Analysis of Train Delays using Bayesian Networks

Valentin Barzal[1*], Matthias Rößler[2], Matthias Wastian[2],
Felix Breitenecker[1], Nikolas Popper[3]

[1]Institute of Analysis and Scientific Computing, TU Wien, Wiedner Hauptstraße 8-10,
1040 Vienna, Austria; [*]*valentin.barzal@tuwien.ac.at*

[2]dwh GmbH, Neustiftgasse 57-59, 1070 Vienna, Austria;

[3]Institute of Information Systems Engineering, TU Wien, Favoritenstraße 11, 1040 Vienna, Austria;

SNE 33(2), 2023, 89-92, DOI: 10.11128/sne.33.sn.10645
Selected ASIM SST 2022 Postconf. Publication: 2023-02-01;
Received Revised: 2023-04-26; Accepted: 2023-05-05
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna
ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. Bayesian networks can be used for analysis and representation of dependencies in large data sets. Due to their property of operating with graphs, they are suitable for analyzing delays in rail networks. After getting an overview of the theory of Bayesian networks, this article deals with recent literature about Bayesian networks and train delays. Furthermore, the presented methods will be applied to data from the Austrian railway network.

Introduction

Since public transport plays a major role in society, it is on societies behalf to keep delays to a minimum. In order to reach the goal of minimizing delays, it is important to understand not only the reasons of delays but also the propagation of delays within the public transport network. The availability of large data sets, powerful computers and suitable mathematical tools allows analyzing such delays in a sensible way.

1 Bayesian Network

A *Bayesian network* is a model represented by a directed acyclic graph (DAG) $\mathscr{G} = (V,E)$, whereat V is the vertex set and E is the edge set. Every vertex represents a random variable. There exists an edge $e \in E$ between two vertices $v,w \in V$, if and only if there is a probabilistic dependency between the two random variables represented by v and w ([1]).

Depending on the model, the random variable (i.e. vertex) can take on different types of values ([2], [3]), e.g. unordered values (e.g. *blue*, *green*, *yellow*), ordered values (e.g. *low*, *medium*, *high*) or continuous values ($\mathbb{R}$). Every random variable also has its probability distribution, which needs to be added to the graphical representation. If the random variables take on continuous values (e.g. when dealing with continuous data), we often consider *multivariate normal variables*, i.e. the random variables are normally distributed and linked by linear constraints. An example of this case is illustrated in Figure 1.

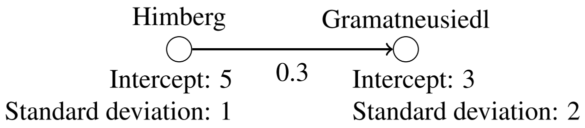

Figure 1: Considering that a train at station *Himberg* is 10 minutes late, then it is expected to be 6 minutes late ($3 + 0.3 \cdot 10 = 6$) at station *Gramatneusiedl*, with a standard deviation of 2 minutes.

When using Bayesian networks for modelling real-world problems, we often just have data describing the values of our random variables, but no information about the graph structure, i.e. the edges/dependencies between the vertices/random variables. In this case, we want to derive the graph structure from the values of the random variables. The *Max-Min Hill-Climbing* method (MMHC) enables this derivation and will be used in Section 2 and Section 3. A detailed introduction to this algorithm can be found in [4].

2 Analyzing the Long Island Rail Road System

In [1], Ulak et al. investigate network-wide pattern of rail transit delays in the Long Island Rail Road System, i.e. a railway system in New York that has a length of 620 miles and consists of a total of 11 rail lines. In Section 3, the methods used in [1] will be modified and used for data of the Austrian railway system. Therefore, Section 2 recapitulates the methods and ideas of [1] without going too deep into detail or presenting the results.

In order to build a Bayesian Network as described in Section 1, every station is considered as a random variable. It takes on the average delay at this station within a certain time interval. There should exist an edge between two vertices (i.e. between two stations) if and only if there is a dependence between those two vertices (i.e. if a delay at one station affects the occurrence of a delay at the other station).

It is possible to classify each train as either eastbound or westbound. When analyzing the delay propagation, i.e. when creating the Bayesian Network, either all eastbound or all westbound trains are considered, but not both east- and westbound trains.

By analyzing the data of an app used by the passengers, the authors receive the delay of every train at every station (with a few exceptions). The observation period is divided into 1-hour intervals. For every station the average delay per train within the examined hour is calculated. A data sample consists of the average delay time at every station within one hour. By using the *Max-Min Hill-Climbing* method (MMHC) and the maximum likelihood method, the graph structure and the corresponding parameters are compiled from the independent data samples. The `R`-package `bnlearn` was used for this task. Detailed documentation about this package can be found in [3].

In order to receive an expressive score for every stations role in delay propagation, two metrics are introduced. The Inducer score of station s indicates how much a delay at station s influences the delays at all the other stations. The Susceptible score of station s indicates how much the delay at station s is influenced by the delays at all the other stations. Similar scores will be introduced in Section 3 in a more precise way.

3 Analyzing the Austrian Railway System

In this section we want to modify the methods of [1] and apply them to data of the Austrian railway network. For this purpose, delay data provided by the Austrian national railway company *Österreichische Bundesbahnen (ÖBB)* is used. It not only contains information about the delay at every train station, but at every operation control point, a so-called "Betriebsstelle". These are specific points within the network. Every train station is also a operation control point. In the following, the word "station" will be used for operation control point.

3.1 Differences to the Long Island Rail Road System

The different topology of the railway network and the different timetables cause different problems that need to be faced:

- In [1], the length of every time interval is chosen to be 1 hour. Considering longer time intervals ”flattens” the delays, since temporal peaks are divided by a higher number of trains. So shorter time intervals provide better ”resolution” of the occurrence of delays.

  On the other hand, it takes some time to propagate the delay within the network. In Austria, there are longer distances between stops and therefore it takes longer to propagate delays. Choosing the length of the time interval to be 1 hour, most trains will ”occur” in more than one time interval. Since the different data sets (i.e. delays within different time intervals) are meant to be independent, this causes some problems.

- There are stations that are not passed by any train in certain time intervals. Calculating the average delay in that period would provoke dividing by zero. This problem can be solved by considering the totaled delay instead of the average delay.

  However, this still might yield meaningful results, since the delays at a station that is approached by a lot of trains, each having just a little delay, might be more relevant than the delay at a station that is only approached by a single train, having a longer delay.

- The Long Island Rail Road System is quite crowded, i.e. if a train is delayed, it is very likely that this affects other trains. In Austria, there are also parts where the temporal distance between different trains is much longer. In this case, delays will still propagate since a delayed train probably will still be delayed at its next stop, but it does not directly affect the delays of other trains.

- By considering only the eastbound or westbound trains, a cycle-free network was obtained in [1]. In Austria, a classification in eastbound or westbound is not possible (e.g. there are also trains going from north to south and vice versa). However, considering all the trains will lead to cycles within the network. This conflicts with the attempt of modelling the network as a directed *acyclic* graph. A strategy to resolve this problem is to delimit the number of trains considered.
- In the Long Island Rail Road System, there are just 124 stops, so it is a relatively small network. Considering all trains in Austria leads to a gigantic model (in terms of vertex numbers, edge numbers and parameters). But when examining only a selection of trains (and its corresponding stations), there are still influences from the non-considered trains. So there is "external" disturbing.
- There are also trains not operated by ÖBB using the same tracks . In the provided data there is only information about the trains operated by ÖBB. So even if no data selection is done, there will be still disturbances by other, not observed trains.

3.2 Methodology

Considering Subsection 3.1, the following restriction will be made:

- Only trains going from Vienna (station *Hauptbahnhof* or *Westbahnhof*) to St. Pölten and vice versa and the corresponding operational points between those two stations will be observed. In this case each train can be classified in eastbound or westbound and we obtain cycle-free networks. This section is also quite crowded, so there will be interactions between different trains.
- Since the distance between Vienna and St. Pölten is relatively short, we can set the length of the time intervals to 1 hour, without dealing too much with the problems mentioned in Subsection 3.1.
- We will not consider the average delay but the totaled delay of every time interval at each station.

Considering only the stations from Vienna to St. Pölten and vice versa is of course a very strong restriction but it enables us receiving suitable results. The considered stations (”Betriebsstellen”) are sketched in Figure 2. In Vienna, there are two branches - one branch going to *Wien Westbahnhof* and one branch going to *Wien Hauptbahnhof*. Between the stations *Knoten Hadersdorf* and *Knoten Wagram* there are also two different routes: In the north (via *Tullnerfeld*) there is the so-called *Neue Westbahn*, in the south (via *Pressbaum*) there is the *Alte Westbahn*.

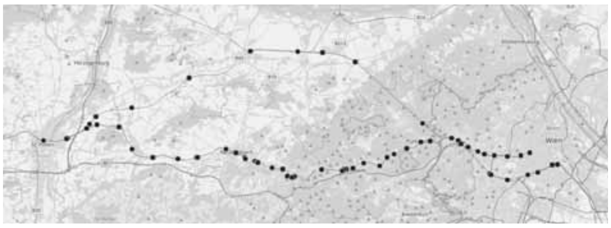

Figure 2: Considered stations (operation control points) between Vienna and St. Pölten

3.3 Network structure

Similar to [1], we use the `R`-package `bnlearn` (and the functions `hc` and `bn.fit`) to obtain the network structure. It was also considered to use the `Python`-packages `Pomegranate` and `PyMC3` for the same task, but those packages can only deal with random variables that take on unordered values . The results can be seen in Figure 3 and Figure 4. In these figures, the directions of the dependencies were not indicated.

When searching for the graph structure, we expect a graph that resembles the geographical topology of the stations, e.g. it would not make sense if a station close to St. Pölten directly affects a station close to Vienna. The topological sorting never coincides with the actual geographical order of the stations. However, we did expect such a behavior, since phenomena like backward propagation (i.e. dependencies going in the opposite direction) and others can cause a mismatch with the geographical order of the stations. Furthermore, some stations are very close and therefore have almost the same delays. In this case, the order of those stations in the obtained network is almost arbitrary.

When observing the plots visually, we detect that there are also connections between stations that are far apart. There are even connections between stations from the *Alte Westbahn* and *Neue Westbahn*, whereat these kind of connections are relatively rare.

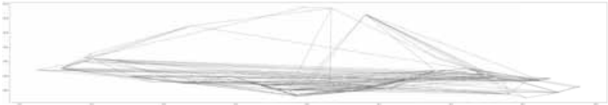

Figure 3: Network connections (eastbound). X-axis and y-axis indicate latitude and longitude.

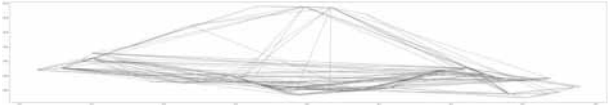

Figure 4: Network connections (westbound). X-axis and y-axis indicate latitude and longitude.

3.4 Metrics

The metrics are inspired by the metrics used in [1], but slightly modified.

The Inducer score indicates how much a delay at a station s influences the delays at all the other stations. Therefor, the delay at the examined station s is set to the median delay of this station (based on the data) and the delays at all the other stations are set to 0. Given this initial condition, the expected delay at every other station is estimated by the Bayesian network. Those values are averaged over all stations. The score of station s can be written as

$$\text{Score}_\text{I}(s) = \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq s}}^{n} \mathbb{E}(X_i | X_s = x_s, \mathbf{X}\backslash(X_s \cup X_i) = 0)$$

where n is the number of random variables, x_s is the median (observed) delay at station s and $\mathbf{X} = (X_i)_{i\in\{1,\dots,n\}}$ is the ensemble of all random variables.

The Susceptible score indicates how much the delay at a station s is influenced by the delays at all the other stations. Therefor, the delay at one station i (with $i \neq s$) is set to its median delay (based on the data) and the delays at all the other stations (including the examined station s) are set to 0. Given this initial condition, the expected delay at the examined station is estimated by the Bayesian network. This procedure is repeated and averaged over all stations $i \neq s$. Using the same notation as above, the score of station s can be written as

$$\text{Score}_\text{S}(s) = \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq s}}^{n} \mathbb{E}(X_s | X_i = x_i, \mathbf{X}\backslash(X_i \cup X_s) = 0)$$

When considering the metrics, we observe that almost all stations with the highest (top 4) scores are either stations that are passed by all trains or stations of the *Alte Westbahn*. Only the station *Tfo* (close to Tullnerfeld) has the highest inducer score (westbound), although it is located at the *Neue Westbahn*. So the *Neue Westbahn* acts well in reducing delay propagation. Going in the westbound direction, two stations at the very beginning of the route reach high susceptible scores (*Wien Hauptbahnhof* and *Wien Westbahnhof Frachtenbahnhof*). Those stations are very susceptible to delays at other stations (although almost all stations are passed *after* those two stations). We also observe that some stations reach negative susceptible scores. According to our model, this means that delays at certain other stations reduce the delay at this observed station.

3.5 Conclusion

When applying the methods introduced in [1] to the Austrian railway system, we need to make restrictions. In order to obtain the inevitable property of an acyclic directed graph, we can only consider an area where every train can be classified into a certain (cycle-free) direction. `R` yields a network structure that seems partly reasonable. There are also connections between stations that are far apart and connections between *Alte Westbahn* and *Neue Westbahn*. However, this does not necessarily need to conflict with the real behaviour.

The inducer score and susceptible score indicate which stations play a big role in delay propagation. It seems reasonable that the *Neue Westbahn* acts well w.r.t. reducing delay propagation, since this section was constructed quite recently. The fact that stations at the beginning of the route have quite high susceptible scores fits to the idea of back-propagation. However, there are also stations with negative susceptible scores. This does not seem to be reasonable.

References

[1] Ulak MB, Yazici A, Zhang Y. Analyzing network-wide patterns of rail transit delays using Bayesian network learning. *Transportation Research Part C: Emerging Technologies*. 2020;119:102749.

[2] Korb KB, Nicholson AE. *Bayesian artificial intelligence*. CRC press. 2010.

[3] Scutari M, Lebre S. Bayesian networks in R: with applications in systems biology. 2013.

[4] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*. 2006;65(1):31–78.