of delay [16]: primary and secondary.

A *primary* or *exogenous* delay is "caused by external stochastic disturbances" [17]. The causes of primary delays are varied and numerous [1], [14], [22], [3], 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, large flows
- *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 sets of data that will be considered.

The relative importance of these factors is poorly studied. A limited study by [16] found that the punctuality of trains is correlated with the number of passengers, occupancy ratio, departure punctuality, and operational priority rules.

The number of passengers affects the *dwell time*, the time "devoted to the loading and unloading processes of the train" [26]. For this dissertation, this refers to the alighting and boarding of passengers. Dwell time is a key parameter of system performance, service reliability and quality [21]. Passenger volume is considered a key factor influencing both dwell time [26] and punctuality [16].

A *secondary* (*knock-on*, *consecutive*) delay is "generated by operations conflicts" [3], 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" [23], or o trains are held according to waiting policies between trains [1].

Although further classifications of secondary delay exist [5], the current level of detail will suffice for this dissertation.

Secondary delays cannot be exactly forecast [1], [14] because they are influenced by multiple interacting factors: the severity of the primary delay, the timetable of the train, the infrastructure, and even the behaviour of the driver, who may drive faster than planned, or reduce dwell time at stations, in order to make up time.

The difficulty of doing so cannot be overstated. Predicting the occurence of a primary delay is easy in comparison.

The goal of this work is to synthesise available research results to inform evidence-based selection of ML techniques.

Siding: a short section of double track on a single track that allowing trains to meet or pass each other [**barbour˙et˙al˙2019**]. A meet or a pass is referred to as a *movement*; movements are directed by human dispatchers.

1.2 Timescales

There are several different timescales at which delays can be predicted, such as short-term (predicted using real-time operating data) and long-time (3 days to a week in advance), as in [28]; tactical (in which models are applied to both timetabling and resource planning), and operational (in which models are used for the real-time prediction of train delays), as in [13]. This dissertation is concerned with the short-term / operational level, henceforth referred to as the *real-time* level.

[**nair˙et˙al˙2019**] merge the real-time (operational) and short-term long-term (non-operational) trains in their ensemble model by utilising different underlying models.

Models for real-time traffic have so far focused on overcoming the combinatorial complexity of train rescheduling, rolling stock and crew scheduling, and delay management [11]. Real-time train delay prediction (RTTDP) models are *online*, i.e. updated as data on train movements becomes periodically available. Many different models have been proposed; they will be discussed later.

For scheduled train services, a trade-off exists between efficiently utilising the capacity of a railway network and improving the reliability and punctuality of train operations. Used 3.25 years of data.

Extensive consultations with SMEs to arrive a set of factors with explanatory power. Started on 2 performance metrics; ended up on 80. Roughly 350 features for operational trains, and 70 for non-operational trains. Five types: train specific (train properties and real-time train state. Information directly influences the delay of each train in the network. Infrastructure: static and RT information regarding statiosn, platofmrs, and tracks. Conveys infrastructure properties such as how busy a station is and how frequently delay happens at a given track. Network-related: a class of features related to delays across the network. Important subset is track and platform occupation conflicts. Feature generation routine aims to determine the likelihood of such downstream conflicts along tracks and platforms by considering arrival times within narrow time windows. Connection: delay may be caused by connecting trains. Determined from data using a rule-based approach. Exteranl features: calendar features, weather information, trainworks, maintenance, holidays.

Trained two families of RF. First for operational trains: a n-stop model $n = 1, 2, ...10$. Limited to 15 trees per model; models for departure and arrival time prediction are trained separately. Increasing had no impact of quality. For static / non-operational trains: one for departure, one for arrival. 22 RF for both. Can be intepretated.

Kernel regression: store a reference catalog of movements for each train. Forecast is then generated by a weighted sum of the reference catalog. Weights are computed by measuring the similarity between the train of interest and the weighted set. Only used for operational trains. Tested for non-operational trains, but it performed poorly.

Catalog for thousands of services. Caching service used (threshold-based heuristic). Discard forecasts generated by catalogs with fewer than 100 trajectories.

And a simulation model! Based on Szabo et al (2017). Run every minute, initialised by setting the train position to the most recent status message received in near real-time. Could not integrate data on delays caused by prolonged boarding and egress processes. Connections are accounted for by estimating a connection matrix of all possible pairs of trains at each station. If deemed significant, waiting time threshold is estimated.

7400 MR models were needed daily.

Edge case when trains with long delays did not report position, as stuck between control points. Results suggest that greatest potential for improvements is in shorter-term forecasts of operatiaonl trains using DDM.

1.3 Metrics

Punctuality is "a feature consisting in that a predefined vehicle arrives, departs, or passes at a predefined point at a predefined time" [25]. 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 [27]; reliability and variability are better metrics [16].

reliability has several measures [24]:

- The probability that a train arrives x minutes late (punctuality)
- The probability of an early departure
- The mean difference between the expected arrival time and the scheduled arrival time
- The mean delay of an arrival given that one arrives late
- The mean delay of an arrival given that one arrives more than x minutes late
- The standard deviation of arrival times

variability is a "measurement of the uncertainty of trip journey times in transportation" [16]. It relates to the distribution of arrival times for a train [15]: a train that arrives the same amount of minutes behind schedule every day has low variability, but not would be considered punctual.

The Office of Road and Rail (ORR), the UK's railway regulator, uses Public Performance Measure (PPM) to assess punctuality, and Cancellations and Significant Lateness (CaSL) to assess reliability.

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" [30]. The specific objectives of this SLR are:

- To identify categories of ML techniques
- To summarise current research solutions for TDP
- To synthesis the current results from ML techniques for TDP
- To identify the research challenges and needs in the area of TDP

2 Research questions

We are interested in answering the following research questions:

- RQ1: What ML models are commonly used for TDP?
- RQ2: What exogenous data is used to improve the performance of those models? / big data analysis
- RQ3: How has ML been used in date in the area of TDP?

3 Overview of literature review method

A full systematic literature review is beyond the scope of this paper. Instead, a recent literature review focusing on big data analytics (BDA) in railway transportation systems (RTS) [9] was used as a basis for the authors' own. The methodology of [9] is therefore thoroughly discussed. The studies identified in [9] 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. The studies found in [9], by database search, and from cited references, are then selected for inclusion in this review by title, abstract, and finally by content. Although the search was by no means exhaustive, the authors are satisfied that the papers gathered represent a comprehensive review of the application of ML to TDP.

3.1 Recent applications of big data analytics in railway transportation systems: A survey [9]

Readers are invited to read the paper themselves; only an overview of content relevant to this literature review is presented here. The authors identified the following data-related keywords: ”Data analytics, Big data, Data mining, Machine learning, Descriptive analytics, Predictive analytics” and the following RTS-related keywords: ”Rail, Railway Engineering, Railway Systems, Railway Operations, Railway Safety, Railway Maintenance”. Of particular significance is ”Rail 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). The authors specifically included only papers with quantitative results, disregarding those about qualitative challenges of BDA in RTS or purely mathematical modelling of RTS problems. The authors 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, simulation, classification, semantic analysis, text analysis, optimisation
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.

3.2 Database search

The papers selected from [9] 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. Where databases allowed Boolean operators, the following formulations were used:

- (((trains OR train) NOT training) OR rail OR railway OR railways OR railroad OR railroads)
- (delay OR delays OR ”event times”)
- (prediction OR predicting OR analysis OR analyzing OR estimation OR estimating)

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. Initially, approximately 3000 studies were identified. Some post-processing was necessary to reduce the number of studies retrieved from IEEE Xplore, in particular, resulting in 69 studies.

3.3 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. Studies were selected by abstract, and then by content. 3 papers were excluded during this process.

If the authors were content from the abstract a paper was relevant, it was selected; otherwise, the paper was read to determine suitability. In total, 19 studies were identified for inclusion in this literature review.

1 was discarded for quality. Of those, 4 were discovered through other means - either from a prior, less structured, search, or from cited references.

4 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][17][14][13]), or even combined multiple models (e.g. the random forests regression of [**wen˙et˙al˙2017**], or emsemble model of [**nair˙et˙al˙2019**]; 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 [20][29]) or which are closer to *algorithmic* models than ML ([**hansen˙goverde˙van˙der˙meer˙2010**]); 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

5 ML models

In this section, each of the distinct ML models identified previously is discussed. It is worth taking studies which compare models with a hint of skepticism. Authors tend to invest considerably more time in the primary model explored; it is unsurprising, therefore, they perform better.

Many studies put great importance on the interpretability of the resulting model. While a desirable characteristic, it is perhaps of such importance here because.... it's real life? bad things could happen otherwise?

5.1 Bayesian networks

We shouldn't have specific sections for each. We're trying to group by model, remember. A Bayesian network (BN) is a "probabilistic graphical model that uses Bayesian inference for probability computations" [**towards˙data˙science˙BN˙intro**]. Each directed edge models a conditional independence, allowing "the incorporation of massive historical data" [12].

The study compared heuristic hill-climbing, primitive linear, and hybrid structure BNs.

There is no common dataset for TDP, unlike other ML areas such as computer vision. That said, data trends can be observed in the papers gathered. Several use the TNV-Extract tool developed by Goverde ? and thus use data from the Netherlands. Several use HSR data from China. Four - those use Italian rail data. The fields of each dataset are explored later on.

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

Uses heuristic hill-climbing, primitive linear, and hybrid structure. Uses real-world train operation from a high-speed railway line, initially to rationalise the dependency graph of the developed structures. Each is then trained with the k-fold cross validation approach to avoid over-fitting and evaluate performance against the others. Used the k-fold cross-validation method to test the train history data in three BN structures (heuristic, naïve, and hybrid) and found that the reconstructed hybrid heuristic BN structure was able to achieve higher prediction performance. Validation results indicate that a BN-based model can be an efficient tool for capturing superposition and interaction effects of train delays. A well-designed hybrid BN structure, developed based on domain knowledge and judgements of expertise and local authorities, can outperform other models. 80% accuracy in predictions within a 60-minute horizon and low prediction errors for MAE, ME, and RMSE. Defines a railway system as several subsystems: network infrastructure, rolling stock, control and communication, and various operational rules and policies. Some delay factors are predictable and controllable; most are neither. A dispatcher's estimation of delays and subsequent decision are strongly dependent on the state of traffic and network and limited to a local geographic area. In large, dense, interconnected networks, such decisions, while locally optimal, may not be globally so. Must support dispatchers by a tool that can account for the interdependencies of train operations and interrelated delay factors. Methodologically, there has been a lack of models capable of simultaneously examining multiple components of delay incidents intertwined with stochastic operations and interaction effects. Technologically, there has been a need for collection and incorporation of massive train operation data. BN are a representational tool meant to capture complex structures. It allows for incorporation of massive historical data in identifying the contingencies between multiple events and updating the state of different variables in real-time. These features, convoluting different factors and fusing massive data, give BNs an advantage. Hybrid BN is tested against different performance measures. First hybrid BN-based delay prediction model in the relevant prediction literature. The main idea behind the hybrid structure introduced here to distinguish between the delay due to the most recent performed operation (e.g. an original delay) and the delay propagated from previous operations (knock-on delays). Made possible by examining the similarities and differences between naïve and heuristic structures supported by domain knowledge and expertise of local authorities. Traditionally scheduled timetables are not adaptive: they often fail to address the time-varying nature of train operations. Each new operational configuration would require re-optimising the timetables, which is computationally expensive. Traditional methods such as regression models require frequent updates of train positions and rich data. Micro- and macro-level simulation tools have been applied to simulate delays at different levels of details. The frequent updates required are mostly due to time-varying operational conditions and the interactions between different subsystems (stations, sections, and trains) under the effects of infrastructure and operational rules. Statistical models are not adaptive enough to incorporate the domain knowledge of local dispatchers and networks' characteristics. To date, identifying which BN architectures are most valid / reliable for predicting train delays for each particular network structure have not been well studied. Need for better predictive models that account for massive real-world train operation data, domain knowledge, and the expertise of local authorities. Built on weather of train operation records, and domain specific knowledge. Proposed model is easy to interpret, generalise, and computationally efficient. Data comes from train operations on the Wuhan-Guangzhou (WH-GZ) high-speed rail (HSR) line in China. 1096km long, with 18 stations. 15 stations and 14 sections are operated by the Guangzhou Railway Bureau and the remained the Wuhan Railway Bureau. Data were extracted from the former's database from February 2015 to November 2015, comprising approximately 380,000 arrival and departure events between stations on the specified line, excluding early arrivals and departures. Operational punctuality is about 85% because of delays. Departure delay is due to the late arrivals or due to disturbances in train operations at stations. Arrival delays are due to departure delays in the previous station or due to a disturbance during traversal time in track sections. Found that arrival and departure delays follow the same distribution, with a linear relationship (chain) with a high correlation between arrival and departure delays at the same station (at least 94%). Used findings to calibrate delay dependencies in the proposed BN structures with different complexity level. Found MAE for all predicted events is around 30s; maximum predicted error is less than 90s. RMSE for both was less than 2 minutes. As this is relatively larger than MAE, suggests the existence of a few outlier prediction errors. Predictions matched only 56% of the time, due to discrete prediction space but continuous variables. So continuous variables were discretised into bins. 3 minutes' width was used for prediction intervals, as late arrivals of less than 90s are not considered delays. As the width of these bins increases, so too does accuracy, as each prediction has a higher probability of falling within the corresponding interval. Overall accuracy to be over 80%, with a no-information rate of 58%. Sensitivity (true positive rate) was ¿ 60%. Inaccurate because of error accumulation which would be addressed fairly easily in real-world operation as predictions could be updated in real-time (e.g. using arrival time at the preceding station, the position of the corresponding train along the track, and the adjusted timetable). Computational time used for training and testing of the model did not exceed 10 minutes. 80% accuracy for a 60-min prediction horizon. It is expected that prediction error could be reduced if the spatio-temporal properties of each track section are also included in the model.

Objectives

introduced the first hybrid BN-based to the area of TDP.

The hybrid heuristic BN, built using naive and heuristic structures and refined by domain knowledge and expert judgements. Achieved 80% accuracy over a 60-minute prediction horizon.

Advantages: it is simple, and so computational efficient. The authors note that results could be improved by including the 'spatiotemporal' properties of each section, which we have taken to mean the speed at which trains can run.

The MAE prediction error was around 30s; the RMSE for both predicted arrival and departure delays was less than 2 minutes

However. Predictions from the hybrid model matched observations only 56% of the time. The authors attribute this to primarily to the discrete prediction space, and so employed discretisation to convert continuous variables into bins.

explore three different Bayesian network schemes: heuristic hill-climbing, primitive linear, and hybrid. Hybrid, incorporating domain knowledge and judgements of local experts, was found to outperform other models, with an accuracy of over 80% in predictions within a 60-minute horizon. The authors define a railway system as several interconnected subsystems: infrastructure, rolling stock, control and communication, and various operational rules and policies. It was found that arrival and departure delays follow the same distribution, with a linear relationship (chain) with a high correlation between arrival and departure delays at the same station (at least 94

5.1.2 Stochastic prediction of train delays in real-time using Bayesian networks [**corman˙kecman˙2018**]

present a stochastic model for predicting the propagation of train delays based on Bayesian networks (BNs). 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. This authors extend this approach by modelling the interdependence between trains that share the same infrastructure or have a scheduled passenger train. The model is tested on historical train realisation data from a bus corridor in Sweden

That is interesting. [**nair˙et˙al˙2019**] trained models on three (apparently) identical measures of delay: travel time, delay, and additional delay (delay accrued since the previous stop). They found that kernels based on the former two substantially outperformed those based on the latter. For computing additional delay, the reference set itself needed to be recomputed each time there was an observation. This proved computationally prohibitive, and so the final model implemented was on delays.

Data-driven methods may not extrapolate well to rare and extreme events such as heavy network disruptions, especially wen few or no examples exist in the training data [**barbour˙et˙al˙2019**]. Trains are heterogenous in respect in tonnage, power, length, and priority.

The validity of all levels of railway operations planning, such as creating feasible and sizeable timetables, predicting real-time traffic, predicting conflicts, and providing reliable passenger information, depends on the accurate estimation of train process times that are subject to delay incidents. [Replicated in Kecman et al, 2015] To minimise the probability of schedule deviation in actual operations, the parameters of train motion equations (with input of the estimated running and dwelling times at individual stations and sections) are usually tuned or optimised based on historical train data. The cost of re-optimising schedules with operational changes can be somewhat overcome by applying data-driven approaches and statistical models to estimate the process times based on various contributing factors. Simulation models, developed based on fixed distributions, require frequent updates from train positions and real-time train data. A generic statistical model for estimating the running and dwelling times. Three global predictive models: robust linear regression, regression trees, and random forests. Based on robust linear regression and some refinements, local models were calibrated for each particular line, station, or block section. Models were evaluated using an aggregated set of historical data on the level of block sections.

5.2 Neural networks

Although initial TDP models used neural networks, their popularity has declined, and they are now used primarily as a benchmark against which to compare other, more sophisticated, ML models.

Perhaps the oldest application of ML to the more general problem of predicting delays in public transport is [**peters˙et˙al˙2005**] (prediction of delays in public transportation using neural networks). This study was not selected for inclusion in this review, however. Instead, the more recent [**yaghini˙et˙al˙2013**] will be discussed.

Proposed a high-precision neural network to predict the late arrival of Iranian railway passenger trains. The authors used decision trees and multiple logistic regression models to evaluate the quality of the results. An artificial neural network model was proposed to predict the delay of passenger trains in Iranian Railways. The accuracy level of the proposed model was found to be superior to other statistical models such as decision tree and multinomial logistic regression models. Managing the consequences of incidents and getting trains running normally again is vital to reducing delays. Data from 2005 to the end of 2009 is used. The average delay in this period was 18174 hours per year (approximately 1.1 million minutes) and 30 minutes per train. Stopping time at interval stations (dwell time) for praying / boarding / alighting passengers are excluded from delay time. Causes for delay: Delay at the origin: difference between the actual train's departure time and the scheduled train's departure time Incidence with another passenger train or freight train. Happens when trains running in opposing directions pass each other at places where loops or sidings are available. Unscheduled waiting time at overtaking points: the train waiting for the arrival and passing of another train with common path according to its priority Engine breakdown

Other causes: wagon breakdowns, infrastructure faults such as track and signal failure, and non-scheduled stops for praying (lol) Data considered a total of approximately 180,000 trains, with a total delay of approximately 5.5 million minutes. Again, bins delay time by inconsistent intervals. For the normalised real number, inputs were an origin-destination pair, the rail corridor, the day, the month, and the year. Also used binary set encoding, and binary. Binary inputs make the network structure size too bigger, requiring more memory and consequently greater time to solve problems.

Quick method: one is trained. Dynamic: topology of the network changes during training, with units added to improve performance until the desired accuracy is achieved. Multiple: train multiple networks in a pseudo-parallel fashion, and choose the best-forming network. Decision tree and multimodel logistic regression were used to evaluate results. Most accurate was the binary quick neural network. The proposed model has great accuracy and low training time. Future research will improve training time and improve prediction accuracy. Accuracy may be improved through meta-heuristic methods such as genetic algorithms, simulated annealing, or hybrid algorithms. Training time can be improved through particle swarm optimisation or continuous ant colony optimisation .

5.3 Support vector regression

Support vector regression (SVR) uses support vector machines (SVMs) as

A SVM maximises the margin between two or more classes to find the optimal *hyperplane*, the seperator between two classes. In SVR, the hyperlaine is used to

SVR is a popular ML algorithm grounded in statistical learning theory and for which training is efficient due to the convexity of the training problem [**barbour˙et˙al˙2019**]. Construct a vertex-edge graph from track infrastructure data.

A kernel

SVRs can be used for continuous values; SVMs are for classification.

[**nair˙et˙al˙2019**] 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.

[**barbour˙et˙al˙2018**] use SVR to estimate freight delays. They use origin-destination specific features, train priority, and train counts as influencing factors. They do not report on prediction error, but show relative improvement over a baselines of historical mean forecasts". Uses real-time data. Uses track geometry: grade and curvature information, single and multi-track territory, length of sidings), historical runtime profiles of trains, properties of trains (length and tonnage) and crew records. Note the difficulty of using features that change: the amount of traffic on the line of the road, the number of available sidings (as trains enter and leave).

Construct distincit 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. [**oneto˙et˙al˙201** do the same; they note that approximately 600,000 models would need to be retrained each day.

In a railway network with k nodes, at minimum k^2 models are required; possibly more if multiple paths exist between each origin-destination pair. In practice, the system is rather more complex. A railway network can be considered a collection of sub-networks, each a route or line (likely operated by a separate entity, as in the UK), and with little engagement between different lines.

The paper needed to train only 140 models, instead of 1225. The authors estimate a total of 10,000 models are necessary for all ETA predicitons.

They predict arrival delays, and note an average improvement of 14% across the study area.

Collection of datasets: freight train movement, train car operations, crew, and locomotive data. Network data is extracted from dispatching, operations, and signalling data. The authors note the importance of crew data. Most countries have a maximum on-duty duration (12 hours in the US) after which they must legally go off duty, despite the large expense of stopping a train and transporting a replacement crew.

However, freight trains are very different from passenger trains: fewer stops, longer journeys, and different times (usually overnight). 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.

The authors 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.

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%.

[**markovic˙et˙al˙2015**] found that their SVR performed better than the ANN. First application of an SVR to TDP.

The paper considers only seven influencing variables: Passenger train category (suburban, regional, long-distance) Scheduled time of arrival at station (continuous) Infrastructure influence defined by expert opinions $(3-9)$ Percent of journey completed distance-wise (continuous) Distance travelled (continuous) Time travelled (continuous) Headway (continuous)

Found that scheduled time of arrival and headway are not strongly correlated with any other covariates. 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.

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

Authors state an interesting extension would be the capture of infrastructure influence varibales thourgh input variables without quantification from SMEs. Or further stratification of data: seperate models for short and long delays, as they are caused by different things.

Capturing the state of the network at a given point in time is the key thing here. Fascinating. Absolutely fascinating. Ag

5.4 Decision trees and random forests

First introduced in [**ho˙1995**] (Random Decision Forests)

Accuracy depends on the number of trees composing the forest, the accuracy of each tree, and the correlation between them [**breiman˙2001**] (random forests). 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.

A random forest (RF) is a collection of individual decision trees (DT). Simply put, each individual tree predicts the class of an input and the class with the most votes is the output of the model. ”A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models” (https://towardsdatascience.com/understanding-random-forest-58381e0602d2). This is the same that underpins ensemble methods; a RF could, in fact, be considered an ensemble method.

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.

A random forest is a meta-estimator. Fits several decisions trees on various subsamples of the training data and uses voting or averaging to improve accuracy and prevent overfitting.

[**nabian˙et˙al˙2019**] uses a novel bi-level RF for TPD. The primary level predicts whether a delay will increase, decrease, or remain unchanged in a specified time frame; the second then estimates the actual delay (in minutes), given the predicted delay category at the primary level. Constructing the model is computationally cheap.

Found that the proposed model provides the best prediction accuracy. Unusual structure is to meet recommendations of the two main railway operators in the Netherlands (ProRail and NS). Hence the coupled classification-regression task.

Compared to linear regression, multi-nomial logistic regression, decision trees, k-NN, and SVM / SVRs.

Slightly different focus: not the prediction of the occurrence of delays, but the prediction of change of *severity* of a delay, given that it has already occurred (i.e. the train is already behind schedule). 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), delays longer than 15 minutes.

Extent to which drivers can compensate for the delay depends on the distance and planned time difference between stations. l is the number of class labels in the classification task. Consists of a RF classifier and $l + 1$ RF regression models.

Assumed that all trains rode on a single track.

[**oneto˙et˙al˙2018**] (dynamic, interpretable) Focuses on the running time, dwell time, train delay, penalty cost (an interesting idea). Perhaps only applicable with different classes of train, right? and train overtaking between two trains in the wrong position relative on the railway network. Two approaches; one is based on the knowledge of the network and the experience of the operators. The other is based on the anaylsis of the historical data of the network with advanced data analytics methods. Former are interpretable and robust, but not very accurate; latter are the opposite.

Former is not computationally demanding, hard to modify to incorporate complex phenomena, and not dynamic.

Propose a hybrid model (HM). This perhaps should be under ensemble methods. Handles infrequent events (e.g. passage of freight trains), fast changes of train movements, easily extensible. First proposal of such a HM. Delayed if 30s late! Predict running time for all subsequent railway sections that it traverses. Dwell time for all the subsequent checkpoints at which it will stop, updating each time it reaches the next checkpoint. Allows TOCs to undestand how much time a train needs to complete the itinerary.

Penalty cost. Fairly novel. NR has too has a ex post facto delay attribution process. Based on train category, operational category, type of railway section, amount of delay, and percentage of responsibility.

A DDM similar to that described in [**oneto˙et˙al˙2016**] is currently in use at RFI, the Italian infrastructure manager (as in [**oneto˙et˙al˙2018**] (dynamic). The RFI DDM consists of many DDMS. For each train and each checkpoint, a DDM is constructed. Can be built using many different learning algos. RF lead to better results. If the timetable changes, it takes at least a month of data before achieving a reasonable accuracy. C

Idea of the EBM is to analyse time each train needs to traverse each seciton, based on speed limits, state of the network, and type of train. A coefficient (gaining time) is computed which represents the time that can be regained in case of delay. This is static.

HM predict running time, dwell time, train delay, penalty cost, and trian overtaking. Encapsulates the experience of operators into a decision tree, but the leaves are constructed following ideas of the DDM. Does not implement one model per train. Groups trains on a series of similarity variables. Grouping increases the number of historical data to exploit during leaf creation, fixes new timetable issue. Leave is a DDM that learns from historical data of all trains that fall into that particular leaf (similar characteristics, iterinary). Each leaf is a RF. So really it’s a fucking DT of RF! The whole HM is constructed and updated incrementally as soon as nwe train movements are recorded in an online fashion.

In the top-level decision tree, a new leaf is added each time a new train movement 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. Forget movements older than 3 months. Based on experience of operators and different window size. Prediction is just consulting the appropriate leaf.

Clearly outperforms the EBM and the DDM. Most evident for freight and regional trains. Constantly better across the whole year. Reaches optimal accuracy after 10 days of results.

Don't predict delays directly: instead use running time and dwell time predictors as building blocks.

HM provides the best trade-off between accuracy and computational requirements.

Lacking the co-operation of the NR.

[**nair˙et˙al˙2019**] selected RFs for the accuracy of forecasts and the possibility of incremental (updating model parameters as fresh data becomes available) and parallel training.

[**wen˙et˙al˙2017**] DD delay-recovery for HSR. Identifies main variables that contribute to delay: total dwell (TD), running buffer (RB) time, magnitude of primary delay (PD), and individual sections' recovery.

Evaluates multiple linear regression and random forest regression.

Delay recovery: how fast a train service can recover from its delay in the subsequent stations. An important performance measure that shows the reliability and robustness of the service being provided.

roughly 30,000 records. 14 stations. roughly 1100km of track.

PDs can be reduced by adjusted running and dwell times. Buffer time = differenced between the scheduled running time (or dwell time) and the minimum required running time (or dwell time) is included in the schedule.

5.5 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.

Regression is a relatively simple technique that has been expanded into many variantss. The most sophisticated of htese are discussed separately (see SVR). Here, we discuss more simplistic regression models.

[**wang˙et˙al˙2019**]

This paper used a three-month dataset of weather, train delay, and train schedule records. The key finding paper was that, 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. It blurs the distinction between short-term and long-term established earlier. Indeed, their division is simply between 'short-term' (real-time, as defined above) and 'long-term', consisting of both short- and long-term as defined above. As the paper explains, long-term prediction is more useful for passengers in planning and adjusting their trips, as a passenger is usually unable to modify their travel times when short-term delay data are released. This paper also lays out the rationale for this dissertation: helping passengers plan more reliable journeys, and for operators developing more efficient train schedules. Short-term prediction is usually estimated from real-time operating data. Long-term train delay prediction model is developed based on advanced weather-forecasting techniques, the close link between weather and train delays, and the relatively consistent rules of train operation. Study combines weather records, historical delay data, and train schedule data. Proposes the new concepts of key train delay stations and the time interval threshold to determine whether a delay to one train results in a delay to the following train. Time interval threshold: the time interval is the difference between the arrival times of two consecutive trains at the same station. The threshold is set to determine whether the delay of a train at a station is likely to propagate to the following train arriving at the same station. The weather data is relatively limited in scope. It includes the lowest daily temperature, highest daily temperature, weather (e.g. overcast, sunny, light rain), Beaufort scale, and air quality index. As might be expected, precipitation caused delays, and, when combined with freezing temperatures, serious delays. 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. In bad weather, the operating speeds of trains are reduced for safety reasons. Average delay was around 10 – 20 minutes in good weather (sunny, cloudy, overcast). The average delay times of adjacent cities are usually similar under the same weather conditions. 75 stations analysed on the line. Most stations have limited overall train delay time and overall number of delays, whily only a small number of stations have large delays. Good positive correlation between the total number of delays experienced by a train operating in the line in March, January, and February. Identified station sequence at which consecutive train delays occurred. The first station of each is the source of the train delay. Stations that often saw original delays are identigide as key stations. If one train is delayed at station k, the following train stopping at the same station may also be delayed. So: delays can propagate. Used density-based clustering algorithm to identify the time interval threshold to determine whether the delay of a train at a given station propagates. The length of the train delay propagation chain can be approximated by exponential distributions. So there exists limited large scale train delay propagation. The paper makes uses of forecasts up to 10 days in advance. This resolution is not available on a regional scale in the UK, but national forecasts are. This, naturally, diminishes the accuracy of the model. Also of interest is the likelihood of that weather occurring. The authors note that snowy weather resulted in greater train delay times in southern cities, which, experiencing snowy weather much more rarely, are more poorly-equipped to deal with it as their northern counterparts. "Identified the station sequence at which consecutive train delays occurred". The first station of each sequence is the source of the train delay. The initial delay usually leads to subsequent delays at subsequent stations: delay propagation. Stations that often saw initial delays were defined as key stations. There is a necessary time interval between two trains passing through he same station. DBSCAN (Density based clustering algorithm) The paper defines a weather score, which quantifies the severity of delay times under particular weather conditions. They also used the number of trains passing through a station, used as a messaged of train service infrastructure Useful for passengers wishing to plan journeys more reliably and for developing more efficient train schedules and more reasonable pricing plans. Attributes delays to factors ranging from severe weather to equipment failure and poor management. Usually caused by temporary speed restrictions imposed for safety reasons. Can be divided into short- and long-term. Network Rail's Darwin system already has an integrated short-term delay prediction mechanism. Long-term prediction is more useful for passengers in planning and adjusting their trips. Three factors: the weather score, the number of trains passing through the station I, and the total number of delays of a train. First quanitifies the severity of delay times under particular weather conditions. Second concerns train service infrastructure, which is also correlated with train delay time. Gradient-boosted regression trees model was used to build the prediction model for train delays. Errors attributed to stochastic equipment failure or other human and operation factors. Could also result from only thee months' of data. Future studies would incorporate as-yet unreleased data. This paper is mainly of interest as the only extant research focusing on the short-term prediction period that this dissertation will also focus on. Main factors: • Natural phenomena (changes of weather, natural disasters) • Human factors (improper operation) • Systemic failures (signal communication failure, cable failure, power outages)

[**wang˙and˙work˙2015**]

Proposes historical regression model to estimate future train delays at each station using only past performance of the train along the route. Several variations of an online regression model are proposed to estimate delay using delay information of the trains at earlier stations along the current trip, as well as delay information of other trains that share the same corridor. Data is from 282 Amtrak trains from 2011 – 2013 (more than 100,000 train trips). Proposed historical regression model improves the RMSE estimate of delay by 12%; online model improves the RMSE estimate of delay by 60%. Delay: the difference between the true running time and the free running time. Variabilities associated with train operations (equipment maintenance, station dwell time, weather). Amtrak trains have priority yet the on-time rate of Amtrak is less than 50%. Average delay for several trains can reach as high as 50 minutes. Analytical approaches are elegant; or simulation approaches are realistic. Application of either constitutes a major model building or calibration task. For complex systems, analytical methods require abstraction to maintain tractability. For simulation approaches, extensive effort is required to accurately calibrate the model. Regression models can be constructed to estimate delay, where parameters are calibrated by learning from historical data. Data-driven approaches prevent application of these methods for scenario planning, for which analytical or simulation approaches are more appropriate. Regression model is for before the trip starts. Online regression model is for after trip ends. Breaks down into analytical delay estimation methods, simulation methods, and data-driven methods. Analytical methods provide explicit mathematical relationships to estimate delays, but cannot fully capture the delays caused by complex interactions among trains, the variabilities among train operations, and operating parameters. Hence the need for simulation based work! Main advantage of simulation models is that they are capable of incorporating the sophisticated interactions of trains on complex infrastructure, and the resulting delays can be easily estimated once the model is calibrated. However, still an approximation. 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 bring 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. Determine parameters of regression model through least squares estimation on the training dataset. Dataset contains all Amtrak passenger train arrival and departure data from each station from 2006 – 2013. Each contains: station code, scheduled arrival day and time, scheduled departure day and time, actual arrival time, actual departure time, and comments. Found to be coarse and missing a lot of records. Like actual train arrival . . . Once a delay occurs on a trip, it is likely to last for several stations. Regression model could not be constructed for all trains as they were subject to a route re-configuration, and so a complete set of training / test data is not available. Both online models before significantly better than the historical model. Predicts train delays at each station before the current trip starts based on the delay recorded in the past trips.

This is barely even regression. I'm going to ditch this one.

Er.... what is this from? 231 trees...? [**nair˙et˙al**] Found that around 200 trees was suitable. Ended up with 231 trees. Each tree hs two variables for each node. For categories of 3 minutes, RFR was 90.9% accurate; for MLR it was 84.4%.

Absolute mystery, this.

Fuzzy Petri nets (FPNs) are "modifications of classical Petri nets for dealing with imprecise, vague, or fuzzy information in knowledge based systems" [**liu˙et˙al˙2017**] (Fuzzy Petri nets for knowledge representation and reasoning: A literature review). 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 and high-level PNs (HLPNs) have much more modelling power. Can effectively analyse concurrent systems by verifying safety rules. Uses a graphical presentation that is easy to understand; the authors adapt the FPN to resemble a dispatcher's interlocking control panel. Easy to modify because of granularity.

Constructe using transitions, places, tokens, and arcs. Uses Sugeno method for the construction of fuzzy models.

ANFIS combines a neural network and fuzzy logic: the ability of an NN to learn and the capability of fuzzy lgoic systems to interpret the imprecise data. Defines a bunch of modules to plug together: station track section, train generation, primary delay calculation.

Used data on train dalys for July 2010. 3710 trains: international passegner, domesic passenger, suburban and regional passenger, international freight, ... and so on.

Fuzzy inference system trained by backprop optimisation; output generated by linear membership function.

[**milinkovic˙et˙al˙2013**] is, to the best of the authors' knowledge, the only application of FPNs to TDP in the literature. Train primary delays were simulated by a FPN module in the model. In the case where there were no historical data on train delays, expert knowledge was used to define fuzzy sets and rules. A model based on the Adaptive Network Fuzzy Inference System (ANFIS) was used for systems where historical data was available; it was used to train the neuro-fuzzy ANFIS model, after which it was replicated by an FPN. It was tested on part of the Belgrade railway node.

If bufffer times between trains are less than the length of the primary delay, delay is propagated to other trains. Three common approaches: analytical, micro-simulation, and statistical analyses.

Fuzzy logic is "a mathematical tool used to model traffic processes that are distinguished by subjectivity, uncertainity, ambiguity, and imprecision". [**milinkovic˙et˙al˙2013**]. Used the following data categories: train category, time of arrival at the station, the distance travelled, and the infrastructure influence. All but time of arrival were inputs.

Used algorithm: analyse railway system. Prepare definitions for fuzzy logic model and data on timetable and infrastructure. ANFIS model is created and trained on collected data. FPN module is constructed according to the structure, realtions, rules and weights of the FL model. Modules that represent distincit types of sections in the model are created and defined. HLPN graph is created b y connecting them.

Sub-5% accuracy for two stations, but over 10% for two others, due to the lower number of trains and many delay outliers. Infrastructure data: section lengths, section plans, restrictied speeds, rack routes.

ANFIS calculates delay for each train. Need to add a module for train route conflict management that uses fuzzy logic to model the strategy of train dispatchers in train control.

5.6 Ensemble methods

Many state-of-the-art algorithms (e.g. bagging, boosting, SVMs) used a weighted combination of simpler classifiers. RF combines bagging to random subset feature selection. Each tree is independently constructed using a bootstrap sample of the dataset. Each node is split using the best among a subset of predictors chosen randomly at that node. Accuracy depends on the number of trees composing the forest. Agreed set of novel KPIs. Treated problem as a time-series forecast, required

Ensemble methods use multiple models to generate forecast. The rationale behind such a framework is simple: gathering forecasts from a diverse set of models reduces bias and error rates. [**nair˙et˙al˙2019**] uses a purely data-driven ensemble method; [**oneto˙et˙al˙2018**] combine a RF and a experience-based model.

[**nair˙et˙al˙2019**] used 3 models for operational trains (those currently running): a RF (n-stop ahead), mesoscopic simulation and kernel regression. And static RF for non-operational trains, and mesoscopic simulation.

5.7 Extreme learning machines

ELMs 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.

Just what is a extreme learning machine?

A family of ML algos that represent solutions in terms of pairwise similarity. Preeminence recently challenged by deep NN and Esnebmle methods.

5.8 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

[**sahin˙et˙al**] has a slightly different focus: assessing the effectiveness of time allowances.

Markov chain focuses only on state changes. Ignores that time at which they occur. Train schedules perform two basic functions: efficiency measuring how well resources (infrastructure, rolling stock, crew, time) is being utilised, and the extent to which goals are being achieved (robustness, punctuality, relaibility). Time allowances are used to recover small delays. Two types:

running time supplement (used to make up a small delay). Added to train paths as a percentage of the minimum running time based on the section-specific speed limit. Extends the running time of a train. buffer time (used to stop delays propagating). Placed between consecutive train paths in addition to minimum headway. Reduces the number of trains that can be scheduled, but increases the quality of the service.

Data is from Turkish state railways. Run-time supplement is 13% of minimum journey time.

[**gaurav˙et˙al˙2018**]

Study systemic delays in train arrivals using n-order Markov frameworks. Uses train operations data for the past two years. An efficient algorithm for estimating delays at railway stations with near accurate results.

Delays are credited to obsolete technology (e.g. dated rail engines), size (e.g. large network structure and high railway traffic), weather (e.g. fog in winter months in north India and rains during summer monsoons countrywide).

Builds a N-order Markov late minutes prediction framework. Builds a scalable, train-agnostic, and Zero-Shot competent framework for predicting train arrival delays. Only a small dataset is used: 135 trains. March 2016 - February 2018.

Zero-shot learning: training and test set classes' data are disjoint. Not a problem with large datasets. Observed that mean late minutes varies monthly.

Used RF regressors and ridge regressors. Avoided building train-specific models for real-time deployment and scalability. Stations with a high traffic and degree strength (number of connections) tend to be bolttelnecks,

n-order Markov depends on up to n previous states. So: leave n-OMPR models for each known station. Use k-NN to identify similar stations to those in the known dataset and the unknown (hence the zero-shot). Not necessarily real-time. Using just 1.2% of trains in India, was able to covert more than 11.3% of stations. A fairly innovative way to escape the limitations of a small dataset.

6 Analytical models

A brief overview of analytical models is provided for context, with emphasis on the shortcomings of such models, and how data-driven approaches can ameliorate these shortcomings. In short: analytical models perform worse, but are easy to explain, understand, and interpret. Analytical models cannot capture the complexity of such models.

An *analytical model* is "primarily quantitative or computational in nature and represents the system in terms of a set of mathematical equations that specific parametric relationships and their associated parameter values as a function of time, space, and/or other system parameters" [8]. Current state-of-the-art TDPS use analytical models [17].

Simplistic early models, such as [7] made overly restrictive assumptions about railway operations by, for example, forbidding overtakes, assuming that departure times are uniformly distributed, and that the speed of each train is unique and constant.

Subsequent work in this area has largely relaxed these assumptions, by including factors such as overtakes, different speeds, priority systems, and uncertainties associated with train departure time [19], [4]. More complex models have also emerged, incorporating stochastic approximation [2] and the impact of dispatching strategies on train delays and passenger waiting time [18]. Although the state-of-the-art advances constantly, a good example of an recent *in-use* system is [1], which is currently used in the German rail network.

7 Datasets

There is no common dataset for TDP, unlike other ML areas such as computer vision. That said, data trends can be observed in the papers gathered. Several use the TNV-Extract tool developed by Goverde ? and thus use data from the Netherlands. Several use HSR data from China. Four - those use Italian rail data. The fields of each dataset are explored later on.

8 Fields

9 Exogenous data

It is widely accepted amongst ML practitioners 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* or exogenous data can be incorporated. This is the realm of 'big data', which involves "multiple datasets and a complicated structure" [**Ghofrani˙et˙al˙2018**]. This section is broken down by the classification defined in the introduction.

Data is exogenous if it is independent of other input data but the output data depends on it. The scope for inclusion is essentially limitless: any source of data which may affect railway operations is a viable candidate. In the studies selected for this review, there are two main sources: infrastructure [**markovic˙et˙al˙2015**] [**milinkovic˙et˙al˙2013**] via *expert opinion*, and weather [**oneto˙et˙al˙2017**] [**oneto˙et˙al˙2018**], [**oneto˙et˙al˙2019**].

Special mention must go to [**nair˙et˙al˙2019**], which used network traffic states, such as likely stretch conflicts and current headways, weather, event information, work zone information, inferred occupation conflicts, train connections, and rolling stock rotations.

9.1 Weather

Weather is a common cause of primary delay. The classification defined earlier included severe heat, flooding, landslips, leaves, snow, and ice. It is expected that weather-induced delays are seasonal. Severe heat is likely to cause delays in summer; leaves in autumn; and snow and ice in winter. [**brazil˙2017**] found that most weather-caused delays occurred in the last third of the year, with a peak in November. Weather is also a popular inclusion for TDP.

Weather was first included in a TDP model, to the best of the authors' knowledge, in [**oneto˙et˙al˙2016**].

Subsequent studies have established the impact of severe weather on train delays "BRAZIL201769, title = "Weather and rail delays: Analysis of metropolitan rail in Dublin". Papers largely agree that, dependent on climate, weather delays most trains during the last third of the year, with November a particular culprit, likely due to the sustained impact of leaf-fall.

[**wang˙et˙al˙2019**] observed that in locations less prepared for specific severe weather - such as snowy weather in southern cities - delays were greater. They found that in severe weather trains delays are determined by mainly the type of bad weather, but in ordinary weather they are determined mainly by historical delay time and the delay frequency of trains.

Fields tend to be largely consistent: there are only so many weather variables of note. [**wang˙et˙al˙2019**] used lowest temperature, highest temperature, weather category (e.g. "overcast", "light rain"), Beaufort scale (wind speed) and air quality index (seemingly unique to China). Data was collected from 344 cities along the route in question. However, the timeframe used was between 1st January and 31st March. As weather-related delays are seasonal, this reduces the validity of conclusion relating to the importance of weather.

[**oneto˙et˙al˙2016**] note that weather conditions can additionally influence passenger flow and consequently dwell times, which have already been described as a key influence on delays.

[**oneto˙et˙al˙2016**] CITE ALL HERE use temperature, relative humidity, wind direction, wind speed, rain level, pressure and solar radiation. [**nair˙et˙al˙2019**] use weather data from 92 weather observatories, including snow conditions, visibility, and temperature. They found that weather has only a small impact of delays; an analysis of delay-attribution showed that less than 3% of delays were directly attributed to weather.

The proposed dataset for this dissertation uses similar fields. For interoperability with forecasts, the comprehensive data provided by the Met Office has been mapped to a more simplistic set of fields: wind gust, relative humidity, visiblity, wind direction, wind speed, temperature, weather type (category), and precipitation probability.

[**wang˙et˙al˙2018**] collected weather data from 344 cities along the route in question, Beijing to Guangzhou. It is worth nothing that the two are approximately 2200km, and so delays are of a magnitude not frequently found

[**oneto˙et˙al˙2016**] 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.

[**nabian˙et˙al˙2019**] only had a daily overview of: maximum wind speed, maximum, minimum, and average temperature, and rain depth. Authors note that the hourly data would have improved the model. Could not be considered significant as a result.

9.2 Infrastructure

Infrastructure naturally matters to trains. Extant work in ML for TDP has been surprisingly lacking incorporating infrastructure characteristics (single or double track, station layouts, interlocking, cant, speed limit), and so on. It is a thoroughly modelled in the analytical models discussed earlier.

[**milinkovic˙et˙al˙2013**] 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.

The authors note that the average track occupancy of a section can indicate possible bottlenecks of a system. This is close to the *tactical* level briefly discussed earlier: the use of data to make decisions on improving infrastructure.

The dataset for the proposed study includes infrastructure characteristics used by ? to actually plan train delays.

Used the Delphi method.

It seems inherently obvious is should have a huge effect on the propagation of delays.

[**markovic˙et˙al˙2015**] explores an expert opinion in much better defined terms. The influence of multiple factors 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) is aggregated into one variable with a value determined by the expert opinions of five dispatchers. Estimates obtained via the Delphi method: experts evaluate a route over multiple rounds until a consensus is reached. Strong correlation found between expert opinions and train delays.

The condition of the Serbian railways is considerably worse than that of many other countries explored. Characterised by: recently renewed lines (enabling maximum speed), lines with sections with TSRs, single and double-track lines, many junctions and railroad crossing, lines split int sections with different signalling and safety equipment.

Evaluated each route on a scale of 1 - 10. 1 denotes a route with the highest number of infrastructural factors that could cause unplanned delays.

The study only considered a limited number of routes: those passing through Rakovica station. Models for larger areas cannot rely on human assessment of infrastructure conditions. 39 lines were evaluated. Specifically: number of stations/stops/junctions/loops/crossings, percentage of single track, percentage with restricted speed, length with restricted speed, block section, track clear section (station distance, braking distance, automatic block system, centralised traffic control, axle counters).

[**nair˙et˙al˙2019**] 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.

9.3 Maintenance

Only one paper was found that incorporated the "maintenance" class: [**nair˙et˙al˙2019**]. The authors used work zone information, indicating location, duration, and the likely impact of different train categories. The proposed dataset uses the RDG Knowledgebase API, which includes data on "incidents": service disruptions and engineering works.

9.4 Other

No papers were found to incorporate accidents, vandalism, trespassing, or fatalities, or strikes. Holidays are, however, included in [**nair˙et˙al˙2019**], and will be included in the proposed dataset. NR categorises the day by weekday, Saturday, Sunday, Christmas, and Bank holiday, reflecting the different timetables used for each.

Right. Now what? I suggest that we go through methodically and tidy up. But first: let's finish my evaluation and conclusion, if there is one. And I'll be damned if this is not considered a goddamn

10 Evaluation

What are the research challenges? How has HOI been used to date?

10.1 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 [**yaghini˙et˙al˙2011**], 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 [**markovic˙et˙al˙2015**] [**barbor˙et˙al˙2019**]. Markov models are also used, but only in limited contexts, in [**gaurav˙et˙al˙2018**] [**sahin˙2017**].

More esoteric ML models, such as fuzzy Petri nets, are also used in [**milinkovic˙et˙al˙2013**], although without further development.

Currently at the forefront of research are three models: random forests [**oneto˙et˙al˙2016**] [**sabian˙et˙al˙2019**] [**oneto˙et˙al˙2018**] [**nair˙et˙al˙2019**], Bayesian networks [**lessan˙et˙al˙2019**] [**corman˙et˙al˙2018**] and extreme learning machines [**oneto˙et˙al˙2016**] [**oneto˙et˙al˙2017**] [**oneto˙et˙al˙2018**]. Increasingly, these models are combined into an ensemble or hybrid model, as in [**nair˙et˙al˙2019**] (random forests, kernel regression, and mesoscopic simulation) [**oneto˙et˙al˙2019**] (random forests, experienced-based model).

For the authors' own work, the methodology and model structure described in [**nair˙et˙al˙2019**] 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 [**oneto˙et˙al˙2019**].

10.2 RQ2: What exogenous data is used to improve the performance of those models? / big data analysis

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

Weather was first used in [**oneto˙et˙al˙2016**], and has subsequently been investigated in [**brazil˙2017**] and applied in [**wang˙et˙al˙2019**][**nair˙et˙al˙2019**][**nabian˙et˙al˙2019**]. [**oneto˙et˙al˙2016**] found 10% performance increase when incorporating weather data into their RF model. However, [**nair˙et˙al˙2019**] 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. [**milinkovic˙et˙al˙2013**] and [**markovic˙et˙al˙2015**] 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.

[**nair˙et˙al˙2019**] 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 [**nair˙et˙al˙2019**] or an existing data-set (if they can finally fucking parse it)!

Maintenance is used only by [**nair˙et˙al˙2019**], 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.

10.3 RQ3: How has ML been used in date in the area of TDP?

This has perhaps been the hardest RQ to answer. Only one data point is available. [**oneto˙et˙al˙2019**] note that, since their earlier paper [**oneto˙et˙al˙2016**], RFI, the Italian railway manager, has adopted their model and a experience-based models similar to that described in [**hansen˙et˙al˙2019**]. It seems clear the the future of TDP is ML, not analytical.

11 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] 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`.

[2] Malachy Carey and Andrzej Kwieciński. “Stochastic approximation to the effects of headways on knock-on delays of trains”. In: *Transportation Research Part B: Methodological* 28.4 (1994), pp. 251–267. DOI: `10.1016/0191-2615(94)90001-9`.

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

[4] Bintong Chen and Patrick T. Harker. “Two Moments Estimation of the Delay on Single-Track Rail Lines with Scheduled Traffic”. In: *Transportation Science* 24.4 (1990), pp. 261–275. DOI: `10.1287/trsc.24.4.261`.

[5] Winnie Daamen, Rob M. P. Goverde, and Ingo A. Hansen. “Non-Discriminatory Automatic Registration of Knock-On Train Delays”. In: *Networks and Spatial Economics* 9.1 (2008), pp. 47–61. DOI: `10.1007/s11067-008-9087-2`.

[6] Mark Dingler et al. “Determining the causes of train delay”. In: *AREMA Annual Conference Proceedings*. 2010.

[7] Ove Frank. “Two-Way Traffic on a Single Line of Railway”. In: *Operations Research* 14.5 (1966), pp. 801–811. DOI: `10.1287/opre.14.5.801`.

[8] Sanford Friedenthal, Alan Moore, and Rick Steiner. *A Practical Guide to SysML*. 2nd ed. Morgan Kaufmann Publishers, 2012.

[9] 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`.

[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] Pavle Kecman, Francesco Corman, and Lingyun Meng. “Train delay evolution as a stochastic process”. In: *Proceedings of the 6th International Conference on Railway Operations Modelling and Analysis: RailTokyo2015*. 2015. URL: `http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-119746`.

[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] 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`.

[14] 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`.

[15] Robert B. Noland and John W. Polak. “Travel time variability: A review of theoretical and empirical issues”. In: *Transport Reviews* 22.1 (2002), pp. 39–54. DOI: `10.1080/01441640010022456`.

[16] 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`.

[17] 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`.

[18] Süleyman Özekici and Selim Şengör. “On a Rail Transportation Model with Scheduled Services”. In: *Transportation Science* 28.3 (1994), pp. 246–255. DOI: `10.1287/trsc.28.3.246`.

[19] E. R. Petersen. “Over-the-Road Transit Time for a Single Track Railway”. In: *Transportation Science* 8.1 (1974), pp. 65–74. DOI: `10.1287/trsc.8.1.65`.

[20] 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`.

[21] Andre Puong. “Dwell Time Model and Analysis for the MBTA Red Line”. In: (Mar. 2000).

[22] 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).

[23] 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).

[24] 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.

[25] 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.

[26] Hor Peay San and Mohd Idrus Mohd Masirin. "Train Dwell Time Models for Rail Passenger Service". In: *MATEC Web of Conferences* 47 (2016), p. 03005. DOI: 10.1051/matecconf/20164703005.

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

[28] 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.

[29] 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.

[30] 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.