Modeling delay relations based on mining historical train monitoring data: a Chinese railway case

Jingwen Guo 1,2 Lingyun Meng 1 Pavle Kecman 3 Francesco Corman 4

1State Key Laboratory of Rail Traffic Control and Safety Beijing Jiaotong University
2Beijing Research Center of Urban Traffic Information Sensing and Service
3Department of Science and Technology, Linköping University
4Section Transport Engineering & Logistics, Maritime and Transport Technology, Delft University of Technology

26 March 2015
Outline

1. Introduction
2. Investigating train delay dependencies
3. Computational experiments
4. Current research
5. Conclusions
Real-time information in railway traffic

Quality of railway service depends on accurate predictions of future train movements on many levels:

- Traffic can be controlled pro-actively by taking actions that prevent delays and delay propagation
- Timetable, passenger transfer plans, rolling-stock and crew circulation plans can be kept up-to-date
- Passengers can be provided with accurate information: in-vehicle, on-platform, online
Real-time information in railway traffic

Important properties of real-time information and short-term predictions:

- Accurate and reliable
- Dynamic and responsive
- Stable over longer prediction horizons

Source: xkcd.com/612
A train run is typically monitored in discrete points in the network:
- track circuit
- signal
- station (timetable points)

Train passing times given with one-second or -minute precision.

Train position updates can be used to derive predictions of future train movements by means of:
- microscopic simulation
- data-driven prediction models
Data-driven prediction models

- Model-based prediction
- Data-driven prediction
- Real-time prediction

training set $\xrightarrow{induction}$ general rule

live data stream $\xrightarrow{transduction}$ prediction

live data stream $\xrightarrow{induction}$ general rule

prediction

deduction
Data-driven prediction models

- Model-based prediction
- Data-driven prediction
- Real-time prediction

live data stream \[\xrightarrow{\text{transduction}}\] prediction

training set \[\xrightarrow{\text{induction}}\] general rule \[\xrightarrow{\text{deduction}}\] prediction
Data-driven prediction models

- Model-based prediction
- Data-driven prediction
- Real-time prediction

```
live data stream ↘︎ transduction
↓
training set ▶︎ induction
↑
general rule
↓
deduction
↓
prediction
```
Outline

1. Introduction

2. Investigating train delay dependencies

3. Computational experiments

4. Current research

5. Conclusions
In our approach...

- A train run is monitored only at timetable points with one minute precision.
- This prevents creating detailed prediction models.
- The dynamics of a train delay over time and space is presented as a sequence of events.
- **Given a current delay of a train, the objective is to estimate the delay of all remaining events along the train route.**
Methodological framework

Model

- A train run given as a sequence of discrete events that model arrival and departure events
- Each event is connected to all remaining events in the sequence
- Interaction between trains is not included in the model
- Historical traffic data used to learn the dependence between delays along the train route
Methodological framework

Model

- A train run given as a sequence of discrete events that model arrival and departure events
- Each event is connected to all remaining events in the sequence
- Interaction between trains is not included in the model
- Historical traffic data used to learn the dependence between delays along the train route
Methodological framework

Model

- A train run given as a sequence of discrete events that model arrival and departure events
- Each event is connected to all remaining events in the sequence
- Interaction between trains is not included in the model
- Historical traffic data used to learn the dependence between delays along the train route

Station 1          Station 2          Station 3          Station N

arr    dep    arr    dep    arr
A train run is given as a sequence of events \( i = 1, \ldots, N \) where each event \( i \) is defined by a tuple (train number, station, event type).

Train runs of the same train number over multiple days are characterised with the same route and stopping pattern.

Historical traffic data are used to calibrate the following linear regression models:

\[
d_j = a_j + b_j d_i, \quad \forall i = 1, \ldots, N, j = i + 1, \ldots, N
\]

where \( a, b \) are linear coefficients.
After an update about train position is received, delay of all remaining events in the sequence is predicted using the derived linear models:

\[
\begin{align*}
  d_{i+1} &= a_{i+1} + b_{i+1} d_i \\
  d_{i+2} &= a_{i+2} + b_{i+2} d_i \\
  d_{i+3} &= a_{i+3} + b_{i+3} d_i \\
  d_{i+4} &= a_{i+4} + b_{i+4} d_i \\
  d_{i+5} &= a_{i+5} + b_{i+5} d_i
\end{align*}
\]
Outline

1. Introduction
2. Investigating train delay dependencies
3. Computational experiments
4. Current research
5. Conclusions
Case study
General description

- Data from the high-speed line between Beijing and Shanghai
- Data is from the northern part - 5 stations
- 58 G (300 km/h) trains and 12 D (250 km/h) trains daily per direction
- Data between the 1st of December, 2013 and the 4th of March 2014
- Only planned and realised time for each departure, arrival and through event (no signal or track data) rounded to full minutes
- Test set contains 20% of randomly selected train runs
- Trains are allowed to depart up to 5 minutes before their scheduled departure time
### Results

#### Results of regression analysis

<table>
<thead>
<tr>
<th></th>
<th>A/D East D</th>
<th>Cangzhou West A</th>
<th>Cangzhou West D</th>
<th>Tianjin South A</th>
<th>Tianjin South D</th>
<th>Langfang A</th>
<th>Langfang D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dezhou East D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cangzhou West A</td>
<td>0.9876</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cangzhou West D</td>
<td>0.9781</td>
<td>0.9879</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tianjin South A</td>
<td>0.9492</td>
<td>0.9569</td>
<td>0.9647</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tianjin South D</td>
<td>0.9423</td>
<td>0.9493</td>
<td>0.9547</td>
<td>0.9899</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Langfang A</td>
<td>0.8787</td>
<td>0.8889</td>
<td>0.8965</td>
<td>0.9578</td>
<td>0.9631</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Langfang D</td>
<td>0.8767</td>
<td>0.8866</td>
<td>0.8938</td>
<td>0.9551</td>
<td>0.9600</td>
<td>0.9938</td>
<td>-</td>
</tr>
<tr>
<td>Beijing South A</td>
<td>0.8575</td>
<td>0.8730</td>
<td>0.8795</td>
<td>0.9371</td>
<td>0.9419</td>
<td>0.9777</td>
<td>0.9818</td>
</tr>
</tbody>
</table>
## Results

### Results of regression analysis

<table>
<thead>
<tr>
<th>A/D</th>
<th>Beijing South D</th>
<th>Langfang A</th>
<th>Langfang D</th>
<th>Tianjin South A</th>
<th>Tianjin South D</th>
<th>Cangzhou West A</th>
<th>Cangzhou West D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing South D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Langfang A</td>
<td>0.9829</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Langfang D</td>
<td>0.9753</td>
<td>0.9910</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tianjin South A</td>
<td>0.9142</td>
<td>0.9265</td>
<td>0.9349</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tianjin South D</td>
<td>0.8983</td>
<td>0.9065</td>
<td>0.9125</td>
<td>0.9735</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cangzhou West A</td>
<td>0.8302</td>
<td>0.8404</td>
<td>0.8451</td>
<td>0.9287</td>
<td>0.9522</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cangzhou West D</td>
<td>0.8283</td>
<td>0.8314</td>
<td>0.8361</td>
<td>0.9215</td>
<td>0.9432</td>
<td>0.9900</td>
<td>-</td>
</tr>
<tr>
<td>Dezhou East A</td>
<td>0.8115</td>
<td>0.8149</td>
<td>0.8194</td>
<td>0.9060</td>
<td>0.9271</td>
<td>0.9752</td>
<td>0.9844</td>
</tr>
</tbody>
</table>
Results
Results of regression analysis

The distribution of R-square

Arrival/departure delays from Cangzhou West to Beijing South station
1 Introduction

2 Investigating train delay dependencies

3 Computational experiments

4 Current research

5 Conclusions
Online reinforcement learning

Observe how much the predictions of the past events were wrong and adapt the future predictions accordingly

- Adapting running and dwell time estimates based on the realised running and dwell times of the same train
- Learning from temporal differences: use the prediction error of the previous train to adapt the prediction for the current train
- Adaptive component may depend on headway between trains - train interactions included without an explicit model
Outline

1 Introduction

2 Investigating train delay dependencies

3 Computational experiments

4 Current research

5 Conclusions
Summary and conclusions

- Analysis of dependence of en-route train delays
- The goal was to determine the how far the real-time information can be propagated with high accuracy
- Future work on modelling the dynamic interrelation of delays and adapting the offline computed functions in real-time
- It is highly recommended to evaluate the approach on other case studies
- Potential applications include integration with online traffic control models, online delay management and passenger information systems
Thank you for your attention