Train Delay Prediction System for Large-Scale Railway Networks based on Big Data Analytics

Emanuele Fumeo, D. Anguita, L. Oneto, G. Clerico, N. Mazzino, F. Papa, R. Canepa

Napoli, 24th November 2016

University of Genoa, Ansaldo STS S.p.A. and Rete Ferroviaria Italiana (RFI)
Presentation Index

- Big Data Analytics for Railways
- The Problem of Train Delay Prediction
- Data-driven Methodology
- Description of the Work Done
- Results
- Conclusions
Presentation Index

• Big Data Analytics for Railways
• The Problem of Train Delay Prediction
• Data-driven Methodology
• Description of the Work Done
• Results
• Conclusions
Big Data Analytics

**Big Data** refers to technologies (both HW and SW) and methodologies able to collect, store, process, analyze and visualize large amounts of heterogeneous data.

Big Data Analytics aims at learning from data, in order to build generalizable data-driven models that give accurate predictions, or aims at finding patterns, particularly with new and unseen similar data.
Big Data Analytics for Railways

Big Data Analytics interest in Railways is constantly growing

Companies are exploiting Big Data Analytics in any sector of railway operations

Maintenance

Traffic Management

Customer Relationship

…and many other aspects

© Patrick Cashin / MTA, © Luca Mazzucco, © Maksym Kozlenko, © Kevin Hadley
Presentation Index

• Big Data Analytics for Railways
• The Problem of Train Delay Prediction
• Data-driven Methodology
• Description of the Work Done
• Results
• Conclusions
Train Delay Prediction Problem

If the train is at **checkpoint A** with 5 minutes of delay, what will be its delay at **checkpoint B**? And at **checkpoint C**?
Current Predictive Methodology

Current method for train delay prediction is based on line characteristics, on trains characteristics and on simple statistics, aiming at computing the amount of time needed to complete a particular section of its trip and exploiting it for predictions.

This method is not able to take into account external factors affecting railway operations!
Data-driven Train Delay Prediction

Study train delays depending on train movements and weather conditions data

Predict the delay of a train at each of the successive checkpoints included in its trip
Presentation Index

• Big Data Analytics for Railways
• The Problem of Train Delay Prediction
• Data-driven Methodology
• Description of the Work Done
• Results
• Conclusions
Data-driven Methodology – General Idea

Use data about train movements and weather conditions to see what happened in the past and try to predict what will happen in the future.
Data-driven Methodology – Implementation

Build a set of generalizable black-box data-driven models that are able to respond to new, previously unseen input data to predict train delays

**Input Data (Current)**
- Current Train Data
- Current Weather Data

**Models (black-box)**

**Output Data (Predictions)**
- Delay in B = 1 min
- Delay in C = 0 min
Presentation Index

• Big Data Analytics for Railways
• The Problem of Train Delay Prediction
• Data-driven Methodology

• Description of the Work Done
• Results
• Conclusions
Outline of the next slides

• Description of Available Data
• Problem Formalization
  • Graphical Example
  • Modelling Approach
• Tested Algorithms
• Models’ Performance Assessment
• Simulation Results
• Conclusions
Description of Available Data

**Train Movements (from PIC by RFI)**

Records about any arrival or departure associated with a particular train at a particular checkpoint.

The data refers to **6 months** of movements in the area of **Milan**, and **1 year** in the area of **Genoa**. It includes:

- TrainID
- Checkpoint ID and name
- Scheduled departure/arrival datetime
- Actual departure/arrival delay
- Event type (origin, destination, passage, stop)

From them, we also derived **Dwell times** at checkpoints and **Running times** between subsequent ones.
Description of Available Data

Weather Conditions (from Italian Weather Services)

This data refers to the same time frame of the train movements data. Weather data contains several parameters from national weather stations:

- **Temperature** (min, max, average)
- **Humidity** (min, max, average)
- **Wind** (min, max, average)
- **Rain Level** (min, max, average)
- **Pressure** (min, max, average)
- **Solar Radiation** (min, max, average)
Problem Formalization

How we approached the train delay prediction problem from a data-driven point of view
Example – Generic Train trip

A train trip starts from station “M” and ends at station “Q”
Example – Selection of weather stations

For each checkpoint in the trip, consider data from the closest weather station.

We consider data from weather station 3 for checkpoints M, L and I.

We consider data from weather station 1 for checkpoint D.

We consider data from weather station 2 for checkpoints N, O and Q.
For each train and for each of the successive checkpoints composing its trip, a **data-driven multivariate regression model** is built, which will output delay predictions for arrival and departures for the corresponding checkpoint.

Each arrow represents a data-driven model.

In this example, 6 different models for this train at checkpoint “M” have to be built.
The same approach is applied each time the train moves to a subsequent station, so to exploit the new information as soon as it is available.

For this train at checkpoint “L”, 5 data-driven models have to be built. For the entire train trip, $6 + 5 + \ldots + 2 + 1 = 21$ different models have to be built.
Train Delay Prediction as a Big Data problem

Every day, in the entire Italy:

- \(\approx 10,000\) trains from the service perspective (i.e. unique train IDs)
- Average of \(\approx 12\) checkpoints for every train
- \(\approx 120,000\) movements
- \(\approx 10\) GB of messages

Therefore, exploiting the proposed approach results in more than 600,000 delay prediction models
Tested Big Data Regression Algorithms

- Extreme Learning Machines (ELM)
- Random Forest (ensemble of Decision Trees)
- Kernel Regularized Least Squares (KRLS)

Implementation for the Big Data framework on Apache Spark (Apache Hadoop)
Performance Assessment Methodology

**Entire Dataset**

- **Training Set**: Data used to build a data-driven model
- **Test Set**: Data used to assess the performance of a model on data that has not been used before

**Test set** is exploited for simulations, so to compare predictions with what really happened in the past.
Simulations for Performance Assessment

Data-driven Models

Update models every new day!

Day 365

Big Data Algorithms

Delay Predictions

Compute (Actual Delays - Predicted Ones) & Save Results

M L I D N O Q

d d d d d d d
d d d d d d d
d d d d d d d
d d d d d d d
d d d d d d d

...
Presentation Index

• Big Data Analytics for Railways
• The Problem of Train Delay Prediction
• Data-driven Methodology
• Description of the Work Done
• Results
• Conclusions
Summary of Simulation Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Current technique</th>
<th>Best data-driven model (without weather)</th>
<th>Best data-driven model (with weather)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVG &amp; VAR (Actual Delay – Predicted Delay)</td>
<td>mean = 3.3 min variance = ±1.6 min</td>
<td>mean = 2.0 min variance = ±0.3 min</td>
<td>mean = 1.9 min variance = ±0.3 min</td>
</tr>
</tbody>
</table>

- **Improvement factor** ≈ ×1.7 (on total average)
- The **accuracy of the data-driven methods is quite homogeneous** (see variance) with respect to different trains and stations, while the current technique suffers from high fluctuations
- The inclusion of **weather data further improves the model**, although slightly
Presentation Index

• Big Data Analytics for Railways
• The Problem of Train Delay Prediction
• Data-driven Methodology
• Description of the Work Done
• Results

• Conclusions
Conclusions

• The new delay prediction system **outperforms the current one**
• The new technique is **designed to cope with large scale railway networks** generating large amounts of data
• It could be **integrated in a TMS solution** to provide advanced and accurate forecasting of train delays

Next Research Steps:

• Find **new railway data to be integrated in data-driven models** (e.g. state of infrastructure assets, train compositions, etc.)
• Consider the integration of **other external data** (e.g. passenger flows for passenger services) in order to analyze whether a further performance improvement is possible
Acknowledgements

In2Rail

Capacity for Rail

Horizon 2020

The Framework Programme for Research and Innovation
THANK YOU FOR YOUR ATTENTION

Train Delay Prediction System for Large-Scale Railway Networks based on Big Data Analytics

Emanuele Fumeo, D. Anguita, L. Oneto, G. Clerico, N. Mazzino, F. Papa, R. Canepa

Napoli, 24th November 2016

Ansaldo STS S.p.A., University of Genoa and Rete Ferroviaria Italiana (RFI)