**Predicting Medium-Term Train Delays**

**Dominic White**

**Noura Al-Mouyabed**

**Abstract**

**Context / background**

Trains are one of the most popular forms of public transport. The perceived likelihood and severity of train delays is a turn-off for many consumers, and so a system predicting medium-term train delays (those occurring up to a certain limit) would improve both customer satisfaction and operating profits. There are currently no solutions to this problem in the literature, nor in practice, although research has been carried out into real-time delay prediction.

**Aims**

The aims of this project are twofold. The primary aim is to identify a suitable model for predicting medium-term train delays. The secondary, supporting aim is to develop a new train delay dataset to train this model on, as there are none widely available.

**Method**

Using historical schedule data, movement data, location data, and weather data from a variety of sources, a dataset is constructed for use in machine learning. A variety of classification and regression models are trained, tested, and optimised. The effect of including weather data, and various encoding schemes, is investigated.

**Results**

An entirely new dataset has been constructed. It is sufficiently robust for general usage, although there remain discrepancies. An improvement on similar works is produced by using random forests for both classification and regression, with x % for classification and y % for regression.

**Conclusions**

To conclude

I managed to achieve this

**Keywords**

Machine learning, train delay, delay prediction, extract-transform-load, etl, classification, regression

**Introduction**

1.7 billion rail journeys were made in the 2018 - 2019 financial year[[1]](#footnote-1).

Trains delays impose a huge cost on passengers and operators by contributing to the inefficiency of train operations (Van Oort, 2011). In 2006 – 07, almost 800,00 incidents caused 14 million minutes of delay to rail journeys, costing a minimum of £1 billion in terms of time lost to passengers (NAO, 2008). Of those, 42% were caused by infrastructure faults, 38% by Train Operating Companies (TOCs), and the remaining 20% by events such as adverse weather, fatalities and vandalism.

Harris and Godward (1997, pg. 130) identify three reasons why punctuality is perceived to be worse than it actually is: passengers tend to recall delayed trains over punctual trains; more passengers travel on late trains than punctual trains (as passengers accumulate while waiting); and operators avoid early running, so late trains are not balanced out by early trains. Furthermore, passengers expect trains to run on time as TOCs are considered to be in greater control of their environment (Murray, 1989, pg. 27-28).

This mismatch in perception places greater importance on reducing the impact of delays on passengers. The benefits are numerous: trains are safer and more environmentally friendly than other forms of transport, such as cars or planes, and are faster than buses.

Furthermore, under a scheme known as ‘Delay Repay’, TOCs are automatically obliged to refund passengers’ tickets if a train is more than 30 minutes late. So there are many incentives to

If passengers could know in advance whether a train would be delayed, how would they modify their behaviour? Could TOCs allocation their resources differently to proactively prevent delays from occurring?

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). The majority of current prediction systems utilise analytical models (Oneto et al, 2016), the focus of this project will be data-driven models, and, in particular, machine learning techniques.

1. **Background**

The motivation to apply machine learning to this problem stems

1. **Objectives**

The research question proposed is: *can medium-term train delays be predicted?* To address this questions, the objectives for this project were divided into three categories: minimum, intermediate, and advanced.

The minimum objective of this project is to construct a high-quality dataset. This dataset will be referred to as DTDD. The purpose of this objective is twofold. It is, first and foremost, a prerequisite for machine learning. Secondly, however, it will establish a strong understanding of the data itself, informing feature engineering in the intermediate and advanced objectives.

The intermediate objective is to develop a classification model to predict *whether* a train will be delayed. A variety of classification models are to be selected, with selection informed by a literature review, then tested on DTDD. The best-performing model is to be further tuned.

The advanced objective is to develop a regression model to predict *by how much* a train will be delayed. This problem is markedly harder than classification.

**Related work**

There has been significant academic effort into producing *real-time* TDP systems. However, for reasons discussed in Solution.Rationale, the medium-term timeframe was chosen as the focus for this project. There are few papers with direct relevance to this timeframe, but much overlap between the two. Solutions to the problem of real-time TDP use a variety of machine learning models, with the field tending towards ensemble models and random forests. Exogenous data used includes weather, infrastructure, and engineering works, in descending order of popularity.

There are many classifications of TDP models, based on scope, model type, and solution methods (Markovic et al, 2015). This project is only concerned with data-driven models.

**Initial attempts**

Peters et al. (2005)

Yaghini et al. (2013)

Milinkovic et al. (2013)

Pongnumkul et al. (2014)

Markovic et al. (2015)

Wang and Work (2015)

**Modern attempts**

Oneto et al. (2016)

Oneto et al. (2017)

Oneto et al. (2018)

Lessan et al. (2019)

Wang and Zhang (2019)

This should be between 2 and 4 pages long. Let’s aim for longer, of course.

**Solution**

**Rationale**

Although 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), of which there are many potential causes (Berger et al, 2011; Milinkovic et al, 2013):

* Signals and points failures
* Severe heat (leading to buckled rails)
* Fatalities
* Flooding
* Landslips
* Leaves
* Snow and ice
* Vandalism and trespass
* Accidents
* Malfunctioning equipment
* Engineering work (repair or construction)
* Prolonged alighting and boarding times

A secondary delay is generated by operational conflict (Cerreto et al, 2016). Primary delays induce a cascade of secondary delays of other trains, which must wait for tracks to be clear, crews to be in the right place, platforms available, and so on. Predicting secondary delays is very difficult, and so this project focuses on predicting primary delays in the medium-term.

Three timescales were considered: real-time / short-term, medium-term, and long-term. Real-time delay prediction models are *online*, i.e. updated as data on train movements become available. Most research focuses on this timescale; models are very complex. Medium-term (referred to as long-term in Wang et al. (2019)) is a period bounded by the availability of exogenous data, such as weather forecasts. Long-term is closer to train scheduling: models in this timescale are applied to timetabling and resource planning, and testing for robustness using historical delay data (Markovic et al. 2015). The medium-term timescale was selected after considerable investigation into the real-time timescale, as it is simpler, poorly explored, and therefore amenable to this project adding value to the area.

This decision meant that secondary delays could be safely ignored. The focus thus shifted to primary delays, and to *exogenous* data. Data is *exogenous* if it is independent of other input data but the output is dependent upon it. 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 recommends using information about passenger flows and about railway asset conditions. Harris (1992), investigating train punctuality in the UK and Netherlands, used proxies for variables:

* 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)

**Extract-Transform-Load**

**Pre-processing**

**The feeds**

Network Rail (NR) provides the following operational data feeds[[2]](#footnote-2):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Acronym** | **Description** | **Frequency** | **Used** |
|  | **BPLAN** | Train planning data, including locations and sectional running times. | Twice a year | N |
| Codes for Operations, Retail, & Planning - a Unified Solution | **CORPUS** | Location reference data. | Monthly | Y |
|  | **MOVEMENT** | Train positioning and movement event data. | Real-time | Y |
| Real-Time Public Performance Measure | **RTPPM** | Performance of trains against the timetable, measured as the percentage of trains arriving at their destination on-time. | One message per minute | N |
|  | **SCHEDULE** | Daily extracts and updates of train schedules from the Integrated Train Planning System (ITPS). | Overnight each night | Y |
|  | **SMART** | Train describer berth offset data used for train reporting. | Monthly | N |
| Train Describer | **TD** | Train positioning data at a signalling berth level. | Real-time | N |
| Temporary Speed Restrictions | **TSR** | Details of temporary reductions in permissible speed across the rail network. | Once a week on a Friday morning | N |
| Very Short Term Plan | **VSTP** | Train schedules created via the VSTP process | Real-time | N |

Three NR feeds are used in this dissertation: CORPUS, MOVEMENT, and SCHEDULE.

Darwin provides schedule information when a train is activated, or when a new schedule, or changes to an existing schedule are received. Schedules cover the complete journey of a single train are always sent in full.

**Choo-choo: TD, TRUST, and Darwin**

The TD feed provides information about the position of trains through a network of *berths*. A berth usually represents a signal. TD was discarded as too low-level for this project.

Train Running Under System TOPS (TRUST) is a NR system used for monitoring the progress of trains and tracking delays in the UK.

Darwin uses both TRUST and TD for real-time data, and also incorporates Darwin workstations, Customer Information Systems (CIS), and internal messaging systems.

There are three train movements. TD feeds into TRUST, and TRUST into DARWIN.

The MOVEMENT feed contains data from TRUST.

# Perhaps this should just be a cold overview. I’m figure it out later. I can always re-integrate the content. I’d rather have a background phase here.

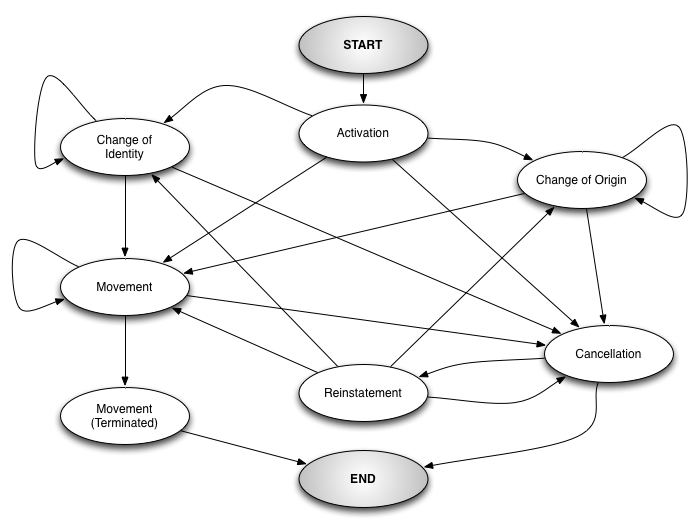


Figure 1. A finite state machine showing the order in which TRUST messages may be received.

|  |  |  |
| --- | --- | --- |
| **Type** | **Name** | **Description** |
| 001 | Train activation |  |
| 002 | Train cancellation |  |
| 003 | Train movement |  |
| 004 | N/A |  |
| 005 | Train reinstatement |  |
| 006 | Change of origin |  |
| 007 | Change of identity |  |
| 008 | Change of location |  |

**Geolocating trains**: **CORPUS (and NaPTAN)**

The Rail Delivery Group (RDG) is in the process of developing a Locations Proof of Concept[[3]](#footnote-3) which unifies multiple RDG and NR location services. However, at time of writing, this is not operational.

CORPUS is location reference data. It is used in conjunction with the National Public Transport Access Node (NaPTAN) database, a nationwide system for uniquely identifying all points of access (nodes) to public transport in the UK. Only rail stations are, naturally, of concern here. A frankly alarming number of codes are used to refer to locations in the UK rail network; they are described below, along with which datasets support them.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CORPUS** | **NaPTAN** | **Description** |
| **STANOX (Station Number)** | Y |  | First two digits are the geographic area. Can refer to non-station locations such as sidings and junctions. Numbers run broadly north-to-south. |
| **UIC** | Y |  |  |
| **3ALPHA / CRS / NRS / TLC** | Y | Y | 3-character limit. Used primarily to identify stations and on seat reservation labels. |
| **TIPLOC (Timing Point Location)** | Y | Y | Relates to points used in deriving train schedules. 7-character limit. A station often has multiple TIPLOCs if it consists of multiple groups of platforms on different lines. |
| **NLC (National Location Code)** | Y |  | 6-character limit. Identifies locations on the railway. Used for retailing and accounting purposes. |
| **NLCDESC** | Y |  | A description of the NLC. No limit. |
| **NLCDESC16** | Y |  | A description of the NLC. 16-character limit. |
| **ATCO** |  | Y |  |
| **EASTING, NORTHING** |  | Y | Geographic Cartesian coordinates for a point. Uses EPSG:27700[[4]](#footnote-4), the British National Grid / Ordnance Survey system. |
| **NAME** |  | Y | The name of the location. |

Each MOVEMENT message, depending on the system, has either a STANOX or TIPLOC code included. Both CORPUS and NaPTAN are therefore necessary to ensure a message can be geolocated.

**The weather**

**Historic data**

The primary source of weather data is the Met Office Integrated Data Archive System (MIDAS). MIDAS is a database of land and marine surface observations, collected from 1853 to the present day, by the Met Office station network. MIDAS offers several datasets. The most comprehensive is the Hourly Weather Observation Data, which contains meteorological values measured on an hourly time scale. These observations include 104 fields[[5]](#footnote-5), though many are for quality control, or too specific to necessitate inclusion.

Station data is available from the Centre for Environmental Data Analysis (CEDA). There are 507 stations, but several are located overseas. Each station is geolocated by latitude and longitude.

**Forecast data**

One of the objectives of this dissertation was the application of a trained model to unseen data. Unseen train data are simply train schedules. Unseen weather data are *forecasts*. Forecasts are available from Met Office Datapoint, a service allow accessing to freely available Met Office data feeds. They may be obtained as 3-hourly site-specific forecasts up to 5 days’ in advance. There are over 5,000 UK forecast sites. There are 10 forecast fields

1. https://dataportal.orr.gov.uk/statistics/usage/passenger-rail-usage/ [↑](#footnote-ref-1)
2. https://www.networkrail.co.uk/who-we-are/transparency-and-ethics/transparency/open-data-feeds/ [↑](#footnote-ref-2)
3. https://wiki.openraildata.com/index.php?title=Locations\_PoC [↑](#footnote-ref-3)
4. https://epsg.io/27700 [↑](#footnote-ref-4)
5. https://artefacts.ceda.ac.uk/badc\_datadocs/ukmo-midas/WH\_Table.html [↑](#footnote-ref-5)