Modeling Delay Propagation and Transmission in Railway Networks

David Rößler[1], Julian Reisch[2] and Natalia Kliewer[1]

[1] Freie Universität Berlin, Department of Information Systems, Berlin, Germany
{david.roessler,natalia.kliewer}@fu-berlin.de
[2] DB Netz AG, Department for Timetabling and Capacity Management, Berlin, Germany
julian.reisch@deutschebahn.com

Abstract: In railway scheduling, the planning of time supplements is crucial to the robustness of the resulting timetable. Time supplements as a means to accommodate for train delays are often distributed according to operation rules and based on experience. A part of the project for strategic schedule optimization at *DB Netze* aims at improving the supplements distribution through learning of structures of delay propagation and transmission from historical railway operation data. The work at hand focuses on delay transmissions between trains. It employs correlations and correlation network analysis to identify and analyze these knock-on delays and to develop logical precedence orders of trains at certain operation points which can in turn be used in a sequential calculation of single train delay propagation. Furthermore, it endeavors to establish a basis to identify strongly connected groups of trains and stations, thus forming relevant subnets for further analysis.

Keywords: data analytics, correlation network analysis, delay management, rail transportation, railway timetabling

1 Introduction

The German railway network is with a total of about 38,000 track kilometers the largest in Europe. The complexity of the timetabling process arises from three conflicting objectives that need to be balanced out. First, the capacity on the infrastructure, that is, allocation of trains to tracks, is to be maximized. Secondly, trains should operate with reasonably tight schedules and thirdly, the timetable must be robust against minor disruptions.

This paper contributes to a research project that aims at optimizing the third aspect, the robustness, meaning that the number of trains and also the traveling times are fixed. One way to cope with minor disruptions and hence improve robustness is to include slack in the timetable, that is time supplements. The trade-off between scheduling slack to achieve robustness against unforeseen events and the goal to realize a schedule as

14th International Conference on Wirtschaftsinformatik,
February 24-27, 2019, Siegen, Germany

efficient as possible and to operate as many trains as possible is evident. To improve upon this trade-off by identifying optimization potentials within the current distribution of supplements is the greater goal.

The project this work is part of is structured in the following way: First, we analyze delay propagation through the network of train operations. Clearly, if a train is delayed at a certain station, then this delay might still have an impact on the amount of delay in its next operation point. Moreover, delay may also be transmitted from one train to another. This happens for manifold reasons, e.g. passenger transits, track blockage and so forth. We implement a correlation network to detect the most important inter-train interdependencies. Then, in the second part of the project, we model delay propagation probabilistically, but only within one train operation and through the formerly detected interdependencies between trains. This gives us the possibility to see what happens, if the probabilities for a delay change, e.g. if trains always depart on time and so on. Furthermore, this model will be the basis of an optimization task, which will be the third and last part of our project. We will apply algorithms that modify the distribution of time supplements and check whether delays might propagate less, thus yielding a marginally increased punctuality.

The work at hand contributes to this project in the following way: We develop and deploy a novel approach to modeling railway train interdependencies with respect to the propagation and absorption of delays. As [1] have presented in their comprehensive survey, railway data analytics can benefit from employing *Big Data* methods. With a focus on scalability for the included development task, our paper explores and selects procedures and tests them on an exemplary large data set from German railway operations. This enables us to detect which trains have a dependency significant enough so that it should be included in the delay propagation model. Furthermore, the delay networks approach yields us the crucial trains and stations where a delay has a huge impact on many other trains in the network. It will prove wise to first optimize the punctuality of these trains, for instance in an initial solution of a future optimization algorithm.

The outline of this paper will be as follows. In Section 2 we introduce the data from railway operations and motivate our choice of data selection. We then give a brief insight in how we clean our data with respect to outliers, missing values and seasonality. What is next, we present our model for delay transmission. More precisely, we analyze the influence of the absolute value of delays of one train to the change of delay of a succeeding train. In Section 3, we start by discussing two measurements of this influence, namely the Pearson and the Kendall correlation coefficient. This gives rise to a delay transmission network that utilizes these measurements as weights on the edges of a network graph model. Section 4 ends with an analysis of a select example and the validation of the overall results. Finally, we give a short conclusion in Section 5.

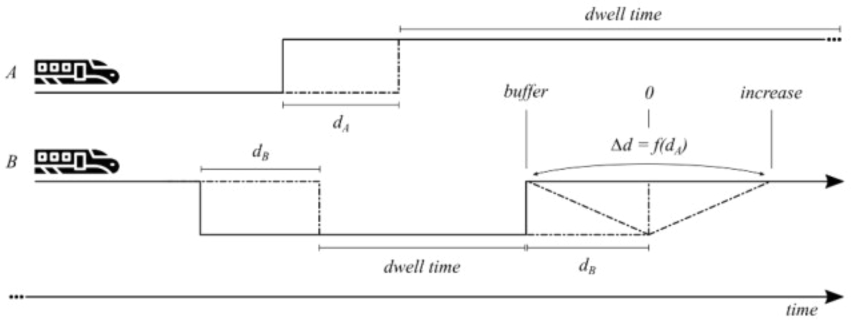


Figure 1. Model for delay transmission between two trains. Train *A* arrives with a delay d_A and train *B* with d_B. The change in delay upon departure of train *B* is ranged between *buffer*, where $\Delta d = -d_B$, no change, where $\Delta d = 0$, and *increase*, in which case Δd.
Source: own compilation.

2 Methods

We focus on the effects of the total delay of an arriving train on the ability of any departing train to reduce prior delay by using up time buffer or the necessity to build-up additional delay. Figure 1 illustrates this idea. We denote the *total delay upon arrival* as d and the *change in delay* as Δd. The suspected relationship can be expressed by $\Delta d_{Dep} = f(d_{Arr}) := \Theta \cdot d_{Arr}$, where Θ is an arbitrary function which yields a measure of association.

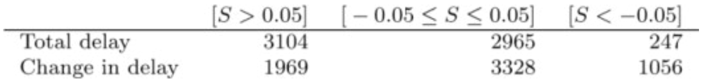
2.1 Data Selection: Region and Traffic Type

As an example, for our analysis, we choose the long-distance train network in the south western region of Germany where we deal with two major railway corridors. The first one is from Basel to Frankfurt and the other one from Stuttgart to Cologne. These corridors meet in Mannheim, where a transfer is possible as the different long-distance train lines are synchronized there. This synchronization is the first reason to choose this region as we expect interdependencies of the long-distance trains there. Furthermore, the two corridors are highly frequented so that the capacities are fully saturated, and an optimization of slack is helpful. We discretize the corridors by flag stops and omit signals which do not involve a regular stop. Accordingly, this study focuses on the southern part of the first corridor with stations *Offenburg (RO)*, *Baden-Baden (RBB)*, *Karlsruhe (RK)*, and *Mannheim (RM)*, the latter two being the cities with the second and third largest population in Baden-Württemberg, whereas the former stations are stereotypical for smaller sized cities. A large variety of train types is moving along the German railway network. We generally distinguish between regular and irregular and long- and short-range passenger trains and freight trains. Since this work's approach is to analyze regular train encounters with a focus on temporal precedence, and since freight traffic behaves structurally irregular, this work will abstract from freight trains and solely consider regularly running passenger trains - both long and short distance.

2.2 Data Cleaning: Outliers, Missing Data, STL

Before the actual data analysis can be performed, data must be cleaned and, if necessary, transformed[1]. The data preparation procedure for the work at hand comprises of the following key activities: *outlier detection and removal*, *handling of missing data* and, finally, *correction for trend and seasonality*.

Outliers. The univariate distribution of the total delay of a train and the change in delay cannot generally be expected to be symmetric. Positive delays occur more frequently than negative ones. The latter would imply that a train runs faster than necessary to maintain punctuality, which is inefficient and generally unwanted, whereas positive delay is an unwanted though inevitable phenomenon. Analyses of train delays of passenger trains in the Netherlands [2, 3] showed right-skewed distributions of delay across trains. In fact, delays for railway trains are commonly modeled as log-normal, exponential, Weibull, or gamma distributed random variables [4]. In our data, the same asymmetry can be observed (see Table 1). Accordingly, the empirical distributions of our random variables are skewed, and their location parameters and moments shifted unevenly. Hence, symmetrical outlier detection methods, like Tukey's fences [5], will falsely classify extreme observations on the heavy tailed side of the distribution as outliers and fail to detect outliers on the steep side. For this, [6] used the *medcouple* (MC) measure to adjust Tukey's fences for skewness. The medcouple, as proposed in [7], is a robust and efficient measure for skewness with a contamination breakdown barrier as high as 25%[2]. It can be calculated in $O(n\ log\ n)$ time. Thus, MC is a best compromise between robustness, complexity and skewness detection performance. For un-skewed distributions, both the basic and adjusted method yield the same results. However, if the data distribution is skewed, then the adjusted boxplot method accounts for the skewness, even in the case of contamination due to the presence of far outliers.



	$[S > 0.05]$	$[-0.05 \leq S \leq 0.05]$	$[S < -0.05]$
Total delay	3104	2965	247
Change in delay	1969	3328	1056

Table 1. Number of trains, which are either right-skewed, symmetrical, or left-skewed. The skewness *S* was measured by means of the medcouple, showing that the majority of trains show exhibit right-skewness or approximate symmetry for both d and Δd.

Using this method *13.95%* of all cases were marked as outliers caused by either of our interesting variables. Out of all observations for the delay upon arrival *5.87%* and for

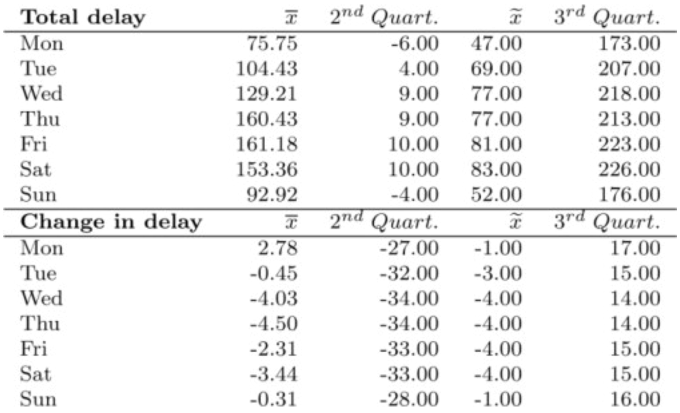
[1] Technologies like *bootstrapping* exist, where these preparation steps are not necessary, however, we expect them not to scale well.

[2] In 8 the asymptotic breakdown point of an estimator T is derived as $\epsilon^* = \epsilon^*(F_0, T) = inf\{\epsilon \mid b(\epsilon; X, T) = \infty\}$, where ε is the amount of deviant (i.e. contaminated) data, *b* the bias function, and ε^* the minimum level of for which the estimate bias becomes infinite.

change in delay *8.34%* where marked as outliers both of which lie well within the breakdown range of MC.

Missing Data. Based on expert knowledge, many occurrences of missing observations in the dataset have been rectified prior to the import into the *RDBMS*[3]. However, missing values for both variables persist. In order to decide upon the correct means of dealing with missing values, first, the cause of their absence must be identified. In data mining, three mechanisms of missing data are distinguished: *missing at random* (MAR), *missing completely at random* (MCAR), *missing not at random* (MNAR) [9]. If the occurrence of missing values is completely unrelated to the manifestations of the variable itself and other observed data, the underlying mechanism is MCAR and can be considered as *ignorable* [10]. The occurrence of missing values in the data seems unrelated to other relevant variables in the data set and is due to the lack of contradicting evidence assumed to be caused by MCAR. MCAR with as low occurrence rates (*3.26%* of the values for *delay upon arrival* and *4.95%* of the values for *change in delay*) as in the present data can be treated with the *complete-case* method, by which only complete observations will be regarded in the analysis[4].

Trend & Seasonality. We can identify systematic differences in delays and the delay differences between days of the week. On average, the total delay upon arrival at an operation point peaks on Fridays. Analogously, the ability to reduce delay seems to



Total delay	$\overline{x}$	2^{nd} *Quart.*	$\widetilde{x}$	3^{rd} *Quart.*
Mon	75.75	-6.00	47.00	173.00
Tue	104.43	4.00	69.00	207.00
Wed	129.21	9.00	77.00	218.00
Thu	160.43	9.00	77.00	213.00
Fri	161.18	10.00	81.00	223.00
Sat	153.36	10.00	83.00	226.00
Sun	92.92	-4.00	52.00	176.00
Change in delay	$\overline{x}$	2^{nd} *Quart.*	$\widetilde{x}$	3^{rd} *Quart.*
Mon	2.78	-27.00	-1.00	17.00
Tue	-0.45	-32.00	-3.00	15.00
Wed	-4.03	-34.00	-4.00	14.00
Thu	-4.50	-34.00	-4.00	14.00
Fri	-2.31	-33.00	-4.00	15.00
Sat	-3.44	-33.00	-4.00	15.00
Sun	-0.31	-28.00	-1.00	16.00

Table 2. Average total delay and change in delay by days of week.

[3] Relational Database Management System: *Maria DB* (https://mariadb.org/).

[4] We have added binary dummies, representing missingness in the total delay and in the change in delay, and then checked, whether they are correlated with each other and with the original values. Furthermore, we have performed a pairwise contingency test for the dummies and cardinal variables (train type, source station), and only found evidence for a, though significant, rather weak association between them (*Cramér's* $v \leq .30$). We found that missingness for the total delay and change in delay are strongly connected ($\rho = .782$). In most occurrences of missingness in either variable the other was missing as well, which supports the use of the complete case method. At the same time, missingness appears to be virtually unrelated to the absolute values of the delay variables.

decrease (see Table 2). The Wilcoxon rank sum test to compare medians of unpaired samples shows that the daily means are significantly shifted[5]. In addition, temporary rescheduling might lead to a-cyclical local and global trends, which must be addressed. To model the described trends and weekly patterns, we use an additive component model. The decomposition is achieved using a combined approach called *seasonal-trend decomposition procedure based on LOESS* (STL), as presented in [11].

2.3 Data Engineering: Cumulated Delay, Train Encounters

As the change in delay of a train after its departure shall be analyzed, data rows which constitute a departure process are merged in order to obtain the cumulated *change in delay*. Furthermore, as this study focuses on associations between train delays at specified operation points and at specific times, encounters of one train with another are obtained by matching the time of arrival of an incoming train with the time of departure the departing train at an operation point. Lastly, relevant and valid subsets must be selected with respect to the size of the resulting sub-samples.

Cumulative Change in Delay. The process of a train arriving and departing at an operation point is established as a sequence of events connected by activities called an *activity-event network* ([12, 13], similarly). A simplified version of such a network is depicted in Figure 2. The nodes in the graph represent signal passing events and links between event nodes indicate the transitions from one activity to the other, e.g. the train moving from signal to signal. Each transition can result in a change in delay Δd. The total delay d of a train i upon arrival can be represented as the difference between its scheduled time of arrival $\hat{t}^{a}{}_{i}$ and its actual time of arrival $t^{a}{}_{i}$, at that time: $d_i(\hat{t}^{a}{}_{i}) = t^{a}{}_{i} - \hat{t}^{a}{}_{i}$ This approach is rather straightforward and the resulting delay indiscriminately encompasses primary and secondary delays, which the train in question has accumulated during its course up until arriving at the respective operation point. The change in delay during the train's departure follows a slightly more complex pattern. As the reader can see in Figure 2, a change in delay can be the result of transitions (2) or (3) simultaneously. The delay change during one transition from t_k to t_j, $\Delta d_i(t_k,t_j)$, can be calculated as $\Delta d_i(t_k,t_j) = d(t_k) - d(t_j)$. For the work at hand, it is considered negligible which transition causes the build-up or decrease in delay, given that it is induced by the delayed arrival of another train. We accept the possibility for causes of the change in delay (e.g. passengers boarding and alighting), however, abstract from it. Because there is a systematic element in how the change in delay is distributed across phases, only the combination of all phases is considered. For simplicity, in this work a linear combination, i.e. the sum, will be applied: $\Delta d_i(t_k, t_{k-n}) = \sum_{j=1}^{n} \Delta d_i(t_k, t_{k-j})$, where $n \in \mathrm{N} \wedge n < k$ Putting the above equations together and simplifying yields the *cumulated change in delay*: $\Delta d_i(t_k,t_{k-n}) = d(t_k) - d(t_{k-n})$.

[5] As can be guessed from Table 2, the medians for Fridays and Saturdays as well as Wednesdays and Thursday are not significantly different.

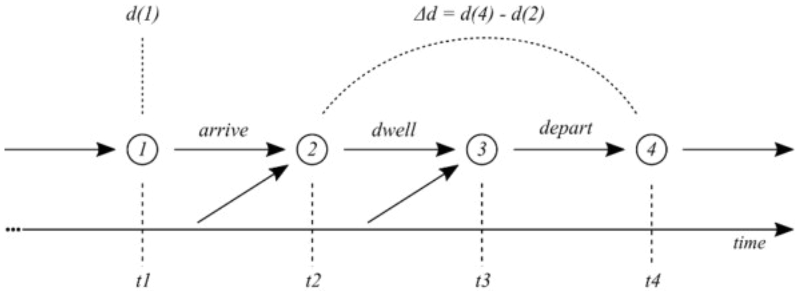


Figure 2. In (1) the train leaves the section prior to the station and enters its area. The train then arrives at the platform/station track (2) and leaves the platform in (3). In (4) the train leaves the station area for the next section. Between events (2) and (3) the train dwells at the platform for boarding and alighting or a conductor change etc. Source: own compilation based on [13].

Train Encounters. Extracting actual train encounters is an important preparatory step for the analysis. The following gives a concise definition of what is meant by the term *encounter*: For the purpose of this work, the arrival of one train *i* and the departure of a train *j* are considered an encounter, if the target operation point of *i* equals the start operation point of *j*, and, if the actual arrival time of i, t_i^a, was less than 10 minutes before the time of departure of j, t_j^a. Following this understanding, encounters imply temporal precedence, e.g. it is logically assumed that a train can only pass its delay on to trains, which depart at a later time, and never to trains that have already left the operation point. This assumption generally holds, however, the opposite direction is not inconceivable, as the dispatcher may always take action to create a situation in which one train passes its delay on to a preceding train and thus works as a time forwarding transmitter. In the further analysis, this inversion does not constitute an encounter.

Exclusion of Small Sub-Samples. For this paper, we consider only pairs of trains which have a minimum number of *30* encounters, using the method presented in [14] to determine exact sample sizes. These sizes depend on pre-estimated correlation coefficients, which are determined either through expert knowledge or prior research. For this work, the overall correlation between the d and Δd serves as an estimate for the expected magnitude of ρ and τ.

3 Constructing the Delay Transmission Network

3.1 Pearson's Product-Moment Correlation Coefficient ρ

For continuous variables which are at least interval-scaled *Pearson's product moment correlation coefficient* is a measure of choice. It gives the change in a random variable X which coincides with an increase or a decrease in another variable Y and vice-versa.

The coefficient is defined for the interval $[-1,+1]$ and measures the strength and direction of a linear association between two variables. A negative sign indicates an anti-proportional relationship and a greater absolute value implies stronger association between the variables. If the coefficient is equal to 0, the two features do not exhibit any *linear* relationship. However, a non-linear function to describe their relationship might still exist [15]. The product-moment correlation coefficient for random variables X and Y is given by $\rho(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$ where x and y are realizations of the variables. Thus, the interpretation of the calculated coefficients is rather straightforward: Let $X = d^k{}_{ij}$ denote the absolute delay upon arrival of a train i at a station k in an encounter with a train j. Furthermore, let $Y = \Delta d^k{}_{ij}$ signify the change in delay of that train j in an encounter with the preceding train i. Then $\rho(X,Y)$ gives the amount of time by which $\Delta d^k{}_{ij}$ would increase or decrease, if $d^k{}_{ij}$ were to be increased or decreased by one second. In other words, ρ is the extent to which the delay of train i proportionally influences the aggregate delay build-up (or reduction) of train j during dwelling and/or departure. $\rho(X,Y) = 1$ means that an increase or a decrease by one second of delay in train i fully translates into an increase or a decrease in the change in delay by that same amount and $\rho(X,Y) = -1$ means the exact opposite.

3.2 Kendall's Rank Correlation Coefficient τ

For ordinal variables, *Kendall's rank correlation coefficient* is an appropriate measure. Its interpretation is very similar to Pearson's coefficient, however, it does not measure the linear relationship between two features, but whether both variables share the number of discordances, instead. We use a modified version of the coefficient, *tau-b*, which accounts for rank ties. The rank correlation coefficient τ_b, corrected for the bilateral presence of ties is defined as $\tau_b(X,Y) = \frac{S}{\sqrt{(\frac{n(n-1)}{2}-T_X)}\cdot\sqrt{(\frac{n(n-1)}{2}-T_Y)}}$ where T_X and T_Y denote the number of ties in the rank pairs of X and Y (in adaptation of the formula given in [16]). In this paper, Kendall's τ will serve as a verification instrument for Pearson's coefficient, as it is non-parametric and, other than ρ, is resilient even in the presence of far outliers [17].

3.3 Graph Theory & Network Analysis

This paper employs elements of graph theory and network analysis to further analyze relationships between trains. While a multitude of applications for network analysis in exploratory data analysis exists, if the network properties of the available data are evident [18], this toolbox has also found use in areas where graphical structures in the underlying data are less ostensible. Examples can be found in qualitative studies where covariates to an interesting outcome might exhibit multiple moderation effects [19], and, for some time now, in genetics where network-graphs are created based on correlations as a similarity measure for gene-expression states ([20, 21]). The foundation of a network structure is a graph G which comprises of *vertices* and *edges*. If two vertices are connected by an edge, they are called *adjacent*. Hence, G can be represented by its *adjacency matrix* $A_G = [a_{ij}]$ in which each a_{ij} represents the number of connections between two nodes i and j. Every graph H with vertices and edges from G restricted by the adjacency matrix A_G is a *subgraph* of G. A network is a graph-based structure and its interpretation specific to the application-domain – the *network graph*. It is formed from *nodes* (vertices) which are connected through *links* (edges) and extends the graph model with additional attributes for vertices and edges. A typical addition in the network context is the assignment of weights w_{ij} to each link. A *weight matrix* W_G would have the same shape as the adjacency matrix of the underlying graph. A network, in which the links have weights of arbitrary value, is called a *weighted* network. If links in the network connect ordered pairs of nodes exclusively, then the network (graph) is *directed* [22].

3.4 Constructing the Delay Transmission Network

The proposed network is based on a graph consisting of a set of vertices ($v_i \in V$) which each represent a respective train (number). Relationships between trains are expressed through a set of directed links $e_{ij} \in E \; \forall \; (v_i, v_j) \in (V, V)$ in the network graph. These directed links represent the dependency of the target node's change in delay from the source node's total delay. Weights in the constructed network are based on the correlation coefficients. In the constructed network graph, link weights are interpreted as similarities or the relative closeness between adjacent nodes. Hence, on the one hand, while negative weights are not generally inconceivable (for example [23]), they are implausible in the current application context. Negative values for the correlation coefficients, on the other hand, are very plausible, and must be dealt with prior to performing the network analysis. Otherwise, the resulting negative and positive weights might bias weighted and distance-based network measures. The application of soft-thresholding produces weights for the proposed network. In this work they are obtained based on the arithmetic means of the correlation coefficients γ: Let γ be the mean of the correlation coefficients ρ and τ and let exist an arbitrary number $\lambda \in <$ then the weight w_{ij} for the link connecting the nodes i and j is given by $w_{ij}^{p} = |\gamma(v_i, v_j)|^{\lambda}$. The resulting values fit the interval *[0,1[*, thus, preserving information on the strength of the respective correlation. Additionally, the exponent λ is included. Performing this operation from the Tukey-ladder of power-transformations [5] adds the ability to reduce tail weights in the coefficients' distributions.

3.5 Measuring Network Properties

In this paper, we use the node *strength* as an indicator for the importance of a node. The node strength respects the strength of ties with other nodes in the network and calculates as the sum of link weights [24]. Of particular interest for this paper is the out-strength $D_{in}(p_k) = \sum_{i \neq k} w_{ik}$ and in-strength $D_{out}(p_k) = \sum_{i \neq k} w_{ki}$ – analogous to the degree-measures.

An adequate means to identify possible moderation-effects, which certain trains might have, is betweenness centrality (C_B) as detailed in [25]. It can be considered a measures of a node's overall connectedness with the network[6]. The betweenness $b_{ij}(p_k)$ of a node *k* with respect to two other nodes *i* and *j* is defined as the ratio of the number of paths between *i* and *j* which contain *k*, and the number of all paths connecting *i* and j: $b_{ij}(p_k) = \frac{1}{g_{ij}} \times g_{ij}(p_k)$. *Betweenness Centrality* (BC) is the sum of a node's betweenness for all node pairs formed from the $n - 1$ other nodes in the network: $C_B(p_k) = \sum_{i=1}^{n} \sum_{j \in [1,i[} b_{ij}(p_k)$

4 Evaluation

In the following section, we take a look at some select examples and evaluate the approach. We have implemented it using *GNU-R*[7]. Furthermore, we have used *Gephi* and *Inkscape* for visualizations.

4.1 Results

In the subset around the four selected operation points, we can observe the following data as follows.

Number of Trains by Operation Point. Figure 2 shows the number of trains arriving and departing at the respective operation points and the distribution of different train types. As was expected, *Mannheim* and *Karlsruhe* handle much more than twice as much traffic as the two smaller operation points.

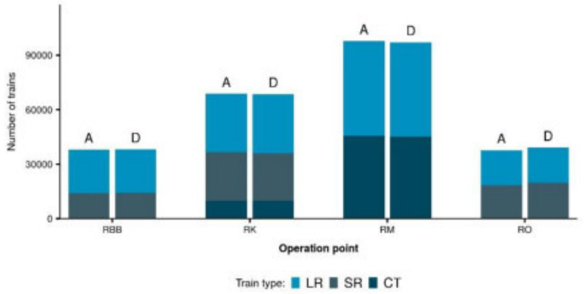


Figure 2. Number of trains arriving (A) at and departing (D) from the selected operation points. Source: own compilation.

[6] The concept of betweenness was originally used as a measure of a person's ability to influence a group. A high betweenness would mean that a person is able to control the flow of information in the network.

[7] To name the central libraries: *reshape2*, *RMySQL*, *dplyr*, *broom* (data handling); *robustbase*, *stlplus* (outlier removal, detrending, de-seasonalization); *igraph* (network modeling); *ggplot2* (visualization).

Number of Encounters by Operation Point. Accordingly, we were able to extract encounters, as presented in Figure 3. The smaller operation points, *RO* and *RBB*, on the one hand, show a similar distribution having most trains being involved in well less than 1000 encounters. The larger operation points, *RM* and *RK*, on the other hand, appear to be very different. The *Mannheim* plot looks like a scaled version of that of the former two. At *RK*, most trains have even fewer encounters than in the smaller operation points, and at the same time, very few trains have a great number of encounters. This would appear to be a result of regular strong peaks in the number of arriving and departing trains.

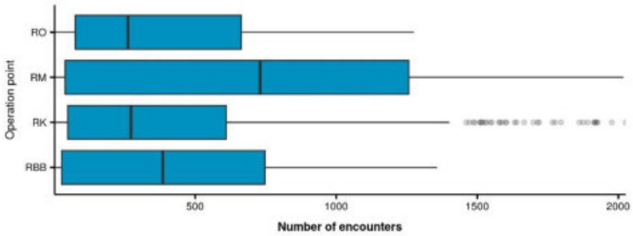


Figure 3. Number of encounters by train at the selected operation points. Given is the count of all encounters a train is involved in (as both arriving and departing). Source: own compilation.

Correlation Coefficients. The Kendall and Pearson correlation coefficients' distributions show similar patterns, with the median indicating a weak negative correlation. Both have a cluster in the weak to medium positive correlation range. In Figure 4, these train couples appear as outliers on the right side of the scale. However, Kendall's τ doesn't reach the extremes of the scale $(-1,1)$ as much as ρ does.

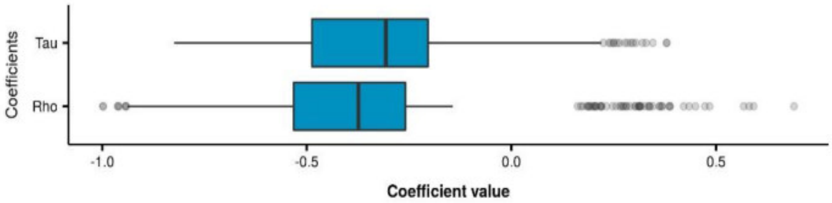
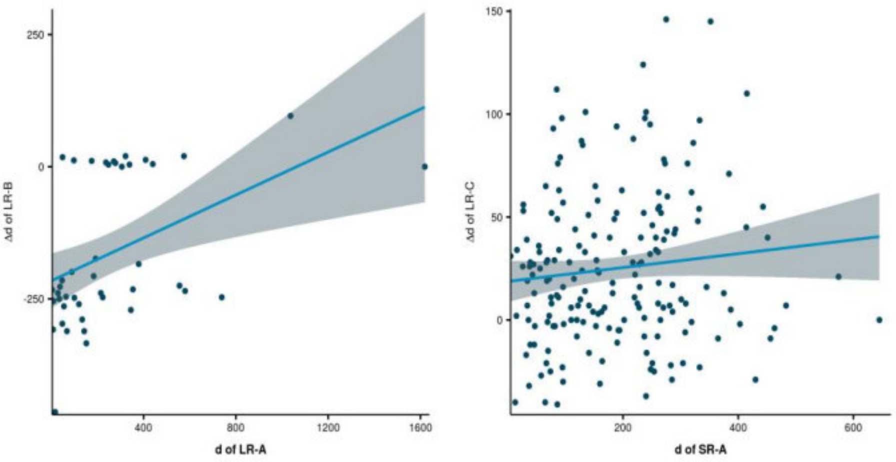


Figure 4. Comparison of the two correlation measures. Source: own compilation.

4.2 Examples

To test the approach, the described network was constructed for the *Mannheim* operation point. As for the link weights, we regard only significant correlation coefficients for τ (with $\alpha = 0.05$) and with maximum CI range of 0.3 for ρ. The resulting density distribution is almost Gaussian with a mean of -0.12. There are several strongly connected trains for which an increase in delay upon arrival in one train coincides with an increase in the change in delay in the other. In Figure 6, the relationship between different trains is presented. There clearly exists a positive relationship between delay upon arrival and the change in delay. Figure 5 (a) plots the total delay of long-distance passenger train *LR-A* against the change in delay of another long-distance passenger train *LR-B*. Figure 5 (b) plots the delay of a regional train *SR-A* against the delay change of a long-distance passenger train *LR-C*. Both situations seem plausible. The trains' respective planned arrival and departure in each coupling are at least 7 minutes apart, qualifying as a connection. While the correlation between the trains in pair (a) is relatively strong by comparison, that in (b) finds week support, as the dots are rather unevenly distributed. It is possible, nonetheless, that the correlation in (b) is due to a feeder train relationship, i.e. that *SR-A* is a feeder for *LR-C*. Table 3 represents the network characteristics of our examples. The out-strength of all long-distance passenger trains is in the 3rd and 4th quartile (median: 0.32) of the strength distribution for *Mannheim* and can be considered as medium to highly influential for this operation point. Most trains, which exhibit strong outgoing links, are long-distance passenger trains as well. *LR-A* is also a highly influential train; however, it has incoming links. Yet, it is the only train with a link to *LR-B*, which is the target of only one incoming edge. This example is remarkable, as *LR-B*, on many days, has no change in delay, at all. On other days, it reduces delay, probably prompted by its own delays. However, its ability to decrease delay seems to be negatively related to the total delay of *LR-A*. Similarly, *LR-C*'s change in delay is correlated only with the total delay of *SR-A*. However, the latter exhibits a higher in-strength and little out-strength. In the respective figure, we can ascertain that the correlation is rather weak ($\rho = .123$ and $\tau = .195$), yet significant.



(a) LR-A on LR-B

(b) SR-A on LR-C

Figure 5. Scatter plot for encounters of trains with positive correlation in *Mannheim*. LR:= long-distance train; SR:= regional train. Source: own compilation.

	Out-Strength	In-Strength	Out-Degree	In-Degree	Betweenness
LR-A	0.68	0.48	3.00	2.00	0.00
LR-B	0.51	0.30	2.00	1.00	0.00
SR-A	0.16	0.57	2.00	2.00	0.00
LR-C	0.65	0.16	2.00	1.00	0.00

Table 3. Exemplary trains and their network characteristics.

Betweenness values are generally very small. With means at *.012* and *.031* and maximum values at *.17* and *.005*, moderation effects for the delay transmission appear negligible. This seems to be due to the fact, that the network is not well connected. The observable formation of cliques indicates that the transmission of delays is restricted to certain "areas" of times during the day.

4.3 Validation

To validate the results, we have extracted a sub-sample from the train encounters by randomly selecting *70%* of all encounters of each pair of trains. This serves as the training set on which we perform the analysis, as described above. The remaining *30%* serve as our test set. Figure 8 shows the distributions of differences at the four focal operation points. As can be seen, the larger operation points *Mannheim* and *Karlsruhe* exhibit slightly higher errors for all train pairs (i.e. lower accuracy) and higher variance in means than the smaller ones. However, the means are kurtotically centered around a gravity center close to zero. The overall *mean squared error* (MSE) for the predicted *Δd* is *149,999.59* (vs. *265,952.12* with mean values). Many of the individual models perform very poorly, however, some perform well - with the best R^2 of *70.70%*.

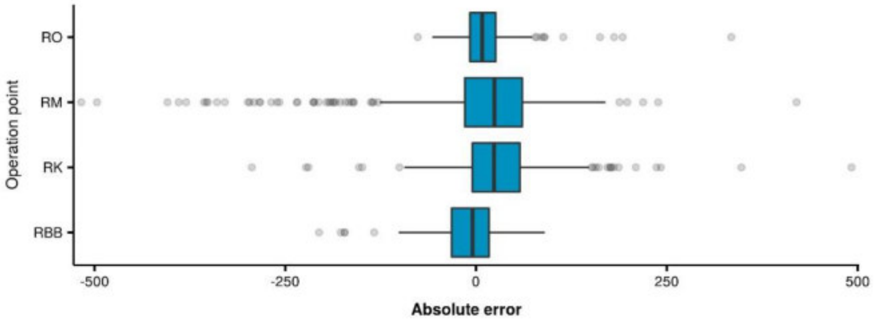


Figure 8. Boxplots for the distribution of differences for predicted vs. observed Δd for each pair of trains at the respective stop (10%-trimmed). Predictions were obtained by multiplying the total delay of the incoming trains with the estimated Pearson correlation coefficients.

5 Discussion & Conclusions

As was stated, as of now, no distinction is made between secondary and primary delays. This is simply due to the fact, that recognizing true secondary delay build up, would go beyond the scope of the research task completed in this work. Discerning primary and secondary delays requires additional modeling, an approach that is discussed further in this section. Furthermore, our understanding of negative correlation coefficients is somewhat inconclusive. A working hypothesis is that negative correlations indicate discontinuities in the development of the change in delay, such as train order swaps; such that up to a certain arrival delay, transmission occurs until a threshold is reach, when transmission decreased towards *0*. The correlation coefficient might then be negative. In addition, further sophistication in the data preparation process or just broadening the data selection might consolidate interpretability and facilitate a better understanding. In the work at hand, an approach for the analysis of train delay propagation was demonstrated. As a result, train interactions can be determined for selected railway networks. These will be used as inputs at subsequent project stages, where we plan to use these inputs to retrieve a computation order for a by-train-optimization of time supplements. Further validation and verification will be part of further project stages. This involves expert evaluations and the inclusion of information on passenger movements, which has not been available to us, so far.

6 Acknowledgements

We are thankful for the historical operational data kindly provided by *DB Netze* and the support we have received.

References

1. Ghofrani, F., He, Q., Goverde, R.M.P., Liu, X.: Recent applications of big data analytics in railway transportation systems: A survey. Transportation Research Part C: Emerging Technologies 90, 226–246 (2018)
2. Goverde, R.M.P.: Punctuality of railway operations and timetable stability analysis. Netherlands TRAIL Research School (2005)
3. Yuan, J.: Stochastic modelling of train delays and delay propagation in stations. Eburon Uitgeverij BV (2006)
4. Schranil, S.: Prognose der Dauer von Störungen des Bahnbetriebs. ETH Zurich (2013)
5. Tukey, J.W.: Exploratory data analysis. Reading, Mass (1977)
6. Hubert, M., Vandervieren, E.: An adjusted boxplot for skewed distributions. Computational statistics & data analysis 52, 5186–5201 (2008)
7. Brys, G., Hubert, M., Struyf, A.: A robust measure of skewness. Journal of Computational and Graphical Statistics 13, 996–1017 (2004)
8. Huber, P.J., Ronchetti, E.M.: Robust statistics. John Wiley & Sons, Inc, New York (2009)
9. Little, R.J.A.: A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 83, 1198–1202 (1988)
10. Rubin, D.B.: Inference and missing data. Biometrika 63, 581–592 (1976)

11. Cleveland, R.B., Cleveland, W.S., Terpenning, I.: STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics 6, 3 (1990)
12. Schöbel, A.: Integer programming approaches for solving the delay management problem. In: Algorithmic methods for railway optimization, pp. 145–170. Springer (2007)
13. Büker, T., Seybold, B.: Stochastic modelling of delay propagation in large networks. Journal of Rail Transport Planning & Management 2, 34–50 (2012)
14. Bonett, D.G., Wright, T.A.: Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika 65, 23–28 (2000)
15. Taylor, R.: Interpretation of the correlation coefficient: a basic review. Journal of diagnostic medical sonography 6, 35–39 (1990)
16. Szmidt, E., Kacprzyk, J.: The Spearman and Kendall rank correlation coefficients between intuitionistic fuzzy sets. In: EUSFLAT Conf, pp. 521–528 (2011)
17. Long, J.D., Cliff, N.: Confidence intervals for Kendall's tau. British Journal of Mathematical and Statistical Psychology 50, 31–41 (1997)
18. Butts, C.T.: Revisiting the foundations of network analysis. science 325, 414–416 (2009)
19. Voth-Gaeddert, L., Cornell, D.: Improving health information systems in Guatemala using weighted correlation network analysis. In: Global Humanitarian Technology Conference (GHTC), 2016, pp. 686–693 (2016)
20. Ruan, J., Dean, A.K., Zhang, W.: A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC systems biology 4, 8 (2010)
21. Zhang, B., Horvath, S.: A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 4, 1–45 (2005)
22. Bondy, J.A., Murty, U.S.R., others: Graph theory with applications. Citeseer (1976)
23. Schwarz, A.J., McGonigle, J.: Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. Neuroimage 55, 1132–1146 (2011)
24. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. Social networks 32, 245–251 (2010)
25. Freeman, L.C.: Centrality in social networks conceptual clarification. Social networks 1, 215–239 (1978)