**Objectives**

By the end of this module, I should have:

* Investigated a problem and produced a survey of existing techniques
* Proposed a plan
* Proposed a solution, described in detail in a design document
* Undertaken a substantial implementation of the proposed solution
* Presented the work, orally, in a group and in one-to-one examination
* Written up their work in a scholarly fashion

Okay. As things currently stand, I have done few of those. I haven’t even proposed a *plan*. But that is why I want to talk to Noura. Let’s have a look at the example plan. Hm. I DO remember writing a plan. Did I ever submit it? If not… where is it? Well, it likely needs massively updating anyway.

Okay. I sound a bit of an arsehole in my project plan, it must be said. Beyond scientific impassivity and into arrogance. I also massively understate the preliminary preparation. I will also go into more depth on my objectives. Done. It’s all things I’ve done already, of course: testing on logistic regression, and random forests, and neural networks with PyTorch. I want another carbonara! Part of this progress must be agreeing with Noura, which is why I’m inclined now to get more data sorted. Once more unto the breach…

I’d refactored all of my download code to make it sexy. My DARWIN archive contains data from 2017-01-01 to 2020-01-02. I do not have my timetable data. I’ll get that downloading now. Okay. I dislike how my repository is structured, but that’s just because I’m anal about this kind of thing. It’s justifiable if fixing the issue makes me better at working, but I think in this case I should just get over it. Nor do I have my weather data. That’s irritating.

Okay, we’re back. I really should go to the library, but I don’t want to leave the house for risk of getting infected. Is that true? No. I am planning on going to the gym, after all. And given a 1-in-100 to a 1-in-25 chance of dying: it’s simply not worth the risk. Good thing that I have weights at home!

Anyway. Download the data, and transfer it to the hard drive. I need to fix a few bugs regarding how I process it, as well. And finalise an acceptable structure.

“””

Initial testing with the simplest of machine learning models – logistic regression – yielded unexpectedly good results. The default SciPy logistic regressor achieved 97% accuracy in a classification task: whether or not a train would be delayed.

5 different models were tested: three linear, and two ensemble. Logistic regression achieved 96% accuracy. The most important feature was ‘category’.

Bayesian Ridge achieved very disappointing results: 6% accuracy. The most important feature was origin\_year and destination\_year.

Gradient boosting classifier achieved 91% accuracy.

Random Forest achieved 97%: destination minutes, origin\_minutes, duration, origin\_day, destination\_day, num\_stops, and category, though interestingly category had the largest error out of all features.

SGD achieved 91% accuracy, with category and seating again the most important.

Okay. Next step is to reformat and refactor the data. Bring it into line with expectations, and remove errors.

“””

**15/03/20**

Okay. Let’s think about my data tasks.

1. Build a map from TIPLOC to MIDAS
2. Commit current pickles.
3. Ensure that STANOX area is included in TIPLOC data (it should be in schedule data as well). We can one-hot encode STANOX area.
4. Combine weather and schedule records
5. Comb over schedule combination logic. Write it up for posteriority. I’d love to identify my discrepancy, but I don’t think it’ll be possible. Don’t care about verbosity; just accuracy.
6. Move extracted timetables to hard drive
7. Combine timetables and Darwin records
8. Move models pre-processing logic into one place
9. Double check and write-up forecast transmutation
10. Encode cyclical variables properly

Preprocessing is necessary. Three terms are generally used: scale, standardise, and normalise. They are defined here for clarity.

To *scale* is to change the range of a set of values, typically to 0 – 1. The shape of the distribution does not change.

To *standardise* is to change values so the standard deviation of a distribution equals one. Scaling is often implied.

To *normalise* is ambiguous. It is not used here.

**Take one train that I know…**

Take one train that I know – exploration data analysis. Check that sensor data marries with the timetable.

Just a day’s worth for a line that I know – confirm that it makes sense.

First predict whether it will be late at its end point.

Either system-wide or a single service.

Will my train be late getting to its destination. Terminus is a good place to start.

ML interesting with lots of features. Maybe a route is better.

Hastings to London Bridge -> late NOT late. > 5 mins.

Time of day, day of week, length of service, number of carriages, operator, weather, public holidays.

Are the features that I have predictive? What simple features could be derived. Build real-time index of delay.

Can just demonstrate that this is a very hard to model. Simplest set of features possible, show that model works. Do it for one line.

Monday – Friday to avoid engineering works. As simple as possible. Here’s some initial variables, box-and-whisker plots against delays.

Make sure that I understand the data.

Correlations on other services that day. Propose expansion. Check that there aren’t big holes in the data.

One train service on one day. A proper EDA across the whole line, and then for a particular service. Given a delay, ho

Is it a typical line? Average propensity to be late across train operators. Something that triggers financial impact. Average length of journey will be a predictive feature. Length should be possible.

Index of overall delays in the vicinity.

**Handling categorical data in machine learning**

Categorical data only takes a limited number of values. Categorical data for the dependent and independent variables are handled differently. Create an independent matrix (X) and dependent vector (Y).

One-hot encoding introduces a new dummy column for each value of a certain category. Usually done when there is no natural ordering between the categories (nominal). Features with some order associated with them are called ordinal. Continuous features are numeric variables with an infinite number of values between any two values (numeric or datetime).

Other such encodings include ordinal, binary, Helmert contrast, sum contrast, polynomial contrast, backward difference contrast, hashing, baseN, leave-one-out, and target encoding.

Independent = response variable.

Transformation into numeric labels followed by encoding.

Issues:

* categorical features may have a very large number of levels (high cardinality) where levels appear in a relatively small number of instances
* input categories must be numeric
* categorical data doesn’t contain the same contextual information for machines

Relationships between a categorical feature and continuous feature can be investigated via boxplot (df\_flights.boxplot('dep\_time','origin',rot = 30,figsize=(5,6))

Should also check frequency distribution of categories within the feature using value\_counts(). Add .count() for the number of distinct categories.

Dummy encoding produces one less category than one-hot, represented by a row of all 0s.

Effect coding uses -1s instead of 1s.

Make sure to typecast categorical features to a category dtype rather than an object.

Label encoding can be done using cat.codes. Numerical values may be misinterpreted by an algorithm.

One-hot is not very useful when there’s lots of categories: the curse of dimensionality. As dimensionality increases, the volume of the space increases so fast that the available data become sparse.

Typical rule of thumb: there should be at least 5 training examples for each dimension in the representation.

Binary encoding encodes categories are ordinal, converts those to binary code, and then splits the binary digits into separate columns.

Backwards difference encoding is a contrast coding system. The mean of the dependent variable for a level is compared with the mean of the dependent variable for the prior level.

Bin-counting sounds like it might be

**Handling dates in machine learning**

* Use the Pandas datetime field. This allows a datetime index to be constructed, if we so desired.
* Common offset aliases <http://pandas.pydata.org/pandas-docs/stable/timeseries.html#offset-aliases>
* Can use resample to convert data to a new frequency is the index is a datetime index.
* Then we could aggregate things on an hourly frequency.

https://www.datacamp.com/community/tutorials/categorical-data

<https://towardsdatascience.com/basic-time-series-manipulation-with-pandas-4432afee64ea>

# This would be done better on paper.

Apply for MIDAS open.

Dataset comparison to the Chinese data.

Lit review categorical data, forecasting prediction, timestamp. Email to Noura.

**INTRODUCTION**

Train delays impose a huge cost on passengers and operators, contributing to the inefficiency of train operations (Van Oort, 2011). A preliminary analysis of historical delay attribution data[[1]](#footnote-1) released by Network Rail (NR) for the 2018-2019 financial year (1st April 2018 – 31st March 2019) shows that 7.2 million incidents caused 35 million minutes of delay. Extrapolation from figures[[2]](#footnote-2) released by the National Audit Office (NAO) in 2008 suggests this cost a total of £2.5 billion in terms of time lost to passengers.

# I haven’t actually found a good definition of what a train delay prediction system is.

Indeed, 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 (Kecman, Corman, and Meng, 2015). Although the majority of current prediction systems utilise analytical models (Oneto et al, 2016), the focus of this paper will be data-driven models, and, in particular, machine learning techniques.

# Okay. 30 references. I should probably cut some down. And I want a meta-study on the value of including exogenous data in machine learning models as well. And I’d like to have actually read all of my papers, thoroughly. For now, I’m going to convert it into LaTeX and re-write it as I do. Wish me luck!

**DELAYS**

Though precise terminology differs, the literature agrees that there are two principal classes of delay (Olsson, 2004): primary (exogenous) and secondary (knock-on, consecutive).

A **primary delay** is “caused by external stochastic disturbances” (Oneto et al, 2016). The causes of primary delays are varied and numerous. Berger et al. (2011) lists the following causes: disruptions in the operations flow, accidents, malfunctioning or damaged equipment, construction work, repair work, and weather conditions. Network Rail[[3]](#footnote-3) additionally lists:

* Signals and points failures
* Severe heat (leading to buckled rails)
* Fatalities
* Flooding
* Landslips
* Leaves
* Snow and ice
* Vandalism and trespass

Milinkovic et al. (2013) add technical failures, lower-than-scheduled running speeds, prolonged alighting and boarding times of passengers, and bad weather conditions, citing Goverde (2010). The relative importance of these factors was explored by Olsson (2004), with the conclusion that, for double track lines, a key factor is the management of boarding and alighting passengers, and for single track lines, the management of train crossings. Goverde (2005) confirmed that non-negative arrival delays of passenger trains follow a negative exponential distribution. The distribution of train delays on the British railway network is accurately described by *q*-exponential functions (Briggs and Beck, 2007).

A **secondary delay** is “generated by operations conflicts” (i.e. primary delays) (Cerreto et al, 2016). As Berger et al. (2011) explain, primary delays induce a cascade of secondary delays of other trains which have to wait according to certain waiting policies between connecting trains.

Nor can these delay knock-ons be exactly forecast, as both Berger et al. (2011) and Milinkovic (2013) note, due to unpredictable influences: the length of primary delays, the timetable of the trains, and infrastructure (e.g. single or double track, station layouts, and interlocking). Additionally, drivers can attempt to reduce delays by driving faster than planned, or reducing their dwell time at stations, to make up for the delay. Daamen, Goverde, and Hansen (2009) further distinguish between two main classes of knock-on delay:

1. hindrance at conflicting track sections
2. waiting for scheduled connection in stations

Much of the literature focuses on one type or the other, or even on the problem of distinguishing between them, as in Lessan et al. (2019), where the key driver behind the implementation of a hybrid Bayesian network is this problem. Hansen et al. (2010) use historical train describer records to automatically classify delays into initial (primary) and consecutive (knock-on) ones. Wang et al. (2019) use a similar technique.

**TIMESCALE**

There are several classifications of timescale. Wang et al. (2019) divided predictions in short-term (i.e. predicted with real-time operating data) and long-time (a relatively long period; 3 days to 1 week in advance). Markovic et al. (2015) distinguish between the tactical level (in which models are applied to both timetabling and resource planning), the operational level (in which models are used for the real-time prediction of train delays), as reinforced by Corman and Goverde (2017). For the purposes of this project, *three* timescales will be defined: real-time, short-term, and long-term.

1. **Real-time**: typically modelled as a time series forecast problem (Oneto et al, 2016), real-time delay prediction models are *online*, i.e. updated as data on train movements becomes periodically available. Most data-driven models fall into this category. As Kecman, Corman, and Meng (2015) note, models for real-time traffic have so far focused on overcoming the great combinatorial complexity of train rescheduling, delay management, and rolling stock and crew scheduling.
2. **Short-term**: the long-term defined by Wang et al. (2019), this period is unusual in that it is bounded by exogenous data: the limited scope of weather forecasts. Wang et al. (2019) use forecasts accurate up to 10 days’ in advance: this project will use forecasts up to 5 days’ in advance. This is a relatively unexplored area, but much of methodology of both bounding timescales can be incorporated.
3. **Long-term**: in this timescale, models are applied to problems of timetabling and resource planning, and are typically tested for robustness using historical delay data (Markovic et al. 2015). For example, Yuan (2002, 2006) developed a delay prediction model that deals with stochastic behaviour, dependency of train delays, and delay propagation. to assess stability and punctuality of a published timetable against primary delays. According to Wang and Work (2015), the need for training data prevents the application of data-driven methods to scenario planning.

It is worth noting that it is possible to convert between timescales, though with diminished accuracy. In the case of converting to real-time to short-term, for example, the accumulated delay, and the run-time of the train, can be estimated.

**METRICS**

**Punctuality** is “a feature consisting in that a predefined vehicle arrives, departs, or passes at a predefined point at a predefined time” (Rudnicki, 1997). This definition, while laxer than that used by others (such as Gylee, 1994), allows greater flexibility in the measurement of punctuality. It also has the interesting effect that trains that arrive *early* cannot be considered punctual. However, the use of punctuality hides a lot of information, as Skagestad (2004) notes; consequently, Olsson (2004) uses reliability and variability. 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 measure reliability.

**Reliability** has several measures. Rietveld et al. (2001) list the following:

* The probability that a train arrives *x* minutes late (punctuality, as defined above)
* The probability of an early departure
* The mean difference between the expected arrival 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” (Olsson, 2004). First introduced by Noland and Polak (2002), where it is related to the distribution of arrival times for a train. A train that arrives the same amount of minutes behind schedule every day has low variability, but would not be considered punctual.

**RAILWAY TERMINOLOGY**

**Capacity utilisation** is “the extent to which the maximum theoretical capacity (for a given scenario) is being used” (Armstrong and Preston, 2017). However, as the International Union of Railways (IUC) notes (2004, leaflet 406, capacity), for any given section of railway: “capacity as such does not exist [and] railway infrastructure capacity depends on the way it is utilised”. Achievable capacity depends upon infrastructure characteristics, the performance characteristics and mix of trains using the route, the timetable in operation, and the target levels of reliability and punctuality to be achieved by timetabled services (Armstrong and Preston, 2017). Capacity utilisation is widely assumed to be positively correlated with the propensity of delays to propagate in a rail network (Olsson, 2004).

**Dwell time** is the time a train stands at the platform, usually for the purpose of allowing passengers to board or alight (Douglas, 2012). Puong (2000) asserts that dwell time is a key parameter of system performance, service reliability, and quality. Passenger volume is considered to be the main factor influencing dwell time (San and Masirin, 2016). Magadelaga et al. (2017), though their work focused on freight trains, found that the main influencing factors were crew shortages, locomotive imbalance and defective perway (i.e. rails).

**Headway** is a measurement of the distance or time between vehicles in a transit system. The minimum headway is the shortest such distance or time achievable by a system without a reduction in the speed of the vehicles.[[4]](#footnote-4) The shorter the minimum headway, the greater the capacity of a transit system.

**Free motion time** is the time a train spends in transit between two scheduled stops.

**Slack time** **(timetable allowance, time supplement)**: the amount of time a train can be delayed without causing another train to be delayed. The greater the capacity utilisation, the shorter the slack time.

**CLASSIFYING MODELS**

There are many classifications of railway delay prediction models, based on scope, model type, and solution methods (Markovic et al, 2015). Wang and Work (2015) distinguish between analytical, simulation, and data-driven models; Markovic et al. (2015) between analytic, multiple regression, and machine learning; Corman and Goverde (2018) into static (offline), dynamic (online) and deterministic or stochastic, depending on “the timespan between the time they are run, and the operations they aim to predict; and on how they tackle uncertainty, respectively”.

For the purposes of this literature review, and simplicity, models will be classified into analytical and data-driven (incorporating multiple regression and machine learning), with a particular focus on data-driven models. The timescale dimension previously discussed will be disregarded.

**EXISTING METHODS – ANALYTICAL MODELS**

An **analytical model** is “primarily quantitative or computational in nature and represents the system in terms of a set of mathematical equations that specify parametric relationships and their associated parameter values as a function of time, space, and/or other system parameters” (Friedenthal and Steiner, 2015). Current state-of-the-art train delay prediction systems use analytical models (Oneto et al, 2016): static rules, built by experts on railway infrastructure, and based on classical univariate statistics.

The paper also notes that the estimation accuracy of secondary delays does not improve. They also note that the potential delay is captured by a single parameter, and advocate separately modelling other contributing factors, such as geometry or weather, which other papers discussed do. Other approaches such as non-linear regression models or calibration via robust least squares (as per Italy’s default model).

Simplistic early models, such as those developed by Frank (1966) make overly restrictive assumptions about railway operations, e.g. no overtakes are allowed, the departing 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: Peterson (Over-the road transit time on a single line of railway, 1974), and Chen and Harker (Two moments estimation of the delay on single-track rail lines with scheduled traffic, 1994) included factors such as overtakes, different speeds, priority systems, and uncertainties associated with train departure time.

Further models have become increasingly advanced, incorporating stochastic approximation (Carey et al. Stochaistic approximation to the effects of headways on knock-on delays of trains, 1994), and the impacts of dispatching strategies on train delays and passenger waiting time.

Current state-of-the-art train delay prediction systems use analytical models (Oneto et al, 2016). Although new models are constantly being developed, a good example of a current *applicable* system is that developed by Berger et al. (2011), which is currently used in the German rail network.

**EXISTING METHODS – DATA DRIVEN (MACHINE LEARNING)**

Machine learning is “the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead”[[5]](#footnote-5). In the context of train delay prediction (and indeed, more generally), such models typically provide a better fit, but less straightforward interpretation, than statistical regression models (Markovic et al, 2015). The distinction between machine learning and data-driven models is increasingly becoming blurred.

Wang et al. (2015) acknowledges that the greatest difficulty to such data-driven approaches is data availability. However, the acceptance of data-driven approaches has greatly increased in the years since the paper was published, and the move towards open-source data feeds has increased both the availability and complexity of the data.

Several different machine learning techniques have been studied thus far in the literature (though this is by no means an exhaustive list): Bayesian networks (Lessan et al, 2019; Corman and Kecman, 2018), support vector regression (Markovic et al. 2015), random forests (Oneto et al, 2016), neural networks (Peters et al, 2005; Yaghini et al, 2013, Wang and Zhang, 2019), fuzzy Petri nets (Milinkovic et al. 2013), and extreme learning machines (Oneto et al. 2017). Attention will be given here to results; a thorough review of the exogenous data used follows.

**Peters et al. (2005)**

Perhaps the earliest work applying machine learning to delay prediction (though in the wider context of public transportation) is Peters et al. (2005), in which the authors develop a simple neural network for a small region of the Deutsche Bahn. Subsequent work on neural networks greatly expanded both the complexity of the model and the quantity of data, as in Yaghini et al. (2011).

**Yaghini et al. (2011)**

Yaghini et al. (2011) explore the implementation of a high-precision neural network to predict the delay of passenger trains in Iran. The authors used decision trees and multiple logistic regression models to evaluate the quality of the NN model and found it superior. The paper also tested three different encodings: normalised real number, binary, and binary set, and three different training methods: quick, dynamic, and multiple, and found the binary quick network to be the most accurate. The authors used a dataset of considerable size: roughly 180,000 trains over the course of four years, and a cumulative 5.4 million minutes of delay.

**Milinkovic et al. (2013)**

Milinkovic et al. (2013) used a novel model: a fuzzy Petri net (FPN). Petri nets are mathematical modelling tools used to analyse and simulate concurrent systems (Murata, 1989). The authors explored two separate FPNs. In the first, expert knowledge was used to define fuzzy sets and rules. In the second, an Adaptive Network Fuzzy Inference System (ANFIS) was trained on historical delay data and then replicated in an FPN.

Both were then tested with real data from a Belgrade station node. The ANFIS-defined FPN produced results within 5% of actual delay values for a subset of the data; slightly worse performance was observed for the expert-defined FPN.

**Markovic et al. (2015)**

Markovic et al. (2015) presented the first use of support vector regression (SVR, a close cousin of SVMs) to delay prediction problem. A comparison with artificial neural networks found that SVR outperformed the ANNs. Data for the analysis was again collected from Serbian Railways. The paper used the expert opinions of dispatchers as a key variable, which is discussed later. They noted that train category was an important factor, as higher-category passenger trains have right of way.

**Oneto et al. (2016)**

Oneto et al. (2016) proposed a train delay prediction system that considered both historical train running conditions and weather conditions. This inclusion is discussed later. The authors map the problem into a multi-variate regression problem and compare the performance of kernel methods, ensemble methods and feed-forward neural networks. They found it was possible to build a reliable and robust data-driven model based solely on historical data about train movements, with almost twice the performance of the existing system. The paper is also notable for its close co-operation with the Italian Railway Manager, which led to the development of a set of novel Key Performance Indicators (KPIs) to assess performance. The authors consider real-time prediction a time series forecast problem. They used Extreme Learning machines (ELM), Kernel Regularized Least Squares (KRLS) and Random Forests (RF) with the Nonparametric Bootstrap (NB) procedure to tune hyperparameters.

**Oneto et al. (2017)**

Oneto at al. (2017) generalised their earlier work to produce a dynamic data-driven train delay system. The performance of the system was tuned through the state-of-the-art thresholdout technique, which relies on differential privacy theory. They compare the performance of two implementations of shallow and deep extreme learning machines and find similar results to their previous paper. However, this scope of this paper must be emphasised: it lays the groundwork for the application of this system to a real-life rail network, rather than just academic work.

**Corman and Kecman (2018)**

Corman and 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

**Lessan et al (2019)**

Lessan et al (2019) 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%). This supports the hypothesis of Wang et al. (2019) of “key train delay stations” at which delays are considerably more likely to propagate, likely a result of an infrastructure bottleneck.

**Wang and Zhang (2019)**

Wang and Zhang (2019) have an unusual time constraint: that of the maximum acceptable resolution of a weather forecast, as discussed early. As discussed previously, their work does not fit in Markovic et al. (2019)’s real-time / operational / tactical trichotomy: rather, it lies between the first two. This is the region that this project intends to explore, so significant attention is paid to this paper in particular in the following section. The authors place emphasis on explaining the necessity of predicting delays in this window. The rationale is simple: passengers cannot easily compensate for delays in real-time. If a passenger can be informed that a train is *likely* to be delayed, they can make an informed choice on catching that train or altering their plans.

The authors present a relatively simple, but effective, gradient-boosted regression trees (GBRT) model. They note that a number of errors are likely due to stochastic errors and the limited timespan of their dataset – only 3 months’ worth of data was collected, which precludes the incorporation of the seasonality of weather.

**INCLUDING EXOGENOUS DATA**

It is widely accepted amongst machine learning practitioners that the greater the quantity of information available for the creation of a model, the greater the performance of the model will be (Oneto et al, 2016). Features can be engineered from existing features or exogenous data can be incorporated.

Data is **exogenous** if it is independent of other input data and the output data depends on it. The scope for inclusion is essentially limitless: any source of data which may affect railway dispatching operations is a viable candidate. In the context of delay prediction, the more of the causes of primary delay that can be incorporated into a model, the better it will perform. Prediction models tend to use either infrastructure (Markovic et al, 2015; Milinkovic et al, 2013) or weather (Oneto et al, 2016; Wang and Zhang, 2018; Oneto et al, 2017), though Oneto et al. (2017) also recommend using information about passenger flows and about railway asset conditions. This project plans to do all of the above.

Primary delay data analyses, such as those performed by Harris (1992), Gorman (2009), and Wen et al. (2017) reveal, in a rudimentary way, the use of proxies for variables. Harris (1992), investigating train punctuality in the UK and the Netherlands, used:

* train length (as a proxy for the number of doors to manage, and passenger demand)
* distance covered (as a proxy for the likelihood of encountering track defects and other technical / operational problems)
* the previous number of stops (as a proxy for cumulative delay resulting from passengers alighting and boarding)
* the age of the motive power unit (as a proxy for reliability; it is industry-held fact that a motive unit’s reliability declines after 20 years)
* track occupation (as a proxy for capacity utilisation the railway, and thus the likelihood of delays propagating)

Gorman (2009) investigated freight rail congestion delay. Factors were broken down into free running time (horsepower per ton, track topography, and slow orders) and congestion-related factors (meets, passes, overtakes, prior time periods’ train counts, total train hours, train spacing variability, and train departure headway). Primary congestion predictive factors were found to have the largest effect on congestion delay.

Wen et al. (2017) found that…

Wang et al. (2019) use three datasets: train schedule data, train delay data, and weather data, all collected from a three-month period between 1st January and 31st March 2018. The schedule data comprised 7172 trains and 2761 stations; each station was additionally geo-located to enable cross-referencing with weather data. Approximately 2.7 million delays were observed in this period, of which 37.4% involved high-speed trains.

Weather data was collected from 344 cities along the route in question, Beijing to Guangzhou. It is worth noting that the two are approximately 2200km apart, and so delays are of a magnitude not often seen in datasets with smaller geographic coverage. The fields were relatively simple, comprising the lowest and highest daily temperature, a categorical weather type, a categorical wind ranking (the Beaufort scale), and an air quality index.

Oneto et al. (2016) found that their random forest model, without the inclusion of weather data, roughly doubled the predictive capability of the original system. Including weather data increased this accuracy by approximately 10%, with the caveat that, the further ahead in the future the forecast was for (and thus the less accurate), the smaller this increase was.

Milinkovic et al (2013) defined their Fuzzy Petri Net (FPN) model by three input parameters: train category (freight, regional, or passenger), timetable influence, and distance travelled by the train. Timetable influence was used as a catch-all of sorts. 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.

Milinkovic et al (2013) defined FPN model by three input parameters:

* Train category (freight, regional, or passenger)
* Timetable influence
* Distance travelled by the train

Timetable influence was used as a catch-all of sorts. 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.

The study used train delay collected in July 2010 over several stations in Belgrade. 3710 trains were included in all categories: international passenger, domestic passenger, suburban and regional passenger, international freight, direct freight, pick-up freight, and other trains. In this period, 826 freight trains and 427 other trains were not registered in the predefined timetable.

The ANFIS model this project plans to replicate functioned best with the following input parameters: train category, arrival time at station, the travelled distance, and the infrastructure influence.

Markovic et al. (2015) used the expert opinion of dispatchers to estimate the likelihood of multiple factors along a rail line (single-tracking, reduced speeds, characteristics of block and interlocking systems, numbers of stations, stops, loops, road-rail level crossings, and junctions) causing a delay. The final estimate was obtained using the Delphi method. A higher score denoted better infrastructure conditions. Their input variables were:

* Passenger train category
* Scheduled time of arrival at station
* Infrastructure influenced defined by expert opinions
* Percent of journey completed distance-wise
* Distance travelled
* Time travelled
* Headway

They found a strong correlation between expert opinions and train delays. As Martin (2016) notes, in real-world train operations, delay prediction relies heavily on the experience and intuition of a local dispatcher, rather than a network-work computational instrument.

Markovic et al. (2015) used the expert opinion of dispatchers to estimate the likelihood of multiple factors along a rail line (single-tracking, reduced speeds, characteristics of block and interlocking systems, numbers of stations, stops, loops, road-rail level crossings, and junctions) causing a delay. A higher score denoted better infrastructure conditions. They found a strong correlation between expert opinions and train delays. As Martin (2016) notes, in real-world train operations, delay prediction relies heavily on the experience and intuition of a local dispatcher, rather than a network-work computational instrument.

The proposed primary dataset consists of historic delay attribution data gathered by Network Rail for the 2018-2019 financial year (1st April 2018 – 31st March 2019). The proposed delay dataset consists of all TRUST messages for the same period. The proposed weather dataset consists of Met Office MIDAS data, across the entirety of the UK, for the same period. Additionally, passenger volume statistics released by the ORR will be incorporated. This is, to the best of the author’s knowledge, the first time such a dataset has been collected in the UK, and likely the most comprehensive gathered to-date in the area of train delay prediction.

Much of the challenge of this dissertation lay in the construction of a dataset amenable to machine learning, a task commonly referred to as ETL (Extract-Transform-Load). The author had little knowledge, and no experience of the details of such a task.

# <https://networkrail.opendata.opentraintimes.com/>

All open data is stored here.

It’s a nicer mirror, for sure.

But if his is there, why does mine need to be? It doesn’t. We’ll have to continue with the ML. Fuck!

**Extraction**

This is the simplest stage. It is the process of collecting data, often from multiple and different sources.

MIDAS data is available under the Open Government License. This allows anyone to copy, publish, distribute, transmit, and adapt the licensed work, and to exploit it both commercially and non-commercially. The user of the licensed work has to acknowledge the source of the work and, if possible, provide a link to the OGL.

Network Rail data is ‘open’. I cannot find an actual license.

National Rail encourage the use and re-use of our feeds in their services. It is available under the Open Government License.

Would it be complete *enough*? Yes.

I am not allowed to alter the content of the feeds themselves, but I do not plan on doing so, so that’s okay.

Much of the relevant material is found in the appendices, but an overview is presented here to provide an understanding of the vagaries of the source data.

Extraction is the process of collecting data, often from multiple and different sources.

Darwin messages switch to a new schema in early 2019, version 16.

Should, for instance, a train be included if the first Darwin message regarding it is of type PASS, rather than DEPARTURE, as should be the case for a departing tra

The first stage was *downloading* the data. Peter Hicks, the founder of RealTimeTrains, and without whom this dissertation would not have been possible, has archived several feeds – SCHEDULE, TD, TRUST, and Darwin – for a number of years.

TD, TRUST, and Darwin all start on 2017-01-01, albeit with missing files several days (2017-01-03, 2017-01-04, 2017-06-11 for TD, 2017-01-03, 2017-01-04 for TRUST).

SCHEDULE starts earlier. However, it is very important with the SCHEDULE feed that no days are missed. Full SCHEDULE extracts are released every Friday at 01:00; on every other day, an *update* extract is provided, which must be applied the previous full extract each day. SCHEDULE only allows inserts up to 48 hours’ before a train is due to run – after which VSTP takes over – so a missing update file on a Saturday could invalidate schedules from Monday to the next full schedule extract on Friday.

SCHEDULE logs start on 2014-10-23. Full extracts are hosted only intermittently. However, this date, naturally, is of no interest to us.

There is a large range of missing schedule data between 2017-07-26 and 2018-03-29 (an update). The next full extract is on 2018-03-30, and so this provides a lower bound of our starting point. Conveniently, this is the start of the UK financial year, which corresponds to a dataset released by ORR pertaining to historical delay attribution, which at the time of writing was thought to be potential avenue of exploration, though this proved infeasible.

However, this range invalidated an ORR dataset, which captures how busy stations are, likely a useful factor.

A year’s worth of data was deemed necessary to capture the seasonality of train delays.

VSTP – Very Short Term Planning – trains are scheduled at most 48 hours in advance. They do NOT appear in the SCHEDULE feed. Instead, Darwin publishes their schedules when the trains are ‘activated’. A corresponding TRUST message may or may not also be published. Their schedules can, therefore, be reconstructed by parsing the Darwin log files. However, much of the metadata included in the SCHEDULE feed is missing, and it seemed unreasonable to attempt to coerce it into this format.

So those trains were discarded. This led to a problem parsing Darwin messages.

Each line in a Darwin log file is simply a Darwin message, of which there are eight types: Train Order, Forecast, Schedule.

**SCHEDULE**

SCHEDULE provides daily extracts and updates of train schedules from the Integrated Train Planning System (ITPS) in CIF format.

The ITPS was introduced in 2010; a report was conducted by the ORR into the resulting disaster (<https://orr.gov.uk/__data/assets/pdf_file/0004/5746/ITPS-ORR-report.pdf>).

Schedules can be uniquely identified by a UID, the schedule start date (SSD), and STP indicator, which defines which schedule should be given priority on a given day.

The creation of a high-quality, open-source dataset could itself have been a valuable dissertation contribution.

Wouldn’t that be cool? I know the dataset very well, after all, I wouldn’t even need to do anything else.

To retrieve the correct schedule for a given UID on a given day, all schedules with that UID must be retrieved. Those not active on the *day* in question – as defined by the *days\_run* field – can be discarded, as well as those with a start date in the future. From those remaining, the schedule with the lowest STP indicator is the correct one, in order `C`, `O`, `V`, `P`. `C` is a planned cancellation. `O` is an overlay from STP. `V` is a variation from LTP. `P` is a permanent, base schedule from

An existing repository (<https://github.com/jhumphry/ukraildata_etl>) was used as an inspiration for a custom CIF parser configured for Peter Hick’s update files.

There are several feeds which were unfortunately not logged, such as the VSTP, and, more interestingly, the TSR (Temporary Speed Restrictions) feed, which details temporary reductions in permissible speeds across the rail network. Fortunately, this data is captured in field in SCHEDULE, namely engineering\_allowance.

Trains from VSTP were therefore disregarded. This, in itself, likely biased the dataset. It seems reasonable to assume that these ‘rushed’ trains are prone to delay. The opposite could also be argued: by their special nature, they are given priority – but we cannot be sure. Any UID in Darwin NOT found in the day’s schedule was discarded.

To ameliorate these concerns, a daily database was defined. For each day, only the ‘winning’ schedule was stored.

Despite my best efforts, there is some data missing. I routinely found that my days (generated by applying updates to full extracts) were smaller in size than ‘full’ database extracts generated for the same day.

Eventually, it was decided to simplify the problem.

Network Rail defines x routes:

* Anglia
* East Midland
* Kent
* London North Eastern
* London North Western
* London Northwestern
* London Overground
* Midland and Continental
* Scotland
* South East
* South West Main Line
* Sussex
* Wales
* Wessex
* Western

Much of the challenge of this dissertation lay in the construction of a dataset amenable to machine learning, a task commonly referred to as ETL (Extract-Transform-Load). The author had little knowledge, and no experience of the details of such a task.

# <https://networkrail.opendata.opentraintimes.com/>

All open data is stored here.

It’s a nicer mirror, for sure.

But if his is there, why does mine need to be? It doesn’t. We’ll have to continue with the ML. Fuck!

**Extraction**

This is the simplest stage. It is the process of collecting data, often from multiple and different sources.

MIDAS data is available under the Open Government License. This allows anyone to copy, publish, distribute, transmit, and adapt the licensed work, and to exploit it both commercially and non-commercially. The user of the licensed work has to acknowledge the source of the work and, if possible, provide a link to the OGL.

Network Rail data is ‘open’. I cannot find an actual license.

National Rail encourage the use and re-use of our feeds in their services. It is available under the Open Government License.

Would it be complete *enough*? Yes.

I am not allowed to alter the content of the feeds themselves, but I do not plan on doing so, so that’s okay.

Much of the relevant material is found in the appendices, but an overview is presented here to provide an understanding of the vagaries of the source data.

Extraction is the process of collecting data, often from multiple and different sources.

Darwin messages switch to a new schema in early 2019, version 16.

Should, for instance, a train be included if the first Darwin message regarding it is of type PASS, rather than DEPARTURE, as should be the case for a departing tra

The first stage was *downloading* the data. Peter Hicks, the founder of RealTimeTrains, and without whom this dissertation would not have been possible, has archived several feeds – SCHEDULE, TD, TRUST, and Darwin – for a number of years.

TD, TRUST, and Darwin all start on 2017-01-01, albeit with missing files several days (2017-01-03, 2017-01-04, 2017-06-11 for TD, 2017-01-03, 2017-01-04 for TRUST).

SCHEDULE starts earlier. However, it is very important with the SCHEDULE feed that no days are missed. Full SCHEDULE extracts are released every Friday at 01:00; on every other day, an *update* extract is provided, which must be applied the previous full extract each day. SCHEDULE only allows inserts up to 48 hours’ before a train is due to run – after which VSTP takes over – so a missing update file on a Saturday could invalidate schedules from Monday to the next full schedule extract on Friday.

SCHEDULE logs start on 2014-10-23. Full extracts are hosted only intermittently. However, this date, naturally, is of no interest to us.

There is a large range of missing schedule data between 2017-07-26 and 2018-03-29 (an update). The next full extract is on 2018-03-30, and so this provides a lower bound of our starting point. Conveniently, this is the start of the UK financial year, which corresponds to a dataset released by ORR pertaining to historical delay attribution, which at the time of writing was thought to be potential avenue of exploration, though this proved infeasible.

However, this range invalidated an ORR dataset, which captures how busy stations are, likely a useful factor.

A year’s worth of data was deemed necessary to capture the seasonality of train delays.

VSTP – Very Short Term Planning – trains are scheduled at most 48 hours in advance. They do NOT appear in the SCHEDULE feed. Instead, Darwin publishes their schedules when the trains are ‘activated’. A corresponding TRUST message may or may not also be published. Their schedules can, therefore, be reconstructed by parsing the Darwin log files. However, much of the metadata included in the SCHEDULE feed is missing, and it seemed unreasonable to attempt to coerce it into this format.

So those trains were discarded. This led to a problem parsing Darwin messages.

Each line in a Darwin log file is simply a Darwin message, of which there are eight types: Train Order, Forecast, Schedule.

**SCHEDULE**

SCHEDULE provides daily extracts and updates of train schedules from the Integrated Train Planning System (ITPS) in CIF format.

The ITPS was introduced in 2010; a report was conducted by the ORR into the resulting disaster (<https://orr.gov.uk/__data/assets/pdf_file/0004/5746/ITPS-ORR-report.pdf>).

Schedules can be uniquely identified by a UID, the schedule start date (SSD), and STP indicator, which defines which schedule should be given priority on a given day.

The creation of a high-quality, open-source dataset could itself have been a valuable dissertation contribution.

Wouldn’t that be cool? I know the dataset very well, after all, I wouldn’t even need to do anything else.

To retrieve the correct schedule for a given UID on a given day, all schedules with that UID must be retrieved. Those not active on the *day* in question – as defined by the *days\_run* field – can be discarded, as well as those with a start date in the future. From those remaining, the schedule with the lowest STP indicator is the correct one, in order `C`, `O`, `V`, `P`. `C` is a planned cancellation. `O` is an overlay from STP. `V` is a variation from LTP. `P` is a permanent, base schedule from

An existing repository (<https://github.com/jhumphry/ukraildata_etl>) was used as an inspiration for a custom CIF parser configured for Peter Hick’s update files.

There are several feeds which were unfortunately not logged, such as the VSTP, and, more interestingly, the TSR (Temporary Speed Restrictions) feed, which details temporary reductions in permissible speeds across the rail network. Fortunately, this data is captured in field in SCHEDULE, namely engineering\_allowance.

Trains from VSTP were therefore disregarded. This, in itself, likely biased the dataset. It seems reasonable to assume that these ‘rushed’ trains are prone to delay. The opposite could also be argued: by their special nature, they are given priority – but we cannot be sure. Any UID in Darwin NOT found in the day’s schedule was discarded.

To ameliorate these concerns, a daily database was defined. For each day, only the ‘winning’ schedule was stored.

Despite my best efforts, there is some data missing. I routinely found that my days (generated by applying updates to full extracts) were smaller in size than ‘full’ database extracts generated for the same day.

Eventually, it was decided to simplify the problem.

Network Rail defines x routes:

* Anglia
* East Midland
* Kent
* London North Eastern
* London North Western
* London Northwestern
* London Overground
* Midland and Continental
* Scotland
* South East
* South West Main Line
* Sussex
* Wales
* Wessex
* Western

**Aims of the literature survey**

* To show that you can do a proper literature survey
* To determine the state-of-the-art in project areas
* To acquire a set of useful techniques / methods / algorithms to make a strong start

**Include:**

* Existing theories about the topic which are accepted universally
* Books written on the topic, both generic and specific
* Research done in the field, usually in the order of oldest to latest
* Challenges being faced, and ongoing work, if available

Should be a quick summary of the state of the art on a subject.

Define terms and give references to where you found the definition.

The next four pages should be used to select themes. Then give a succinct, careful description of each of these as if you were addressing an experienced computer scientist who was not an expert in the area. Use the last pages to assess the themes, draw some conclusions, and show how you’ll use it in your project.

Two-way traffic on a single line of railway (Frank, 1966)

Causal Analysis of Railway Running Delays Cerreto, Fabrizio; Nielsen, Otto Anker; Harrod, Steven; Nielsen, Bo Friis

**A hybrid Bayesian network model for predicting delays in train operations (Lessan et al, 2019)**

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.

**Stochastic modelling of train delays and delay propagation in stations (Yuan, 2006)**

A delay prediction model that deals with stochastic behaviour, dependency of train delays, and delay propagation to assess stability and punctuality of a published timetable against primary delays. [Replicated in Yuan et al, 2002]

Developed an important stochastic model for train delays and train propagations in stations. The most important contribution is an innovative analytical probability model that accurately predicts the knock-on delays of trains, including the impact of train punctuality at stations on the basis of an extension of the blocking time theory of railway operations to stochastic phenomena.

**Propagation of train delays in stations (Yuan et al, 2002)**

A delay prediction model that deals with stochastic behaviour, dependency of train delays, and delay propagation to assess stability and punctuality of a published timetable against primary delays. [Replicated in Yuan et al, 2002]

**Predictive reasoning and machine learning for the enhancement of reliability in railway systems (Martin, 2016)**

In real-world train operations, delay prediction relies heavily on the experience and intuition of a local dispatcher rather than a network-wide computational instrument.

Proposed a prototype rail advisory system that applies a series of predictive reasoning and machine learning tools to predict the effects of various disruptions.

**Real-time optimal train regulation design for metro lines with energy-saving (Zhang et al, 2018)**

Delays should be predicted and compensated in time, otherwise there may be a disruption or domino effect of the propagated delays.

**Train delay evolution as a stochastic process (Kecman, Corman, and Meng, 2015)**

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]

A model for real-time prediction of train delays using Bayesian reasoning. Used two months of historical traffic realisation data from the Swedish infrastructure manager in a simulated real-time environment. Predictions are reliable for up to 30-min horizons. Main assumption, however, is that the train orders and routers within the prediction horizon are known, which is often not the case IRL.

Uncertainty modelled based on a Markov stochastic process.

” (Friedenthal and Steiner, 2015) A Practical Guide to SysML, 2015).

Accurate prediction of train delays is an important requirement for proactive and anticipative real-time control of railway traffic and transport.

Valid estimates of arrival and departure times are therefore important for preventing or reducing delay parapgation, manainging connections,

Achieved prediction results of 71%.

**Stochastic prediction of train delays in real-time using Bayesian networks (Kecman, Corman, Peterson, and Joborn, 2015)**

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.

**Statistical investigation on train primary delay based on real records: evidence from Wuhan-Guangzhou HSR (Wen et al, 2017)**

Many uncertainties may arise from these subsystems that can disturb the planned activities and operations, resulting in unexpected delays.

**Service reliability and urban public transport design service reliability (Van Oort, 2011)**

Train delays impose a huge cost on passengers and operators, contributing to the inefficiency of train operations.

**Stochastic delay propagation in railway networks and phase-type distributions (Meester et al, 2007)**

The underlying problem is related to the delay prediction practice that has received considerable attention due to its vital importance to train operations management and passenger information provision.

Pros and cons, and then how they got through it, suggestions for further research. Split it categories, use paper as evidence into the story. Not a blow-by-blow. Story of the problem.

Puong (2000) – Dwell time model and analysis for the MBTA red line

Douglas (2012) – Modelling train and passenger capacity

**Data driven approaches for passenger train delay estimation (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. # Code can be found here: <https://github.com/Lab-Work/TrainDelayEstimation_IEEEITSC>

Predicts train delays at each station before the current trip starts based on the delay recorded in the past trips.

**A structured model for rail line simulation and optimisation (Peterson et al, 1982)**

Proposed a structured model for the movement of trains over the train line which supports arbitrary numbers of trains with different speeds and priorities on single or multiple tracks on rail networks with sidings, switches and cross-overs.

**A simulation modelling methodology for analysing large complex rail networks (Dessouky et al, 1995)**

Proposed simulation models that are able to simulate train movements over single and double track lines, junctions, terminals, and model rail networks consists of multiple trackage configurations and speed limits. [Replicated in Dessouky et al, 2004]

**Modelling train movements through complex rail networks (Dessouky et al, 2004)**

Proposed simulation models that are able to simulate train movements over single and double track lines, junctions, terminals, and model rail networks consists of multiple trackage configurations and speed limits. [Replicated in Dessouky et al, 2004]

**Predicting on-time line-haul performance in scheduled railroad operations (Hallowell et al, 1996)**

Simulation models have also been used to calibrate the parameters of the analytical models for delay estimation,

**Statistical estimation of railroad congestion delay (Gorman, 2009)**

Considered factors such as horsepower per ton, track geometry, train priorities, meets, passes, overtakes, and train spacing variabilities for a first-class freight railway.

**Online Train Delay Recognition and Running Time Prediction (Hansen, Goverde, van der Meer, 2010)**

Compared the actual blocking time with the predicted blocking time based on train delay data. Developed an online model to accurately predict the running time between two stations considering route conflict, vehicle type, weather conditions, and other factors.

Trained an online model using historical track occupation data and implement it on a section of the Dutch railway corridor.

Event graphs were used to forecast running and arrival times.

“Based on (timed event) graph modelling with train routes and process times as input parameters

This paper describes train operations as a set of processes (which take a certain amount of time), and events, which form the beginning or end of a process. This may be insufficient. Events are divided into three types: train departures, train arrivals, or train passages at stations.

The data used is slightly lower-level than that used for this dissertation, with a closer correspondence to the TD datafeed than the TRUST feed.

This paper defines two mechanisms by which a departure delay may influence the running time of the affected train:

1. The train driver will try to use the running time supplement by running at the maimum speed allowed in order to reduce the delay
2. At intermediate stops delay trains require less time (if at all) waiting for their scheduled departure time.

“Dwell time is the duration that trains must spend for certain activities to be carried out in a particular yard station”

“System factors include signal communication failure, cable failure, and power outages”. Such failures are harder to predict using the dataset to hand, as it contains no information on infrastructure condition.

Uses data-mining and analysis of standard track occupation data to compute accurate actual train delays at stations. Delays can be automatically classified into initial and consecutive via comparison of actual with scheduled blocking times.

The distributions of running times and dwell times of each line and direction can be estimated conditional on the amount of original delays, route conflicts, and factors such as type of rolling stock, peak hours, or weather conditions.

The running times of actual trains until the next downstream station can be accurately predicted at a high precision. Parameters calibrated and tested in the Dutch railway corridor Rotterdam – the Hague

Standard registration of train punctuality in current railway registration performance reports is based on automatic registration of revealed differences between the arrival times of scheduled and realised train trips at a number of stations in the network. Threshold for delays differs between networks.

Performance contracts between the government and TOCs specify some minimal required punctuality and include fines in case of bad performance. A prerequisite for the non-discriminatory assignment of delays and allocation of fines is a clear and objective determination of the cause and responsibility of train delays.

Current system doesn’t distinguish between original and knock-on delays.

Describes a recently developed prototype software tool for automatic distinct assignment of initial and consecutive delays and accurate online prediction of running times until the next stations. Model obtains its parameters by statistical analysis of historical track occupation data and takes into account possible downstream conflicts.

Train operations are affected by driver behaviour, passenger volumes, weather conditions, etc. Data can be exploited to find stochastic distributions and dependencies amongst delays, process times, and categorical factors.

Model is based on (timed event) graph modelling with train routes and process times as input parameters.

Requires a topological description of the underlying rail network including the measurement locations of actual train positions. Tool obtains train path realisations on a track section level with a precision of one second.

Graph model is based on the topological description of the railway network according to the train described configurations. Divided into track sections by track-free detection devices. Microscopic level is used to find statistical dependencies and parameter estimates offline. Online prediction model computes scheduled station events (arrival, departures, and through passages) and the train position steps from the describers.

Processes: take a certain amount of time (train running times, dwell times, waiting times caused by conflicting train routes, etc).

Events: form the beginning or end of a process (FPN style!) Three types of event: train departures, arrivals, or passages at stations.

Dependencies between two are represented as timed event graphs. Modelled by an arc connecting dependent events I and j; weight represents the minimum time duration between events i and j. Can thus be modelled by a set of events and a list of arcs. By mining occupation data, a timed event graph of all recorded train operations can be generated.

The scheduled running time between two main stations consists of four components: minimum running time between scheduled stops, dwell time at intermediate scheduled short stops, running time supplement, and scheduled waiting time.

Departure delay may influence the running time of that train: the train driver may use the running time supplement by running at the maximum speed allowed in order to reduce the delay, and at intermediate stops delay trains required less time waiting for their scheduled time.

Scheduled running time is calculated until the actual arrival (i.e. standstill at the platform) of the train, whereas the observed running times are calculated using the section occupation time.

When trains run ahead of their schedule, they have to wait on their scheduled departure time at station stops. Therefore, delays have a large influence on dwell times of trains at stations.

Distinction between trains that actually experienced their minimum dwell time and those that had to wait for their departure time, leading to longer dwell times. Trains experience their minimum dwell time only if they arrived with some delay. So use scheduled dwell time minus 60 seconds as a limit for this minimum arrival delay.

Statistical analysis revealed a linear dependency between delays and running times for small delays up to some threshold, and in particular for trains with large running time supplement. More advanced minimum running time estimations will therefore be used as a piecewise linear function.

**Train delay analysis and prediction based on big data fusion (Wang and Zhang, 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)

**Identification of delay factors that affect high dwell time of freight trains (Magadalega et al, 2017)**

**Alighting and Boarding Times of Passengers at Dutch Rail Stations (Wiggenraad, 2001)**

Considers the distribution of passengers along the platform, alighting and boarding times, type of station and train service, vehicle characteristics, and period of the day.

**Influencing factors on train punctuality – results from some Norwegian studies (Olsson et al, 2004)**

Found that the punctuality of trains is correlated with the number of passengers, occupancy ratio, departure punctuality, and operational priority rules.

High capacity utilisation is widely assumed to reduce punctuality. Studies from Oslo show that capacity utilisation alone cannot explain all variations in punctuality during the day. The number of passengers in the train is highlighted as an additional explanation factor. The key success factor for punctuality on local and regional trains in congested areas is the management of boarding and alighting passengers.

On single track lines, such as long distance lines and regional lines not in contested areas, the key success factor seems to lie in theSchwa management of train crossings.

Punctuality is usually related to deviations, primarily negative, from the timetable. Often used as a discrete measurement, related to a predefined level of accepted deviation. Even though it normally refers to trains running behind schedule, trains running *ahead* of schedule are not punctual either. Usually presented as a percentage.

Delay is a continuous measurement of the negative (late) deviation from the timetables. A train can be more or less delayed.

Though punctuality is most commonly measured at the final destination, punctuality can relate to any point along the train route.

Unreliability and variability enhance the understand of punctuality. Unreliability can be used when discussing deviations from the official timetable.

The PPM measure used in the UK combines figures for punctuality and reliability into a single performance measure.

**Capacity utilisation and performance at railway stations (Armstrong and Preston, 2017)**

Should come in very useful:

<https://www.sciencedirect.com/science/article/pii/S2210970617300409#bib23>

**(Harris, 1992)**

Used least-squares multiple linear regression to analyse how a selected set of variables affects punctuality in the UK and the Netherlands. Variables were train length, previous number of stops, distance covered, age of motive power unit, and track occupation. Length of train (measured by number of carriages) was chosen because it assumed speed characteristics and that there are more doors to manage. Train length is also used as an indicator of passenger demand. The previous number of station stops was included in the analysis as a potential cause of delays due to boarding and alighting passengers. Chosen because the likelihood of encountering track defects and other technical / operational problems increases with the distance traversed. Age of the motive power unit was included based on experience that these become increasingly unreliable after 20 years or so. Track occupation: the busier a railway is, the more likely I is that one delay will cause problems for other services.

**Gibson et al. (2002)**

Use exogenous delays to mean primary delays: caused by direct influence on the train.

(**Rietveld et al, 2001)**

List a number of different measurements of reliability (or punctuality) such as:

* The probability that a train arrives x minutes late
* The probability of an early departure
* The mean difference between the expected arrival 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

**(Gylee, 2004)**

Punctuality is ‘the ability to achieve a safe arrival at a destination to an advertised timetable’

**(Rudnicki, 1997)**

Punctuality is ‘a feature consisting in that a predefined vehicle arrives, departs, or passes at a predefined point at a predefined time’.

**(Carey, 1999)**

Distinguishes between exogenous delays (those due to events such as failure of equipment or infrastructure, delays in passengers boarding or alighting) and knock-on delays (those due to exogenous delays and the interdependence in the schedule).

**Causal analysis of railway running delays (Cerreto et al, 2016)**

Found that train infrastructure layout, the performance and reliability of train vehicles, and the train stopping time at stations will affect the train delay time, the minimum running time, and the buffer time of trains.

**Extreme weather impacts on freight railways in Europe (Ludvigsen et al, 2014)**

Correlation between severe weather and equipment failure. Proposed a train delay analysis model.

**Determining the causes of train delay (Dingler et al, 2010)**

Used a railway traffic controller (RTC) to classify and quantify train delays and identify the factors causing train delays. Relationships among speed, siding dwell time, and train delay were discussed. Methods of increasing train operation speed and the number of tracks were proposed to reduce train delay time.

**Train delay prediction systems: a big data analytics perspective (Oneto et al, 2018)**

Used machine learning algorithms and statistical tools to construct a train delay prediction system (TDPS) for large-scale railway networks.

**Railway passenger train delay prediction via neural network model (Yaghini et al, 2013)**

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 multi-model 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 .

**Stochastic approximation to the effects of headways on knock-on delays of trains (Carey et al, 2007)**

Developed a simple stochastic model for knock-on train delays (a delay caused by other trains).

**Running times on railway sections with heterogenous train traffic (Huisman et al, 2001)**

Developed a stochastic model to predict the delays of trains with different speeds. Can capture both scheduled and unscheduled movements. A case study of a railway section in the Dutch railway network illustrates the practical value of the model, both for long and short-term railway planning.

**Prediction of delays in public transportation using neural networks (Peters et al, 2005)**

Developed an intelligent delay predictor model for real-time delay monitoring and timetable optimisation in the range of train networks. The system is responsible for processing existing delays in the network to generate delay predictions for dependent trains in the near future. Rules-based system was used as a comparison to the specially developed neural network in order to evaluate the accuracy and faculty of abstraction of such an artificially intelligent component.

**Dealing with stochastic dependence in the modelling of train delays and delay propagation (Yuan, 2009)**

Developed a model that deals with stochastic dependence in the modelling of train delays and delay propagation. The proposed model can be used in assessing timetable stability and predicting train punctuality given primary delays. Model validation reveals that the delay estimates match with real-world data very well.

**Modelling trains delays with *q*-exponential functions (Briggs and Beck, 2007)**

Demonstrated that the distribution of train delays on the British railway network is accurately described by *q*-exponential functions. Used data on departure times for 23 major stations for September 2005 – October 2006.

**Non-discriminatory automatic registration of knock-on train delays (Daamen, Goverde, Hansen, 2009)**

Proposed a method to predict knock-on delays in an accurate and non-discriminative way. Two main classes of knock-on delays are distinguished: hindrance at conflicting track sections and waiting for scheduled connections in stations.

**Hybrid model for prediction of bus arrival times at next station (Yu et al, 2010)**

Proposed a hybrid model based on support vector machines (SVM) and the Kalman filtering technique to predict bus arrival times. The SVM model predicts the baseline travel times on the basis of historical trips, using time-of-day, weather conditions, route segment, the travel times on the current segment, and the latest travel times on the predicted segment. The Kalman filtering-based dynamic algorithm uses the latest bus arrival information, together with the estimated baseline travel times, to predict arrival times at the next point. Results show that this model is feasible and generally provides better perf

ormance than ANN-based models.

**Murata 1989: Petri nets: properties, analysis, and applications**

**Jang 1993: Adaptive-network-based fuzzy inference systems (ANFIS)**

**A fuzzy Petri net model to estimate train delays (Milinkovic et al, 2013)**

Train delays are commonly used parameter for determining timetables and solving infrastructure problems. Also important for train dispatching the railway traffic operations. Unpredictability of delays makes efficient planning of railway operations very difficult.

Primary delays are delays caused by external stochastic disturbances. When they develop inside the observed network, they are called original delays. If the buffer times between trains are less than the length of the primary disturbance, delay is propagated to other trains. This creates knock-on or secondary delays. Very difficult to calculate and predict because they depend on the length of primary delays, the timetable of the trains, and the infrastructure (such as single or double track, station layouts, and interlocking) (Hansen et al, 2010)

Primary delays are caused by technical failures, lower-than-scheduled running speeds, prolonged alighting and boarding times of passengers, and bad weather conditions.

Identify three common approaches to determining train delays: analytical methods, micro-simulation methods, and statistical analyses based on empirical data.

The authors also caution that, in systems with many possible sources of disruption, and a relatively high probability of external influences that could induce primary delays, it is difficult to find a relationship that can be used to calculate train delays.

FPN model with characteristics of hierarchy, colour, time, and fuzzy reasoning are used to simulate traffic processes and train movements in a railway system. Used expert knowledge to define fuzzy sets and rules when lacking train data.

Used an Adaptive Network Fuzzy Inference System when historical data n train delays were available, which is used to train the neuro-fuzzy ANFIS model. After it was verified, it was replicated by a Fuzzy Petri net, and validating by animating train movement and plotting the time-distance graph of the trains. The FPN model was tested on part of the Belgrade railway node.

Modelling the propagation of train delays focuses on specific track layouts, the signalling and train protection system, and the timetable design.

The module used to calculate primary train delays based on fuzzy logic was the first subsystem that the toen entered. The structure, relations, rules, and weights of this module were defined by experts’ knowledge when no historical data was present. Their experience was use to define input variables, the rules bases, and output variables, with the Delphi method used to gather data on the causes of delays and average delays, collected via questionnaire.

**Primary delays**: delays caused by external stochastic disturbances, such as technical failures, lower-than-scheduled running speeds, prolonged alighting and boarding time of passengers, and bad weather conditions. Distribution of primary delays can be obtained by a statistical analysis of existing empirical data (aha! That’s what I could use that dataset for!) With this knowledge, what could we do?

Three common approaches: analytical methods, micro-simulation methods, and statistical analyses.

Simulations require data regarding the infrastructure, timetable, and train performance (where can I find infrastructure data?)

Statistical analysis of arrival delay data suggests that many factors influence train delays.

Fuzzy logic: a mathematical tool to model traffic processes that are distinguished by subjectivity uncertainty, ambiguity, and imprecision.

Input categories were: train category, time of arrival at the station, the distance travelled, and the infrastructure influence.

HLPNs have much more modelling power, though he same computational power, as a standard PN.

FPN is used to estimate knock-on delays. Primary delays were calculated in the FPN subsystem module as inputs to the simulation model and secondary delays were delays caused inside the model due to the propagation of primary delays.

Places = track sections

Transitions: conditions for train movement

Tokens: trains

Categories: freight, regional, passenger

The paper considers three categories of trains present in the Belgrade rail network: freight, regional, and passenger. Network Rail has similar TSPEED categories: Passengers and parcels in WTT, Freight trains in WTT, Trips and agreed pathways, and Special Trains (Freight or Passenger).

Section occupancy time depends on the train length, train acceleration and deceleration, as well as on its maximum speed.

Principles of traffic organisation and train movement:

* Trains can be dispatched from the station if the exit signal allows train movement and the protective path signals are set to forbid movement
* The train can occupy the next station if the next two block sections are free
* When entering the station, the condition needed to create the route is that all sections on the route must be free.

Model was defined by three input parameters: the train category, the timetable influence, and the distance travelled by the train. The timetable influence included the influence of operation time, the types of locomotives, local conditions, technological solutions, principles for safety and signalling, and weather conditions. Long travelling distances increased the probability of train delays.

Passenger trains operated according to the pre-defined timetable, and freight trains operated according to the timetable. Moments of freight trains arrival were generated by a random generator.

The best ANFIS model has input data that were defined by the following parameters: the train category, the arrival time at the station, the travelled distance, and the infrastructure influence.

Pairs of input-target data were used to calculate the membership functions of the Takagi-Sugeno type fuzzy logic system.

Delays of trains greater than 60mins were not predicted as precisely because the occurrence of those delays is smaller, and so are more difficult to forecast.

When they develop inside the observed network, they are called original delays. If the buffer times between train are less than the length of the primary disturbance, delay is propagated to other trains (the so-called train delay threshold defined in Big Data Fusion).

This results in knock-on or secondary delays. These are very hard to calculate or predict, as they depend on the length of primary delays, the ti

**Analysing passenger train arrival delays with support vector regression (Markovic et al, 2015)**

They found support vector regression to outperform artificial neural networks.

Reducing delays is of great importance to train operators and desirable to passengers.

Real-time prediction of train delays was used to detect instabilities in the timetable and retrieve a feasible train schedule.

A number of prediction models have been developed in the literature which can be classified by their scope, model types and solution methods.

Presented a comparison between the performance of support vector regression and neural networks for analysing passenger train arrival delays and the influence of infrastructure on arrival delays. Uses numerous test instances to show that SVG outperforms other models in predicting arrival delays.

Reducing train delays in an economically viable way would be highly desirable. So: establish a functional relationship between train delays and various characteristics (infrastructure, timetables, and trains).

Such a functional relationship would be helpful in tactical planning, e.g. investment planning. Managers are interested in reducing train arrival delays at a station by investing in the infrastructure, e.g. track renewal to increase the speed profile along some routes, building an overpass to avoid train conflicts, or introducing more advanced signalling systems that resolve rights of way more effectively.

Simulation analysis:

This paper focuses primarily on the effect of *infrastructure* on train *arrival* delays. They note several ways infrastructure can be improved to reduce train arrival delays:

* Track renewal, to increase the speed profile along some routes
* Building an overpass to avoid train conflicts
* Introducing more advanced signalling systems that resolve rights-of-way more efficiently

Analytic models. Probability distributions of train delays are derived from track occupancy and release records. The obtained distributions of the underlying random variables are then used within stochastic queueing models. The model is then used to compute delay propagation at platform tracks and junctions. For the train arrival delays, different distributions are considered, including Weibull, chi-squared, Erlang and log-normal, but mostly exponential.

Train arrival delays develop due to irregularities in internal sources within the railway system (e.g. infrastructure, train operators, traffic management), as well as exogenous sources (e.g. weather conditions, environment).

Initial models were analytic (Schwanhaeusser, 1974), confirmed that non-negative arrival delays of passenger trains follow a negative exponential distribution.

The work most closely aligns with ‘long-term’ (‘tactical’, in the words of the authors): rail infrastructure changes little day-to-day in the same manner that weather does.

They consider real-time delay prediction to be the “operational level”.

Timetables are tested for robustness using the probability distribution of process durations which are derived from historical traffic realisation data.

They also found superior results for a support vector machine (SVM).

The paper used the expert opinions of dispatchers to estimate the likelihood of multiple factors along a rail line (single-tracking, reduced speeds, characteristics of block and interlocking systems, numbers of stations, stops, loops, road-rail level crossings, and junctions) causing a delay.

They noted that train category was an important factor, as higher-category passenger trains have right of way.

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)

We will definitely need to integrate schedule data. Let’s hop88e that I can find it.

Negative coefficient for infrastructure, as a higher score denote better conditions.

Aggregate 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) into one variable, with value determined through the expert opinion of five dispatchers, with final estimate obtained using the Delphi method.

**An Identification of delay factors that affect high dwell times of freight trains (Magadagela, 2017)**

**Influencing factors on train punctuality (Olsson NOE)**

**Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data (Oneto et al, 2016)**

Proposed a train delay prediction system that considered both historical train running conditions and weather conditions. Weather conditions may affect passenger flow and therefore the stopping time of trains.

For any checkpoint *C*, a train should arrive at time t^CA and depart at time t^CD, as defined in the timetable. Timetables typically have 30-second precision. The actual arrival time and actual departure time are defined as t^^CA and t^^DA respectively. The difference between t^^CA and t^CA (t^^CA – t^CA) and (t^^CD – t^CD), either of arrival or departure, is defined as delay.

According to the paper, state-of-the-art delay prediction systems neither use historical delay data or useful exogenous data about phenomena that can affect railway operations. They instead rely on static rules built by experts of the railways infrastructure based on classical univariate studies.

Problem is mapped into a multivariate regression problem. Compares the performance of kernel methods, ensemble methods, and feed-forward neural networks.

It is possible to build a reliable and robust data-driven model based only on the historical data about train movements. Found remarkable improvements on current state-of-the-art train delay prediction systems. Inclusion of weather data show a significant positive impact on performance.

Focuses on predicting train delays to improve traffic management and dispatching using advanced analytics techniques able to integrate heterogenous / exogenous data.

Delays have various causes: disruptions in the operations flow, accidents, malfunctioning or damaged equipment, construction work, repair work, and weather conditions like snow and ice, floods, and landslides.

Train attempt to follow a fixed schedule called the ‘nominal timetable’.

Rail Traffic Management Systems (TMS) support managing the inherent complexity of rail services. Accurate train delay predictions to TMSs greatly improve traffic management and dispatching in terms of:

* Passenger information systems
* Freight tracking systems
* Timetable planning

They cite as other factors affecting railway operations driver behaviour, passenger volumes, strikes, and holidays. For models based on classical univariate statistics, other factors can only be indirectly considered (e.g. specific models for weekends), not considered, or cannot be easily integrated into models.

Possible to perform a multivariate analysis over data coming from different sources but related o the same phenomena.

Considers the problem as a time series forecast problem, where every train movement represents an event in time. Solution is extended by including data about weather conditions related to the itineraries of the considered trains.

For a train, a dataset of delay profiles can be identified. A set of data-driven models can be built from these profiles which, working together, perform a regression analysis on the past delay profiles and consequently predict future ones.

Oneto et al. (2016) worked closely with Rete Ferroviari Italiana (RFI), the Italian railway authority, and so devised a set of novel Key Performance Indicators (KPIs).

Use Extreme Learning Machines (ELM), Kernel Regularized Least Squares (KRLS), and Random Forests (RF), with the Nonparametric Bootstrap (NB) procedure used to tune hyperparameters.

Considers a rail network a graph where nodes represent a series of checkpoints connected one to each other.

A train follows an itinerary composed of N\_c checkpoints, characterised by a station of origin, a station of destination, some stops, and some transits at checkpoints in between.

Dwell time: the difference between the departure time and the arrival time for a fixed checkpoint.

Running time: the amount of time needed to depart from the first of two subsequent checkpoints and to arrive at the second one.

Time series forecast, where a set of predictive models perform a regression analysis over the delay profiles for each train.

Weather can influence passenger flow and consequently dwell time.

Retrains models when a complete set of messages describing the entire planned itinerary of a particular train for one day is retrieved. Training phases are done during the night when just a few trains are flowing through the network.

Variable of interest is the delay profile of the train. Possible correlated variables include information about other trains travelling on the network, weather conditions, etc. For some correlated variables, the forecasted values cold be available in addition to historical values, i.e. forecasts at future times made at past times.

Input space comprises: weather conditions, delays, dwell times, and running times for the train in question. The weather conditions, delays, dwell times, and running times for all the other trains running over the railnetwork in this time period.

Empirical Risk Minimisation (ERM) is normally avoided, as it leads to severely overfitting the model on the training dataset.

The ELM approach was introduced to overcome problems posed by back-propagation training algorithms: potentially slow convergence rates, critical tuning of optimisation parameters, and the presence of local minima that call for multi-start and re-training strategies.

It is well-known that combining the output of several classifiers results in a much better performance than using any one of them alone.

Used more than 6 months of data related to two main areas in Italy, including more than 1000 trains and several checkpoints.

Each record contains:

* Date
* TrainID
* Checkpoint ID
* Checkpoint Name
* Arrival Time
* Arrival Delay
* Departure Time
* Departure Delay
* Event Type (Origin, Destination, Stop, Transit)

Weather conditions include:

* Temperature
* Relative humidity
* Wind direction
* Wind intensity
* Rain level (mm)
* Pressure (bar)
* Solar radiation (W / m^2)

No WI: no weather information

WI: both historical and weather data have been used

Found that RF with WI was 2x better than the extant system. Weather information improved accuracy by approximately 10%.

**Rail traffic management systems (TMS) (Davey, 2012)**

**Condition-based maintenance in railway transportation systems based on big data streaming analysis (Fumeo et al, 2015)**

Condition-based maintenance of railway assets

**Improving rail network velocity: a machine learning approach to predictive maintenance (Li et al, 2014)**

Condition-based maintenance of railway assets

**Automatic fastener classification and defect detection in vision-based railway inspection systems (Feng et al, 2014)**

Automatic visualisation systems

**Graph methods for estimation of railway capacity (Branishtov et al, 2014)**

Network capacity estimation

**Energy-efficient locomotive operation for Chinese mainline railways by fuzzy predictive control (Bai et al, 2014)**

Optimisation for energy-efficient railway operations

**A delay propagation algorithm for large-scale railway traffic networks (Goverde, 2010)**

**Models for predictive railway traffic management (Kecman, 2014)**

Proposed a real-time delay prediction model based on historical arrival and departure data

**Online data-driven adaptive prediction of train event times (Kecman et al, 2015)**

Presented a microscopic model to predict train travel time and delay for railroad networks. Historical track occupation data are used to predict train travel time and delay for railroad networks.

**Train Delay Prediction Systems (TDPS): A Big Data Analytics Perspective (Oneto et al, 2017)**

# I really, really need to look at schedules.

# We may as well have as much location data as possible, too: there’s no harm in it, and we can always cut it out later

**Improving arrival time prediction of Thailand’s passenger trains using historical travel times (Pongnumkul et al, 2014)**

Worked on data-driven models for train delay predictions, treating the problem as a time-series forecast one. System was based on ARIMA and k-NN models, although their work reports the application of their models over a limited set of data from a few trains.

1. <https://www.networkrail.co.uk/who-we-are/transparency-and-ethics/transparency/our-information-and-data/> [↑](#footnote-ref-1)
2. <https://www.nao.org.uk/report/reducing-passenger-rail-delays-by-better-management-of-incidents/> [↑](#footnote-ref-2)
3. <https://www.networkrail.co.uk/running-the-railway/looking-after-the-railway/delays-explained/> [↑](#footnote-ref-3)
4. <https://en.wikipedia.org/wiki/Headway> [↑](#footnote-ref-4)
5. <https://en.wikipedia.org/wiki/Machine_learning> [↑](#footnote-ref-5)