

Spatial, Temporal and Size Distribution of Freight Train Delays: Evidence from Sweden

Niclas A. Krüger*^{a,b,c}
Inge Vierth^{a,b}
Farzad Fakhraei Roudsari^{b,c}

^a Centre of Transport Studies, Stockholm

^b Swedish National Road and Transport Research Institute

^c TRENOP, KTH Royal Institute of Technology, Sweden

*Corresponding author:
Phone: +46 70 33 54 599
E-mail: niclas.kruger@vti.se

Abstract: This paper analyzes how freight train delays are distributed with respect to size, location and time of their occurrence. Arrival delays are analyzed in detail using data covering all freight train departures and arrivals during 2008 and 2009 in Sweden. Moreover, the link between capacity usage and expected delay is analyzed using the fact that demand fluctuates on different time scales, especially due to the economic chock in 2009. Since the distribution of delays on different scales describe reliability and vulnerability in the rail transport system, the results have potentially important policy implications for rail investment appraisal.

Keywords: Fat-tails; Freight; Delays; Rail; Reliability; Vulnerability

1 Introduction

Trains that are either delayed, early or cancelled constitute a problem for companies that transport goods by rail. Lack of reliability in the rail transport system means that rail transports have a disadvantage, all things equal, compared to the other modes. Generally reliability is defined as deviation of the transport time compared to the scheduled transport time. To date there is only limited quantitative information in Sweden about freight trains not arriving on time, with major consequences, since without information it is not possible to target appropriately the problem of unreliability. To some extent the lack of reliability is taken into account by planners by building in slack in the scheduled arrival times. However, too much slack will eventually lead to many train arrivals being early, which is also a problem for scheduling the unloading and for the additional storage space and labor required. Cancelled trains (that are often not included in delay statistics) can be seen as very long delays that can cause high costs for firms sending and receiving goods.

Except from extreme situations like very strong winters or large scale accidents, there is also limited knowledge about the reasons for the trains not being on time and the consequences for the rail transport operators and the companies sending and receiving goods. There is moreover limited knowledge about how the capacity utilization, track maintenance and investments in infrastructure influence the probabilities and size of delays. Many networks, because of interdependence in the links, have shown to be vulnerable to extreme events, that is, infrequent but large negative impacts on the functioning of the networks, which makes it hard to identify single causes and remedies to the problem.

This means that benefits of improved reliability, that is, costs saved due to reduced variability in transportation time resulting from investments in rail infrastructure, are not taken into account in a proper way in cost-benefit analysis carried out in Sweden and elsewhere. The benefits of reduced variability in freight transport are not calculated at all or calculated in a very simplified way. A more accurate calculation of the benefits, that is, estimating the value of transportation time variability (short VITV) requires several steps: a) identification and description of deviations from scheduled times, b) finding appropriate ways to measure benefits from reduction of deviation size, probability and variability, and c) monetary valuation of the benefits.

The Swedish Transport Administration (Trafikverket) has developed a monitoring system covering all trains in the Swedish national rail system. The database includes information about deviations (actual time compared to scheduled time) of both freight trains and passenger trains at

a detailed rail segment level. To date information for 2008 – characterized by an economic boom and 2009 – characterized by a recession due to the financial crisis in late 2008 – is available.

This paper studies the distribution of freight trains arriving late at final stations. Trains that depart too late, arrive too early or are cancelled are not analyzed in detail. We investigate reliability – or lack thereof – of rail transportation using a descriptive approach, which is, as we argue above, a prerequisite for valuation of reduced transport time variability in cost-benefit analysis and hence for appraisal of infrastructure investments.

This paper is structured as follows: Previous research is summarized in section 2 and the data obtained from the Swedish Transport Administration is described in section 3. Section 4 analyzes the data. In section 5 we analyze the implications for the estimation of the value of transportation time variability (VTTV) and discuss general policy implications. Section 6 concludes the paper.

2 Previous research

The paper addresses several topics that have been studied in detail by previous papers. In order to keep the section tractable, we divide the literature in different topics.

Distribution of delays: The distribution of delays in rail networks is dependent on train routes and types of train, which may vary over time. Delay distributions can be used as an input for analyzing the robustness of timetables and reliability. However, finding a distribution which can generally be used is difficult. Schwanhäußer (1994) concludes that the tail of arrival delays follows an exponential distribution. This result is confirmed by Goverde et al (2001), Yuan et al (2002), Goverde (2005) and Haris (2006). The exponential distribution for arrival delays is therefore widely used as input for delay propagation models. More flexible distributions are sometimes used to capture the distribution of late arrival and departure times. For example, the Weibull distribution, the gamma distribution and the lognormal distributions have been adopted (Higgins & Kozan, 1998; Bruinsma et al, 1999; Yuan, 2006). Yuan (2006) compares the goodness-of-fit among several distribution models selected for train event times and process times, by fine-tuning the distribution parameters for data recorded at The Hague railway station. The Weibull distribution is found to be the best fitting distribution model for arrival delays, departure delays and the free dwell times of trains. The model can be used to determine the number of trains that can be accommodating in the network for a given accepted level of knock-on delays. Güttler (2006) fits a normal-lognormal mixed distribution to assess running times of trains between two stations using the data for the German railway. Briggs and Beck (2007) find in a study on the British railway network that the distribution of train delays can be described by so-called q-exponential functions (closely related to the exponential distribution).

Primary and secondary delays: Delays can be classified in primary and secondary (knock-on) delays. Lindfeldt (2008) studies the distribution of primary delays for passenger and freight trains on the heavily used route between Stockholm and Gothenburg in Sweden and finds that the primary delays are widely spread along the route. A further result is that a reduction of delays of high speed trains would not affect the overall delays in the network significantly, simply because high speed trains make up a small share of total rail transportation. According to Yuan (2006), the propagation of train delays mainly occur during the arrival or departure of trains at stations, since the crossing or merging of tracks are in most cases bottlenecks in highly used rail networks. For passenger trains, the bottlenecks of a rail network are generally situated at stations (Carey & Carville, 2000). An extensive amount of literature addresses sources, impact and nature of secondary delays. Schwanhäußer (1974) calculates an equation for additional delays as a

function of buffer time, initial delays, mix of priority classes, headways and overtaking possibilities. The exponential distribution of arrival delays is widely used as input for delay propagation models, including models for propagation of departure delays at stations. Carey and Kwiecinski (1994) estimate the effects of secondary delays between two trains on a single-track due to tight headway by nonlinear regression and heuristic approximations. Higgins and Kozan (1998) present an analytical model that quantifies the expected delay for a passenger train on a track section in an urban area. The authors assume that delayed trains follow with minimum headway speed when a conflict occurs, although in the real world delayed trains need to stop in front of red signals due to occupied route sections or platform tracks ahead. Yuan and Hansen (2007) improve the model developed by Higgins and Kozan by estimating knock-on delays of trains due to speed deceleration and acceleration and the use of track capacity. The model reflects speed fluctuations caused by signals, dependencies of dwell times at stations and stochastic interdependencies due to train movements.

Capacity utilization: The first model that addresses capacity in connection with assessment of delays was developed by Frank (1966). Frank (1966) studies delays on single tracks with both unidirectional and bidirectional traffic. The assumption is that only one train can occupy single-track lines between sidings and by restricting train speeds and travel time to be deterministic, the total number of trains that can travel in the network is estimated. Petersen (1974) improves the model by taking into account two different train speeds and by specifying delays as a function of single and partial double track lines. Petersen (1974) assumes that the departure time is independently and uniformly distributed for a defined period of time and that sidings have equal space and that there is a constant delay for every encounter between two trains. These assumptions are applied to construct a simple expression that shows the number of conflicts that a specific train will meet before arriving at the final station. Chen and Harker (1990) improve the model in Petersen (1974) by relaxing the assumption that trains are uniformly distributed. This makes it possible to provide single track delay functions for different train types with given train schedules and distributional information concerning operational uncertainties. Moreover, the model can estimate the delays for a given traffic flow using a stochastic approach. In order to determine the delay probability for a single conflict, the dispatcher's behavior is modeled, using a district choice model. The model can estimate the mean and variance of a train delay on a single track, which can be used to measure reliability of given schedules. They show that the mean delay and the variance have a direct relationship with inter-station distances (a shorter distance leads to lower means and variance). Harker and Hong (1990) include double-track lines and dynamic priorities, that is, train priorities dependent on expected delays. The

capacity utilization in the Swedish network is studied by Törnquist (2006), Nyström (2008) and Lindfeldt (2010).

Causes for delays: Nyström (2006) investigates the identification and categorization of causes of delay of a single-commodity train. Delays are categorized into short and long delays and the main finding is that many large delays are caused by failures due to infrastructure items such as turnouts and/or by adverse weather and that many short delays are caused by minor locomotive problems or low contact wire voltage due to many nearby trains. Nyström (2007) analyzes also how different employees at the Swedish Rail Administration assign codes to pictures that describe different delay situations. The finding is that there are quite large differences in how same types of delays are reported by different persons and hence, that there is a low level of consistency.

Infrastructure and delays: With respect to rail infrastructure, a range of analytical models focus on the operation of single or double tracks, capacity problems, delay propagation and re-scheduling on isolated lines or entire networks (Lindfeldt, 2010). Greenberg et al (1988) develops queuing models to estimate dispatching delays for a low speed, single track rail network with widely spaced passing locations. It is assumed that train arrivals follow a Poisson distribution and that trains follow each other with minimal headway. Using these assumptions they calculate the expected delays for single track segments.

Simulation methods for delay estimation: Murali et al (2010) develop a simulation-based technique to generate delay estimates as a function of traffic conditions and network topology. Their model is tested using delay data in the Los Angeles sub-network and the data can be used to measure expected delays on each train route. A given train can be scheduled on a route that has minimum expected delay. The dispatchers can schedule train routes using their delay estimation procedure in order to minimize the delay propagation. Marinov and Viegas (2011) develop a mesoscopic simulation modeling method for analyzing freight train operations. To capture the global impact of the freight train operations, the rail network is separated into lines, yards, stations, terminals and junctions. Interconnected queuing systems interact and influence each component. Marinov and Viegas study the network with freight train movements characterized by both insignificant deviations and significant deviations from schedules. Corman et al (2011) studies the problem of coordinating several dispatchers to find the optimal solution in case of disturbances. The objective is to minimize delay propagation; in order to do so the problem is formulated as a bi-level program. Constraints are imposed on the border of each dispatch area. Dispatchers have to schedule the trains within their area by producing a locally feasible solution compliant with the border constraints. The coordination problem can also be

solved by branch and bound procedures. The coordination algorithm has been tested on a large rail network in the Netherlands with busy traffic conditions, yielding a proven optimal solution for various network decisions. In simulation software like RailSys (Radtke, 2005), several processes related to delays can be modeled at a high level of detail: primary delays, conflicts (knock-on delays), catch-up effects etc.

Scheduling methods and delays: The optimization of capacity utilization and timetable design requires a certain degree of reliability of train operations. D'Ariano and Pranzo (2009) categorize the problems in rail traffic management into two major groups regarding time: a) the long term timetable design that should provide a robust schedule for a large network and a long time horizon and b) the real-time dispatching that considers the current timetable and deals with delay management. D'Ariano et al (2008) studies the implementation of the traffic management system ROMA (Railway traffic Optimization by Means of Alternative graphs) to optimize rail traffic also when the timetable is not conflict-free. Branch-and-bound algorithms are used to sequence train movements and local search algorithms are applied to optimize the rerouting processes. Corman et al (2010) evaluate the performance of the ROMA-system by inspecting the quality of the dispatch solutions when input data is manipulated by small stochastic variations. The authors find that the first-in-first-out-strategy in most cases is not able to solve conflicts without causing delay propagations. Optimized dispatching solutions, computed by the ROMA system, can on the other hand give a better delay reduction than straightforward dispatching rules. Mu and Dessouky (2011) compare different methods that can solve or improve the freight train scheduling problem and show that the heuristics algorithms provide better solutions than the existing procedures.

In summary, we conclude that the distribution and propagation of train delays as well as the impact of infrastructure and capacity utilization on delays has been analyzed during the last two decades using a wide range of methods. In many of the latest papers, scheduling and simulation methods are developed. This paper contributes mainly to the first category (Distribution of delays), even though we focus on freight trains and not on passenger trains. However, our data covers a complete national rail network, whereas earlier studies use a much more limited data. In addition, we analyze spatial and temporal distributions not previously studied. Moreover, we contribute empirically to the topics delay propagation and capacity utilization. The approach here is on a rather aggregate level in contrast to many other previous studies. However, the results from disaggregate models have implications for aggregate data and vice versa. Hence, the empirical evidence presented by this paper in section 4 provides new insights, inputs and validation cues for other research methods and research areas related to rail delays.

3 Data

According to official statistics, ton-kilometers transported by rail in Sweden decreased between 2008 (23.1 billion ton-kilometers) and 2009 (19.4 billion ton-kilometers), a decrease by 17 percent due to the economic bust following the financial crisis in late 2008. Sweden has a relatively high share of rail freight compared to other countries; the share is about twice as high as the average share for the European Union. Most of the rail infrastructure in Sweden is used by both passenger and freight trains.

The Swedish Transport Administration is responsible for the overall planning of the traffic on rail tracks and for the maintenance of the national rail network. This paper utilizes a database from the Swedish Transport Administration for 2008 and 2009 comprising all trains. The timetables for freight trains differ from the timetables for passenger trains as the number of freight trains has to be adjusted to fluctuating demand. Regular routes have to be booked several months in advance and can be cancelled with short notice (to date paths can be cancelled without costs). Ad hoc routes can be booked a short time before the departure of the train. Routes are allocated to trains using specific unique train numbers for trains that go between specified origins and destinations (OD-pairs). A destination can also be a marshaling yard, where wagons are recombined (and a new train number is used for the new train). Hence, it is not always possible to follow goods/wagons to the final destination within the rail system.

The database contains more than six million freight train observations per year. For each train, event times along the route are recorded, including arrival and departure time at the origin and destination and all sections in between. Codes for causes of delays and of cancellations (at the OD-level or section level) are also recorded in the database, but incomplete. Based on information in the database, OD-distance on scheduled and actual travel time, speed as well as arrival and departure deviations can be calculated.

In total we have 6,766,331 observations in the raw data for 2008 and 6,140,445 for 2009. We disregard observations that are obviously wrong (departure time after arrival time, trips with an average speed of more than 120 kilometers per hour), observations for cancelled trains and observations for very short distances. The purpose of this paper necessitates the inclusion of extreme values, so we do not exclude so-called outliers from the data. We focus on the arrival delay at the destinations, hence we disregard the data for segments between origin and destination; in total our data contains 157,537 OD-pairs with a unique train number for 2008 and 125,219 OD-pairs for 2009. As mentioned above, the train numbers do not necessarily represent the total route for the cargo from sender to receiver. The main variable of interest in this paper is

arrival delays, computed as difference between actual arrival time and scheduled arrival time at the destinations within the rail network. Hence, positive arrival deviation corresponds to late arrival and negative values to early arrival.

According to the earlier definition of the Swedish Transport Administration trains are considered to be on time if they arrive less than five minutes behind schedule at the terminal station, but recently the definition was changed to 15 minutes. Based on the earlier definition (5 minutes), 16 (15) percent of the freight trains in 2008 (2009) arrived in time at the terminal station (OD-pairs). 27 (25) percent in 2008 (2009) of the freight trains arrived too late and 56 (60) percent arrived too early. The high share of trains arriving before scheduled time illustrates the high amount of slack in the freight timetable. In 2008 5.8 percent of the freight trains were cancelled by the Swedish Transport Administration and in 2009 2.6 percent of the trains. The cancelled trains can cause severe extra costs for the operators and/or shippers similar as delayed trains.

In Figure 1 we show the frequency of different categories of causes for delays for the year 2009 (the results for 2008 are almost identical). We see that causes related to traffic management and to operators are most often mentioned as a cause. Another way of looking at causes of delay is to look at their share of total delay minutes (see Figure 2). Interestingly, although operator errors are more frequent, their share of total delays is less than for delays caused by traffic management, which account for 55% of all delay minutes but only 37% of all delay events. The five most mentioned delay causes due to operator error were in descending order late departure from freight terminal, train linkage, circulation/train turning, late from abroad and extra wagon service. The five most mentioned delay causes due to traffic management error are in descending order meeting train, train ahead, bypass, crossing the train route and scarcity of track.

Figure 1: Total number of delays and causes (share in %)

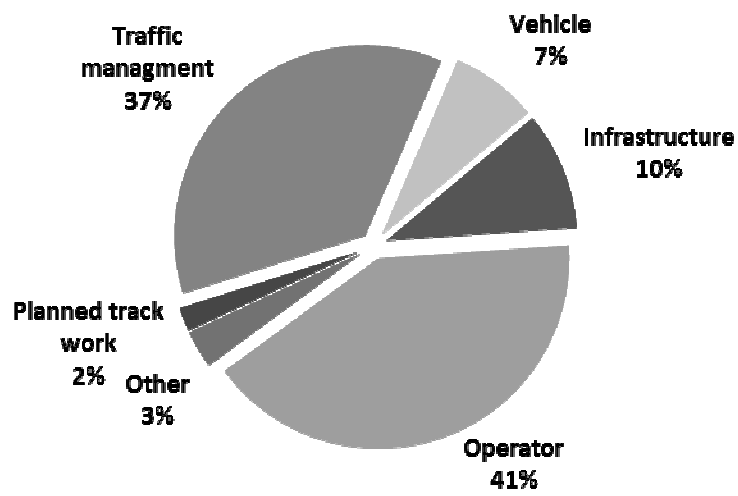
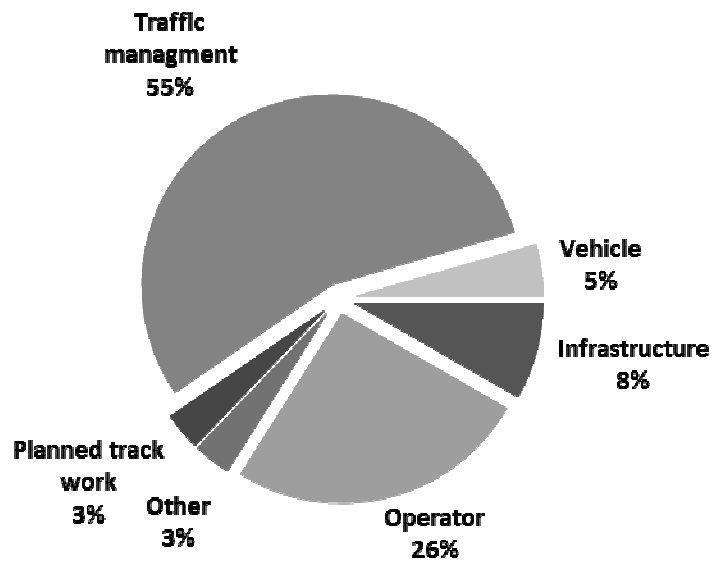


Figure 2: Total delay minutes and causes (share in %)



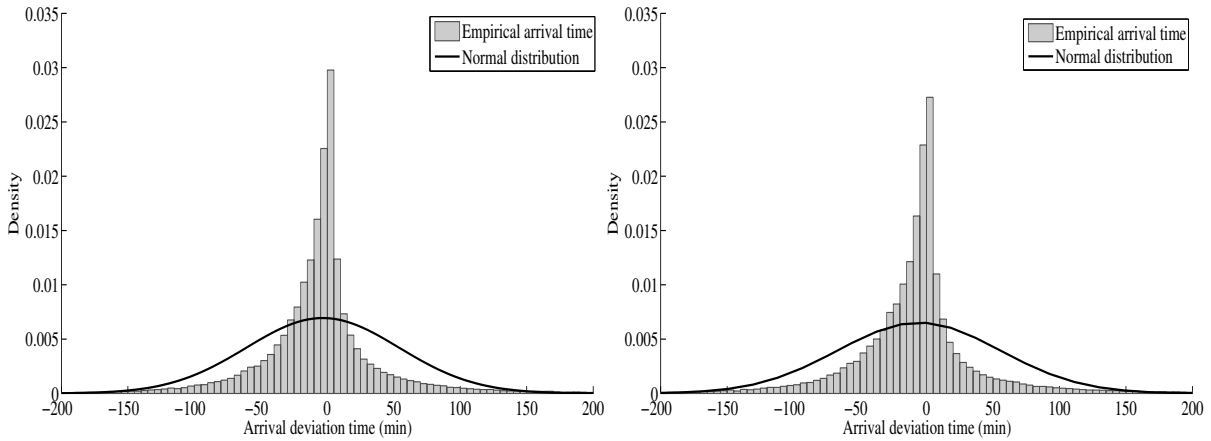
4 Analysis

4.1 Delay distribution on size-frequency scale

4.1.1 Analysis of percentiles

The histogram of freight train arrival deviations at final destinations exhibit non-normal properties; the histogram has a sharp peak around zero and fat tails (see Figure).

Figure 3: Arrival deviation at destination in 2008 (left) and 2009 (right)



In order to check whether many small deviations from timetable or a few extremely large deviations matter most to total freight train delay time, we examine different percentiles of arrival delays at the final destinations, in order to compute their share of total delay minutes. For example, by examining the 90th percentile we can see how much the 10% largest delays account for as percentage of total delay. Table 1 shows the share of each percentile as percentage of total delay minutes in 2009 (the results for 2008 resemble those for 2009).

Table 1: Percentile share in percentage of total delay minutes

Percentile	Total no. of observations	Average delay (min)	Percent of total delay
50 th	25090	65.89	95%
80 th	10026	127.74	74%
90 th	5022	187.65	54%
95 th	2507	259.42	37%
99 th	500	454.23	13%

For each percentile in the first column of Table 1, we show in the second column the total number of observations that fall within each percentile. The third column shows the average of

arrival delay within each percentile and the fourth column reports the share of each percentile of total delay minutes. For example, we can see that 74% share of delays in total are caused by only 20% of all observations. Regarding these calculations we see that extreme delays are important since their contribution to the total delay is larger than the many small delays. This result is stable across the two years included in our database.

However, for a quantitative measure it is useful to compute the kurtosis. Kurtosis measures the degree that values, much larger or smaller than the average, occur more frequently (high kurtosis) or less frequently (low kurtosis) than in a normal (bell shaped) distribution; that is, kurtosis allows you to assess whether the distribution curve has thin (few extreme events) or fat tails (many extreme events). For our data the kurtosis of arrival delays at destination is 24, in comparison, the (standard) normal distribution has a kurtosis of 3. This means that large delays are much more common in our data compared to a normal distribution.

4.1.2 Identification of distribution

Since we ruled out the normal distribution in the previous section, we want to explore what candidate distribution will fit our data better. Distribution fitting can be a valid method for several reasons: First, the distribution gives a hint on the mechanism behind the causes of delays or, to put it the other way around, any mechanism explaining delays should lead to a similar distribution as the empirical delay distribution. Related to this is that by means of distribution fitting it can be examined whether delays of different sizes are caused by the same mechanism. Second, distribution fitting allows us to predict probabilities for very low probability events (e.g. large delays). Third, examining the distribution helps us find a general pattern that potentially can be applied to and compared to regional and international data for delays.

There are different candidate distributions that exhibit fat tails. The lognormal distribution is the simplest one to check, since it implies that the natural logarithm of delays is normally distributed. Another candidate distribution is the so-called power-law distribution. For a continuous variable as train delays, the power-law distribution is defined as follows (Clauset et al., 2009):

$$f(x)dx = \Pr(x \leq X < x + dx) = cx^{-\alpha} dx \quad (1)$$

Where X is the observed value, $f(x)$ is the density function and c is a normalization constant. The normalization constant is needed since the density function diverges as x approaches zero. Therefore we must have a lower limit for the power law process. The complementary cumulative distribution function is defined as:

$$1 - F(x) = \Pr(x \geq X) = \int_x^{\infty} f(x)dx = Cx^{1-\alpha} \quad (2)$$

Hence, we have:

$$\frac{d \ln(1 - F(x))}{d \ln(x)} = 1 - \alpha \quad (3)$$

Therefore, we can use the slope in a log-log plot to estimate the scaling parameter a once we identify the lower limit. For identification purposes of the lower limit, the data is divided into different intervals and the lower limit is identified as the interval with the best power law fit. The method can be criticized on grounds that the OLS-estimation of the slope coefficient for a given lower limit does not imply that the estimated distribution parameters do satisfy the properties of a proper cumulative distribution function. Therefore, maximum likelihood estimation simultaneously of the power law coefficient and the lower limit is necessary. Furthermore, the power law hypothesis should be tested against the main alternative distribution that we regard to be the exponential distribution, which is defined as:

$$f(x)dx = \Pr(x \leq X < x + dx) = \beta e^{-\beta x} dx \quad (4)$$

The complementary cumulative distribution function is thus:

$$1 - F(x) = \Pr(x \geq X) = \int_x^{\infty} f(x)dx = e^{-\beta x} \quad (5)$$

Whereas we can identify a power-law by a straight line in a double-logarithmic plot, an exponential distribution will exhibit a straight line in a semi-logarithmic plot instead:

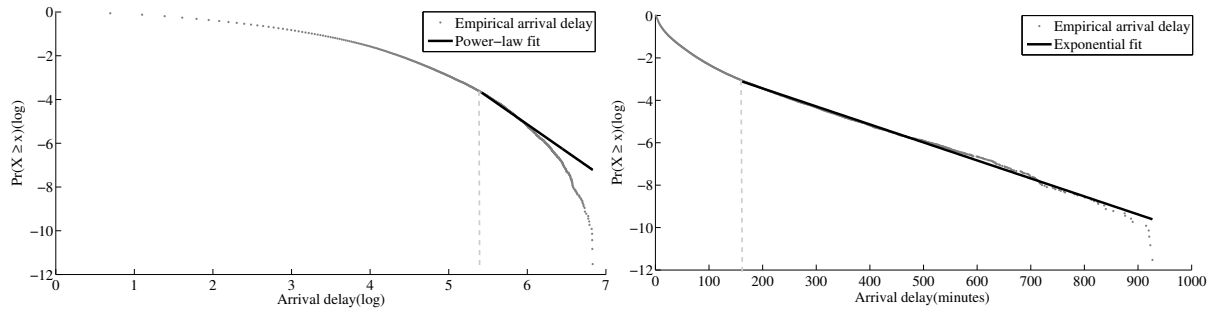
$$\frac{d \ln(1 - F(x))}{dx} = -\beta \quad (6)$$

There are related distributions as the power law with exponential cutoff and the stretched exponential. Hence, it is difficult to definitely rule out the power law or to distinguish it from the exponential distribution. The standard way to identify power-law behavior is by spotting a straight line if plotting the logarithm of the inverse cumulative distribution against the logarithm of seize (see Figure). We pool all observations in order to detect very large delays. As mentioned before, a power law distribution is necessarily bounded by a minimum value and hence applicable only to the tail of the distribution. Additionally, empirical distributions often exhibit signs of boundary effects so that power laws only can be confirmed for certain intervals.

Based on the method outlined in Clauset et al (2009) and based on maximum-likelihood estimation, different lower boundaries are selected and the power law coefficient is estimated; goodness-of-fit tests are used to select the best fitting lower boundary. Figure 4 shows the result of fitting the data to the power law distribution. It seems that the power law hypothesis cannot be confirmed. For very large delays the power law overestimates the probability for having a delay of that magnitude. However, we cannot rule out that the tail is power law distributed with an exponential cutoff, that is, a combination of both distributions. One reason for this differing

behavior for extremely large values might be a boundary effect, caused by the fact that very large departure delays are registered as cancellation of the train trip instead.

Figure 4: Power law and exponential distribution fit for arrival delays at final destination



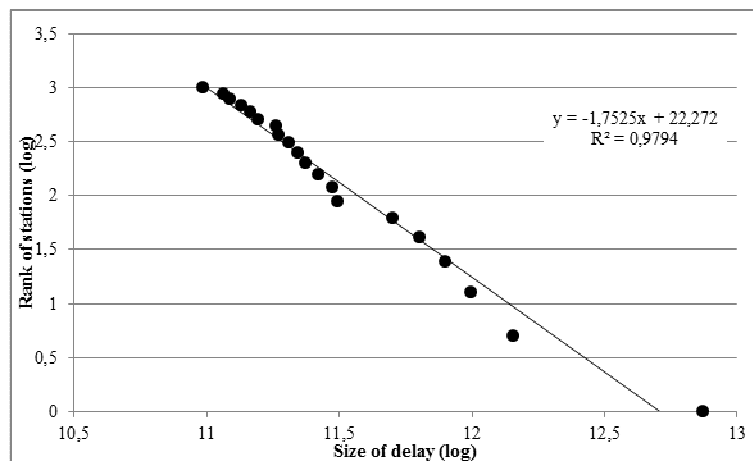
The exponential distribution gives a better fit for the extreme values in the tail (the power law coefficient is 3.212 and for the exponential distribution the coefficient is .584). The goodness-of-fit test confirms the visual inspection so that we conclude that the exponential distribution can be used to describe the tail of arrival delays of freight trains in the Swedish rail network.

4.2 Delay distribution on spatial scale

4.2.1 Rank-Size distribution of stations

In this part we examine arrival delays (at final destinations) regarding to their aggregated size in order to find a pattern of how the delay proneness is distributed among terminal stations. In characterizing and analyzing many phenomena, double-logarithmic rank-size distributions have been used. The reason is that it will unveil a linear pattern if it follows the discrete version of the power law. In order to compute the rank-size distribution, first we need to rank the size of total delays in a given station in a descending order and calculate the natural logarithm of the rank and delay size. Figure ranks the 20 stations which have the largest arrival delays. The station with the largest cumulative delay is assigned the rank 1 which is zero on the logarithmic scale.

Figure 5: Rank-size distribution for both years



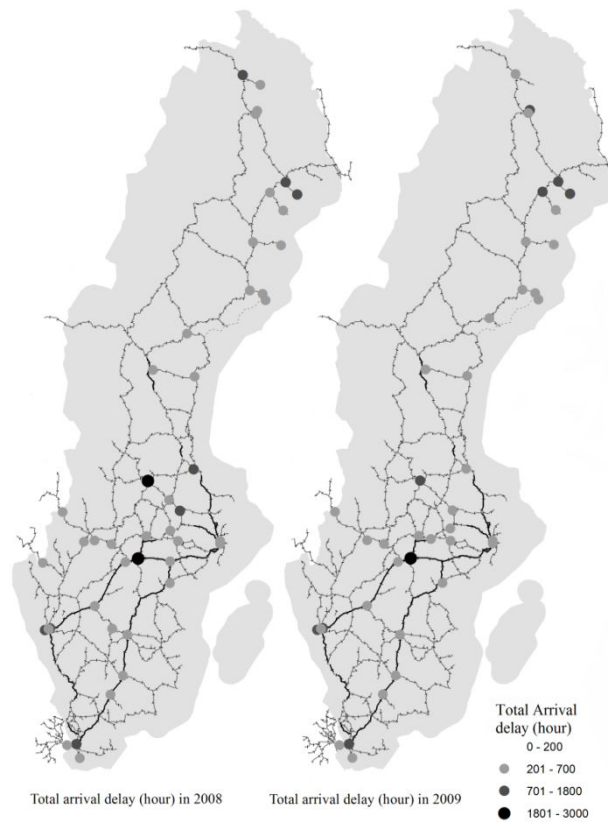
A double-log model is fitted to the data via OLS-regression and the very high value of R-square shows that the rank-size distribution can be described by the estimated linear relationship.

Table 2 ranks the 5 stations most prone to delay out of 270 stations in total. As can be seen, there are some differences in the ranking of stations for year 2008 and 2009. Although Hallsberg marshaling yard is the station in both years with the largest cumulative delay time followed by Borlänge, the following ranks change both with respect to the order of stations and the stations among the top 5 between the two years. Among the 20 stations most prone to delay we find 7 marshaling yards in 2008 and 6 marshaling yards in 2009 out of in total 13 marshaling yards in Sweden. The reason is possibly the high degree of connectivity of the marshaling yards with many different origins. The last column in Table 2 shows how many different origins are linked to the stations (in column 1) during the year. There are 52 (59) stations that have their destination in Hallsberg marshaling yard and 42 (43) that have their destination Borlänge in 2008 (2009). The significantly higher number of links associated to Hallsberg and Borlänge makes them central hubs in the rail network. Hallsberg alone experiences 10 percent of all delays, the top 5 stations together experience 25% off all delays and the top 20 stations (out of 270) experience more than 50% of all delays. These results reveal in what way the network is vulnerable in case any problem occurs in these connected stations. Figure shows the total arrival delays in hours geographically.

Table 2: Rank-Size distribution of delays in 20 stations for both 2008 and 2009

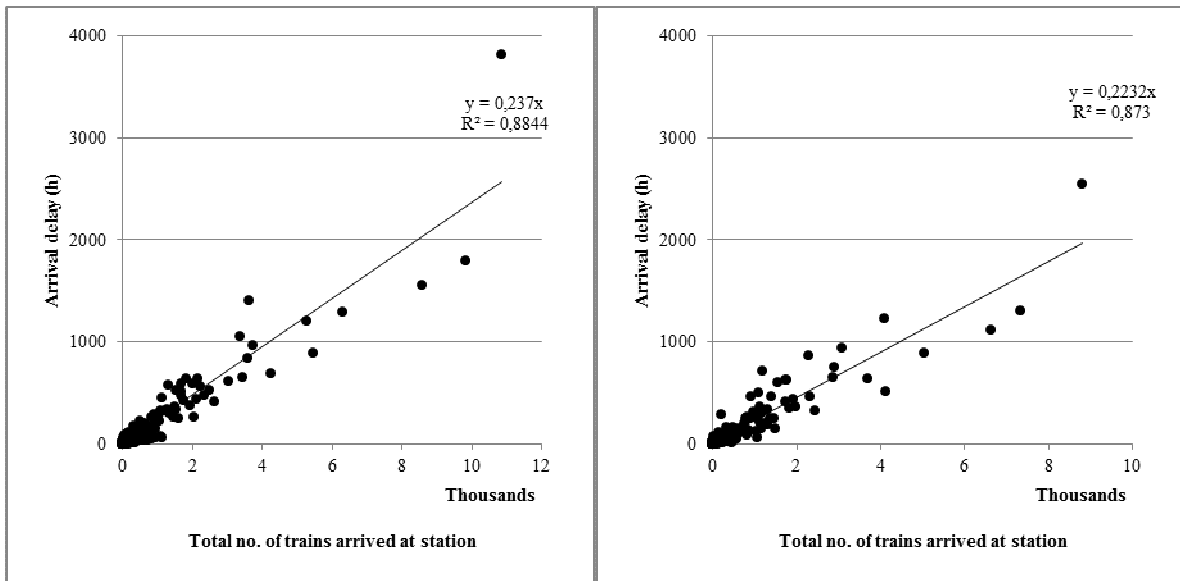
2008					
Station	Rank	Total arrival delay (hour)	% share of total delay stations	Cumulative	No. of O-D pairs
Hallsberg marshaling yard	1	3883	11%	11%	52
Borlänge C	2	1838	5%	16%	42
Malmö freight yard	3	1577	4%	20%	34
Kiruna ore yard	4	1424	4%	24%	6
Sävenäs marshaling yard	5	1317	4%	27%	35
2009					
Station	Rank	Total arrival delay (hour)	% share of total delay stations	Cumulative	No. of O-D pairs
Hallsberg marshaling yard	1	2596	9%	9%	59
Borlänge C	2	1335	5%	14%	43
Boden C	3	1229	4%	18%	29
Malmö freight yard	4	1133	4%	22%	29
Göteborg Skandia harbor	5	952	3%	25%	20

Figure 6: Total arrival delay per station in 2008 and 2009



We expect a high degree of correlation between the total number of freight trains arriving at a certain station and the total arrival delay in that station. In order to explore this issue we use scatter plots and regression analysis (see Figure 7). As expected, there is a linear relationship between the number of trains and total arrival delay that fits relatively well. However, total arrival delay in Hallsberg is higher than expected, given the number of trains arriving. One possible explanation is that passenger trains contribute to congestion in Hallsberg.

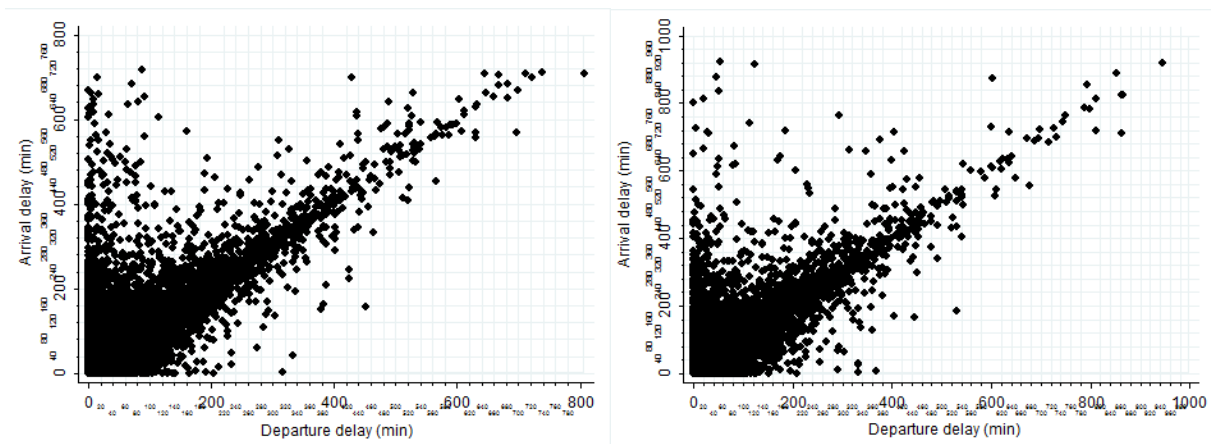
Figure 7: Scatter plot of no. of trains vs. arrival delay in 2008 (left) and 2009 (right)



4.2.2 Transmission of delays in the freight train network

In this section we explore how a delay is propagated within the network. In a highly used rail network close to capacity limits we would expect that delays increase along the trip to the final destination. The simplest way to examine this question is to estimate the impact of delays in the origin on the delay in the final destination. Figure shows the scatter plot for departure and arrival delays.

Figure 8: Departure delay vs. arrival delay in 2008 (left) and 2009 (right)



One difference between the years we can see by plotting the data is that in 2009 the number of trains with no or small departure delay but with high arrival delay is smaller than in 2008. The

observations of both years were merged together. In order to quantify a relationship between departure time from origin and relative arrival time at the destination OLS-estimation is used. The distance is included to control for that the probability for arrival delay is higher for longer distance because of the accumulation of disturbances (see Equation 4). Table 3 shows the regression results.

$$arrival_delay = \beta_0 + \beta_1 departure_delay + \beta_2 distance \quad (4)$$

Table 3: Linear regression model results

	Coefficient	Std. error	t-statistic
β_0	3.957	0.222	17.84
β_1	0.946	0.003	358.73
β_2	0.037	0.001	53.02
R-square: 0.6562			

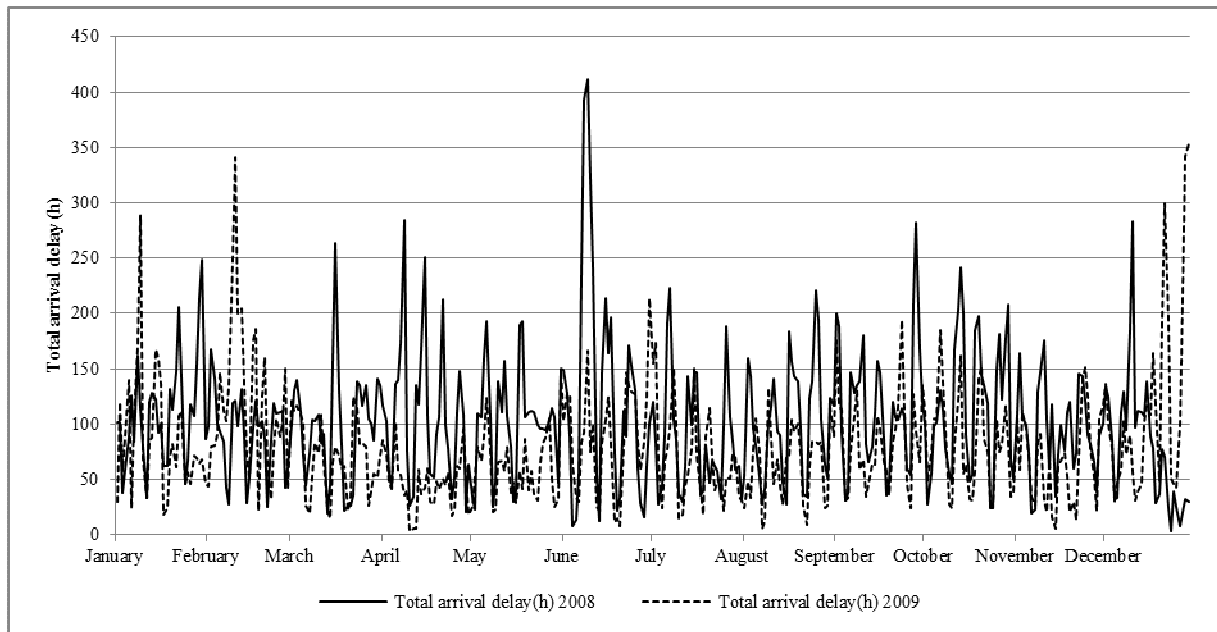
A positive intercept in the regression model shows the expected delay at the destination given that the departure delay is zero. If the arrival delay increases one minute for every minute the departure is late, the slope coefficient of the regression would be equal to one. However, the slope coefficient for departure delay is significantly smaller than one. This implies that some part of the initial delay is recovered along the trip. This is contrary to the hypothesis outlined above, according to that, we would have expected that train delays are subjected to a vicious cycle, so that once a train is delayed, the train will be delayed even more along the travel to the destination. One explanation is the high number of early arrivals in the data, implying that the timetable allows early departures due to a considerable number of available slots, suggesting that the network is not heavily utilized in certain parts of the Swedish rail network and during parts of the day (i.e. during the night when many freight trains go). Hence, it would be fruitful to analyze delay data for passenger trains (that have less slack) in order to reveal the true nature of delay propagation, but this is outside the scope of the present paper.

4.3 Temporal delay distribution

In this section we want to explore the distributions of delays considering daily variations. Figure 9 shows the result of this daily aggregation as a time series for 2008 and 2009. The most

pronounced differences are the large delays in December 2009, a combination of harsh winter conditions and high demand (many trains) due to the Christmas holiday.

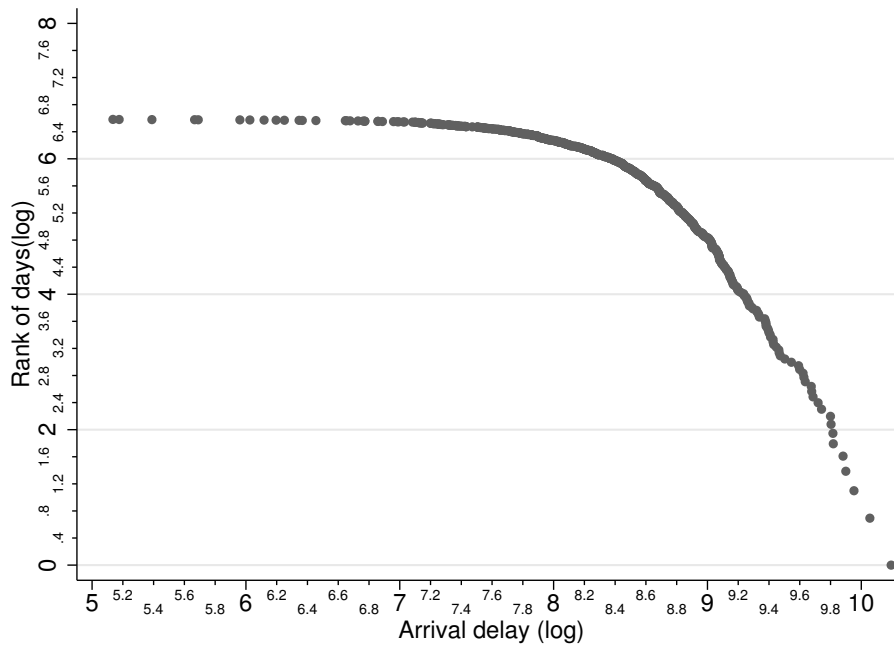
Figure 9: Total delay per day in the Swedish rail network



By superimposing the delays for 2008 and 2009 on each other, we can see that there is a high degree of periodicity in the data, showing the same pattern in both years. However, there are deviations from the periodicities. Among the different periodicities (within weeks, months, seasons and years), the weekly variations seem to be most pronounced.

We sum the delays for every day and use the same procedure described above for delays at stations, that is, we assign rank 1 to the day with most delay minutes and rank the other days accordingly. Figure 10 shows the rank-size distribution of all days in the data set on a log-log scale. Interestingly, the tail of the rank-size distribution seems to follow a power-law. This is a difference in comparison for delay distributions per train (size distribution) and per station (spatial distribution) both apparently following an exponential distribution. Hence, the worst days are very different from a normal day and this might indicate that there is a cascading spread of delays within the network. We can also calculate the probability for an event (certain total amount of delay any given day) to occur in the network. There is a small but positive probability that a daily disturbance double the size that occurred during 2008 and 2009 will occur during a certain future year. The probability is determined by the slope of the linear fit to the tail of the rank size distribution in Figure 10.

Figure 10: Rank size distribution of days



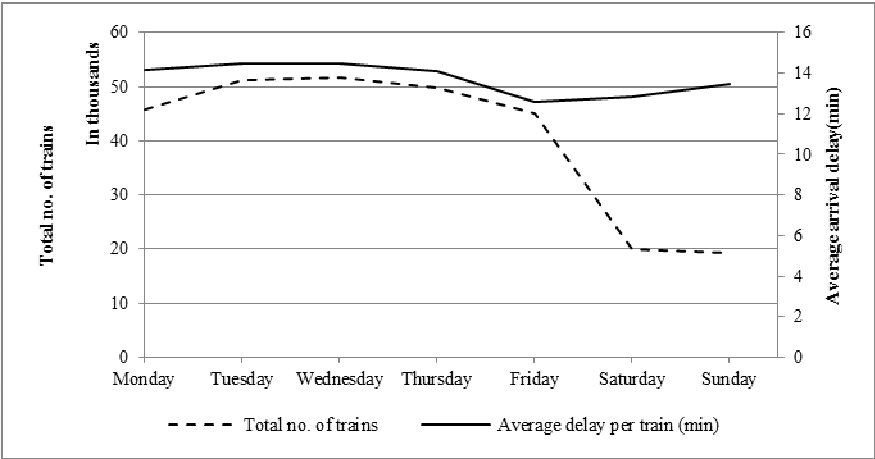
4.4 Capacity utilization and expected delay

In Section 4.2 we saw a strong but imperfect relationship between the amount of delay and the number of trains, that is, more trains means simply more delays. However, as capacity utilization increases due to the increased number of trains we would expect that average delay per train increases since congestion would lead to knock-on effects. Capacity utilization differs over time and rail demand exhibits periodic cycles within days, weeks and months. In fact, for the years examined here, the ton-kilometers transported on rail was 17% lower in 2009 compared to 2008 due to the economic contraction. In a sense, these exogenous variations allow us to examine the link between capacity utilization and the expected delay per train. All things equal, we would expect that the average delay increases as capacity utilization increases.

First, we examine the within week variation (see Figure 11). We expect to have more delays during workdays, since there would be more passenger trains commuting, which increases the chance of meeting that leads to more delay. In addition, there are also more companies sending and receiving goods during week days. Figure 11 shows total arrival delays at terminal stations regarding days of week and relatively total number of trains. It seems Tuesdays and Wednesdays are the busiest weekdays which result in higher total level of delays. As can be seen even if the number of freight trains decreases by more than half during weekends, the average delay is almost unaffected. The total number of passenger trains is constant across weekdays at circa 230 thousands whereas during weekends the number falls just below 150 thousands, which

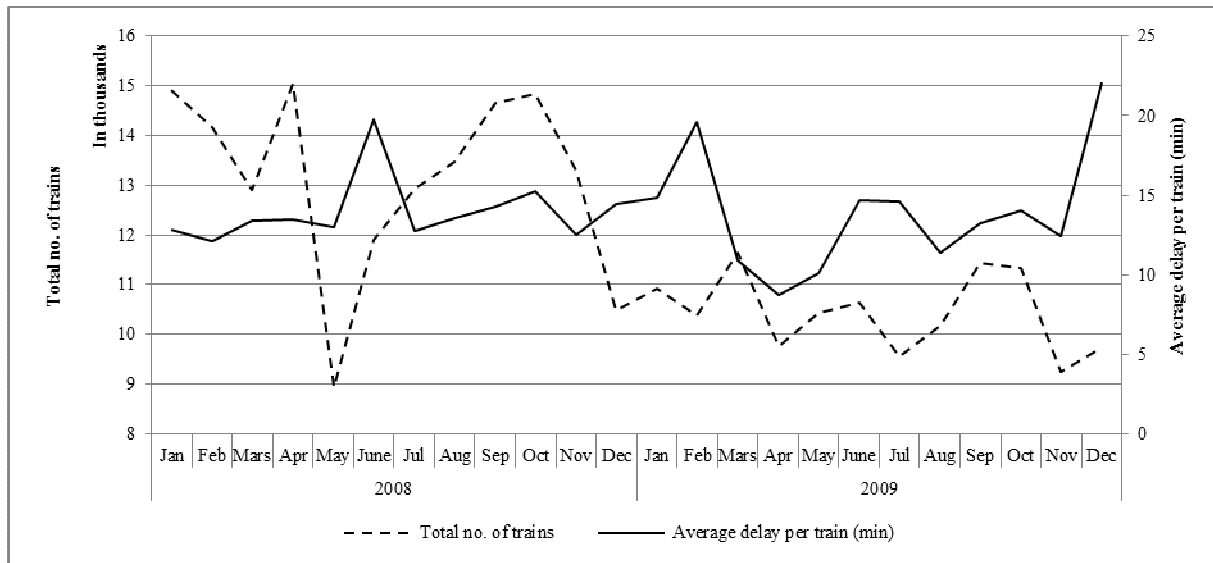
corresponds to a decreased capacity utilization of about 35 percent during weekends. Hence, we would expect a significant dip in average delays since capacity is freed up during weekends, but no such effect can be observed.

Figure 11: Total number of trains and average delay per weekday



Next we aggregate arrival delays over months so that the strong within-week periodicity is filtered from the data. In Figure 12 we aggregate the data for months and compare total number of trains and average delay per train per month. We can see from Figure 12 that the annual trend which is negative due to the economic bust of 2009, is more pronounced than the seasonal variation (for example, there is a peak during autumn in both years). Interestingly, even as the total number of freight trains is falling during 2009, there is no visible effect on the average delay. No clear-cut relationship between the number of trains and average delay emerges. The correlation coefficient is -0.07, implying that there is a slight tendency that delay per train is decreasing as capacity utilization is increasing.

Figure 12: Arrival delay and total number of trains aggregated over months



It could be argued that the weak (and actually negative) link between capacity utilization and average delay is due to the fact that the traffic is decreasing in off-peak times whereas it is the same during peak-times and hence that the missing link is just an illusion. In order to investigate this issue we identify peak-times by plotting the total number of trains using the rail network at any given hour (see Figure 13). Figure 13 shows that the only peak time is around 6am for both years. It can be seen that there is a reduction of freight trains in 2009 both during peak hours and off-peak hours. Figure 14 analyzes the impact of the decrease in freight transportation demand between 2008 and 2009 on average delay in 2009. In general, there is no evidence in Figure 14 that traffic demand reductions either during peak or during off-peak decreases average delay in 2009, the correlation coefficient is -0.31. Hence, the link between capacity utilization and expected delay per train is best described as weak, although extreme delays might bias upward the average delay estimates, hence the negative relationship.

We also analyze the relationship between total number of trains including passenger trains in order to control for potential interaction effects, that is, whether passenger trains affect average delay for freight trains. The correlation coefficient for all train arrivals during a certain time of the day and average delay per freight train is 0.09.

Figure 13: Total number of trains and time of day

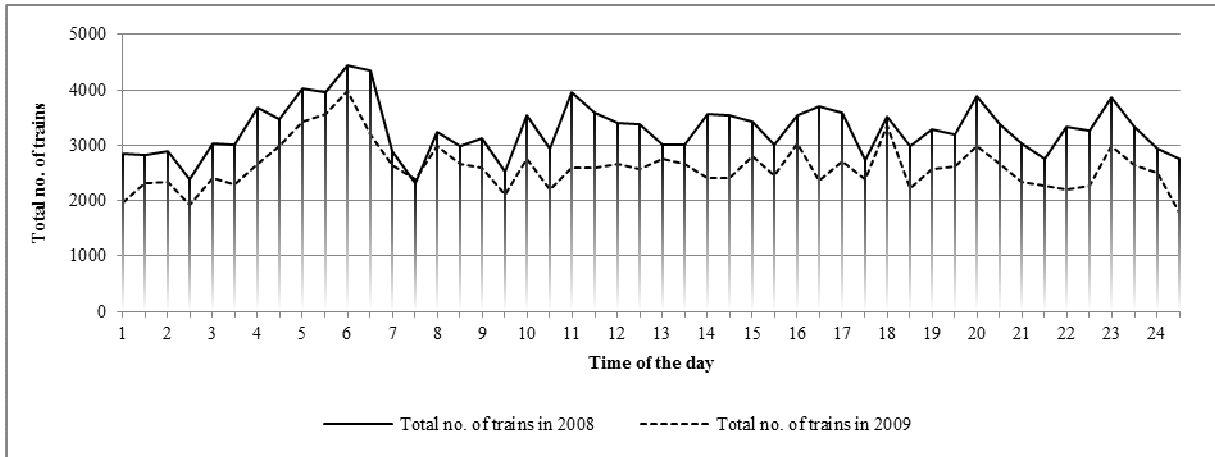
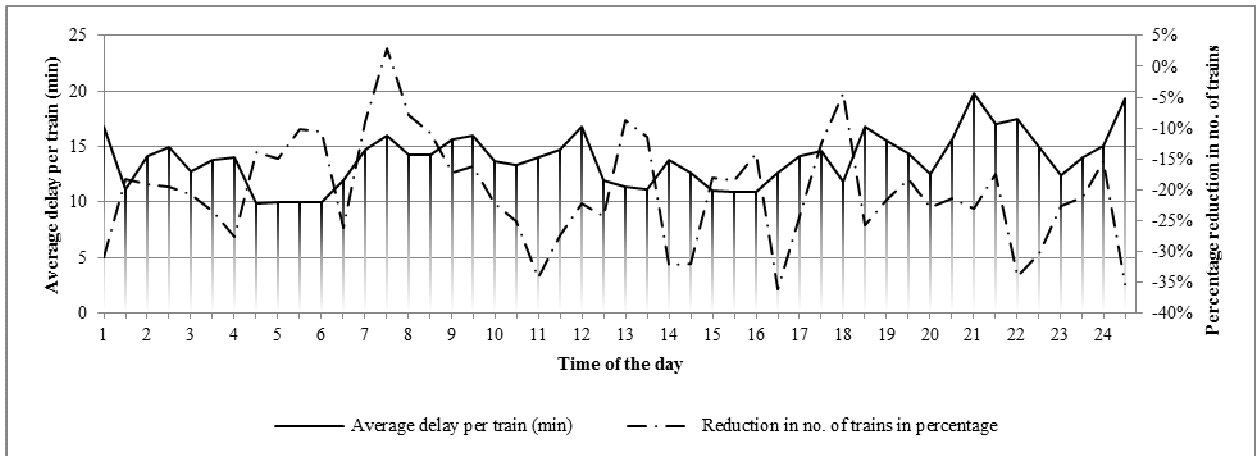


Figure 14: Average delay in 2009 and reduction in number of trains in percent



5 Implications

5.1 Implications for inventory management and the value of transportation time variability

The results have implications for how many parts companies have to keep in buffer stocks in order to provide a sufficient service level. The buffer stock size is also related to the value of transport time variability (VTTV), since in the absence of variability no buffer stock would be needed and hence no additional costs of variability would arise.

A representative company determines a required service level a , defined as the ratio of the number of on time deliveries and total deliveries. Based on this service level the company reacts to a stochastic delivery time by holding a buffer stock s , which is a function of standard deviation (or more general: uncertainty) in transport time: $s=f(\sigma)$ ceteris paribus. The cost for holding a buffer stock and hence the cost of variability in transport times is the cost of physical storage of the goods and the capital costs of the goods stored. Hence we can compute the cost of variability under certain simplifying conditions.

More specific, we illustrate with the following example: if we consider a monthly production of 4400 units it implies that the monthly need for a certain input is 4400 as well (assuming here that the ratio is one input unit per one unit output). If production consists of 22 working days a month it implies that there is a daily need for 200 parts. If the production is based on Just-in-time deliveries the required service level would be about 99%, that is, 99% of all deliveries should be on time in order to avoid stockout cost. It can be shown (Hansmann, 2006), that if transport time is normally distributed, this means that the buffer stock required is (e is the value of a given percentile for the standardized normal distribution and q is the demand in production per time unit):

$$s = e_{99} \cdot \sigma \cdot q \quad (5)$$

Note that σ and q have to be expressed in the same time unit; hence if the standard deviation is expressed in hours, we need to compute the input demand per hour as well. Between Malmö and Hallsberg we can compute from the data that the average planned transport time for rail freight is 6 hours and that the standard deviation is 1 hour. Assuming that the company faces this average transport time and variability the buffer stock becomes (we need to order 100 parts at each time if the company produces during 12 hours per day, that is 200/12 parts per hour):

$$s = 2.3263 \cdot 1 \cdot 200 / 12 = 39 \quad (6)$$

If the standard deviation is 1 hour, in 99 % of all cases the deliveries will be within 8.23 hours, that is, at most 2.32 hours delayed. The buffer stock is thus what is needed in production during the 2.32 hours of maximum delay, roughly 39 parts (since we need 200/12 parts per hour as input in production). Thus, in total we have to order 139 parts each time, 100 parts for production during the expected transport time and 39 as a buffer for the stochastic part of transport time.

The finding in this paper is however that rail transports are not normally distributed. This affects the calculation of buffer stocks twofold: once, the standard deviation itself might be biased downwards and second, the standard deviation is not a good risk measure for variability. The empirical delay distribution shows that the 99th percentile between Malmö and Hallsberg is more than 2.32 times the standard deviation of 1 hour (that is 2 hours and 20 minutes); it is instead almost 4 hours in total. Hence, the necessary buffer stock becomes 67 parts (demand in production per hour times 4 hours of maximum delay) instead of 39 parts, an increase by 70 percent compared to the calculation based on standard deviation and the assumption of normality.

Thus, there are huge amounts to be saved by reducing the extreme delays. However, because of the possible complex causal-structure behind large delays we do not know if it will be effective to target them. Still, the buffer stock example here illustrates the importance of the tail for freight transports. It can be argued that we will seldom see rail-based just-in-time production and that most transport chains include collection and distribution transports by road. However, this is an adaption to the present state of the rail network by companies. Keeping buffer stocks, localisation of supply industry close to factory (agglomeration buyer-supplier) and transporting freight by other modes are possible strategies to cope with rail transport time variability. However, all these measures tentatively increase costs for companies. Using standard deviation and normality of delays would be misleading for calculation of this cost.

5.2 General transport policy implications

We believe that this paper has important, more general policy implications. The results suggest that the general association between rail breakdowns and weather conditions might be just a revelation of symptoms and not the sole cause of such breakdowns. For example, at many times and places bad weather occur without train operations failing. Hence, we think that even if exogenous factors matter for the probability and severance of breakdowns, the causes are to some extent endogenous to the rail transport system and most probably there is some interdependence and combinations between various endogenous and exogenous causes.

Endogenous in this context means that the rail transport system has a complex structure that is prone to be robust with respect to disturbances at most times and places but that at some times and places the network is fragile to disturbances, propagating initial small and geographically limited disturbances through connecting links, nodes and the rolling stock itself to larger and more widespread breakdowns in the transport network. This is the reason why we consider them as endogenous: the network itself causes the breakdowns because of its own structure. One example is the track capacity utilization that is assumed to have a pronounced effect on freight train delays. But one can argue that large delays occur where capacity is used the most, because there are many links, nodes and train that can be affected and once affected influence other links, nodes and trains. In other words, the extremely unevenly capacity utilization that is characteristic to rail networks (and other complex networks) is in itself a cause for large delays. Once we account for the number of trains the average delay per train (and per train kilometre) is not significantly affected. Hence, large delays are just a sign that we increased transport demand and there will be no easy way to prevent delays. But there are nevertheless some possible policy measures that can be used to target the problem. The reason is the revealed extremely uneven distribution of transport demand and delays over time and space. These distributions suggest that investment in capacity, reinvestments and ordinary maintenance measures should be spread not uniformly across the rail network but be focused on problematic hotspots. Even now investments and maintenance are not carried out evenly but capacity investments and reinvestments should probably be even more extremely distributed.

Often, but not always, we can identify a certain cause for how a major break-down started.¹ We have to acknowledge that many delays are unpredictable. Our data analysis reveals that the delays that matter most have extremely low probability. Even if we know based on our analysis that during the next year a major breakdown of a certain size will occur at least one time, we still do not know where and when. If we could predict them they would not occur. However, based on our analysis of distributions delay distributions at stations we can say that measures undertaken to reduce the importance of central nodes (regardless whether it is a steering computer or a marshalling yard) would milder the consequences of disturbances. For example, backup capacity at steering and information centrals that can overtake responsibility temporarily. Similarly, measures undertaken for improving reliability outside the central nodes are with a high

¹ An example: On 23rd July 2011 the city of Norrköping south of Stockholm received 43 mm of rain within just two hours. Due to this heavy rainfall the computer central of the Swedish Transport Administration, responsible for steering and providing information for rail traffic in large parts of Southern Sweden located in Norrköping, were flooded. As a result, train operations in major parts of Sweden between Malmö and Stockholm had to be cancelled and redirected (with large delays) for several days.

probability without any major effect on overall performance of the rail network since they are of minor importance of the functioning of the whole network. For this reason, our results provide some (even if limited) insights on the impact of measures on performance outcomes and are therefore of some value for cost-benefit analysis of measures targeted at improving reliability and increasing capacity.

6 Conclusions

This paper analyzes the distribution of freight train arrival delays on the spatial, temporal and frequency-size scale. Since the spatial and frequency-size and temporal distributions describe the vulnerability of a rail transport system it has potentially important policy implications. Considering the tail of arrival delays at the final destinations within the rail system we find that it is exponentially distributed. This implies that the tail makes up the biggest part of total delay time. The 20 % largest delays contribute to about 74% of total delay minutes. For the spatial scale, we find that more than 50% of the total arrival delay per year occurs in just 7% of stations, all of those being marshaling yards. The contributions of the paper to the literature are manifold, mainly because this paper: i) analyzes all freight train delays in a national rail network covering 2 full years, ii) analyzes the spatial and temporal concentration of delays, iii) uses the exogenous shock caused by the financial crisis to investigate the relationship between capacity utilization and average delay and iv) analyzes the implications of observed delay distribution on the inventory management of firms which has, as we argue, implications for the valuation of transportation time variability as well.

With the help of regression analysis we analyze how delays are propagated in the network. We find that delays at the origin increase arrival delay but that some part of the initial delay is gained at arrival, probably due to large slack in the timetables. Finally, in the temporal scale we analyze arrival delays in different time scales such as monthly, weekly and daily delays. We expected that the significant reduction of ton-kilometers transported due to the economic recession would reduce not only total but also the average arrival delay since there would be more free capacity. The results show, however, that the average delay did not decrease as the number of freight trains decreased due to the economic contraction in 2009.

To the best of our knowledge this is the first study in its kind related to rail freight. Based on our findings we would advocate further research in this area with focus on passenger trains, the interaction between passenger and freight trains and further analysis of Swedish rail network structure. The analysis of delay distribution facilitates the selection of appropriate methods for valuation of rail reliability and thus for cost-benefit analysis of rail investments and other measurements. More specific, our analysis suggests that standard deviation is not an appropriate measure of transport time variability of rail freights. Moreover, it seems that very few extremely large delays matter most for the total amount of delay and not the many small delays. However, only a sound valuation method can answer the question whether society would benefit from reducing the number of delays in general or from preventing the extremely large delays.

References

- Bruinsma, F., Rietveld, P. & Vuuren, D. v., 1999. Unreliability in Public Transport Chains. World transport research - selected proceedings of the 8th world conference on transport research, Volume 1, pp. 59-72.
- Burdett, R. L. & Kozan, E., 2006. Techniques for absolute capacity determination in railways. *Transportation Research Part B: Methodological*, 40(8), pp. 616-632.
- Carey, M. & Carville, S., 2000. Testing schedule performance and reliability for train stations. *Journal of the Operational Research Society* 51, pp. 666-682.
- Carey, M. & Kwiecinski, A., 1994. Stochastic approximation to the effects of headways on knock-on delays of trains. *Transportation Research Part B*, 28B(4), p. 251–267.
- Chen, B. & Harker, P. T., 1990. Two moments estimation of the delay on single-track rail line with scheduled traffic. *Transportation Science*, Volume 24, pp. 261-275.
- Clauset, A., Rohilla Shalizi, C. & Newman, M. E. J., 2009. Power-law distributions in empirical data. *SIAM Rev.*, 51(4), pp. 661-703.
- Corman, F., D'Ariano, A., Longo, G. & Medeossi, G., 2010. Robustness and delay reduction of advanced train dispatching solutions under disturbances. Lisbon, Proceedings of the 12th World Conference on Transport Research, WCTR 2010, pp. 1-12.
- Corman, F., D'Ariano, A., Pacciarelli, D. & Pranzo, M., 2011. A bilevel rescheduling framework for optimal inter-area train coordination. Saarbrücken, ATMOS 2011 - 11th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems, pp. 15-26.
- D'Ariano, A. & Pranzo, M., 2009. An Advanced Real-Time Train Dispatching System for Minimizing the Propagation of Delays in a Dispatching Area Under Severe Disturbances. *Networks and Spatial Economics*, 01 03, 9(1), pp. 63-84.
- D'Ariano, A., Corman, F., Pacciarelli, D. & Pranzo, M., 2008. Reordering and local rerouting strategies to manage train traffic in real time. *Transportation Science*, 42(4), pp. 405-419.
- Frank, O., 1966. Two-Way Traffic on a Single Line of Railway. *Operations Research*, Volume 14, pp. 801-811.

- Goverde, R., 2005. Punctuality of Railway Operations and Timetable Stability Analysis, s.l.: TRAIL Research School for Transport, Infrastructure and Logistics, Delft University of Technology.
- Goverde, R., Hooghiemstra, G. & Lopuhaä, H., 2001. Statistical Analysis of Train Traffic. Delf: Delft University Press.
- Greenberg, B., Leachman, R. & Wolff, R., 1988. Predicting dispatching delays on a low speed single track railroad. *Transportation Science*, 22(1), pp. 31-38.
- Güttler, S., 2006. Statistical modelling of Railway Data, s.l.: M.Sc. thesis, Georg-August-Universität zu Göttingen.
- Hansmann, K.-W. (2006), *Industrielles Management*, 8th ed, Oldenbourg.
- Haris, M., 2006, *Analysis and Modeling of train delay data*, s.l.: University of York.
- Harker, P. & Hong, S., 1990. Two moments estimation of the delay on a partially double-track rail line with scheduled traffic. *Transportation Research Forum*, Volume 30, p. 38–49.
- Higgins, A. & Kozan, E., 1998. Modeling train delays in urban networks. *Transportation Science*, 32(4), p. 251–356.
- Lindfeldt, A., 2010. A study of the performance and utilization of the Swedish railway network. Opatija, s.n.
- Lindfeldt, O., 2008. Evaluation of punctuality on a heavily utilised railway line with mixed traffic. *Computers in Railways XI*, WIT Press, pp. 545-553.
- Lindfeldt, O., 2010. *Railway operation analysis: Evaluation of quality, infrastructure and timetable on single and double-track lines with analytical models and simulation*, Stockholm: KTH.
- Marinov, M. & Viegas, J., 2011. A mesoscopic simulation modelling methodology for analyzing and evaluating freight train operations in a rail network.. *Simulation Modelling Practice and Theory*, Volume 19, pp. 516-539.
- Mu, S. & Dessouky, M., 2011. Scheduling freight trains traveling on complex networks. *Transportation Research Part B: Methodological*, 45(7), pp. 1103-1123.

- Murali, P., Dessouky, M., Ordóñez, F. & Palmer, K., 2010. A delay estimation technique for single and double-track railroads. *Transportation Research Part E: Logistics and Transportation Review*, July, 46(4), pp. 483-495.
- Nyström, B., 2006. Delay analysis of a freight train – an improvement case study, Luleå: Division of Operation and Maintenance Engineering, Luleå University of Technology.
- Nyström, B., 2007. Consistency of Railroad Delay Attribution in Sweden- Measurement and Analysis. Division of Operation and Maintenance Engineering, Luleå University of Technology.
- Nyström, B., 2008. Aspects of improving punctuality from data to decision in railway maintenance, Luleå: Luleå University of Technology.
- Petersen, E., 1974. Over the road transit time for a single track railway. *Transportation Science*, Volume 8, pp. 65-74.
- Radtke, A., 2005. Software tools to model the railway operation, s.l.: Habilitation at the University of Hannover.
- Schwanhäußer, W., 1974. Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn, Aachen: Veröffentlichungen des verkehrswissenschaftlichen Instituts RWTH.
- Schwanhäußer, W., 1994. The status of the German railway operations management in research and practice. *Transportation Research Part A: Policy and Practice*, 28(6), pp. Pages 495-500.
- Törnquist, J., 2006. Railway traffic disturbance management, Karlskrona: Blekinge Institute of Technology.
- Yuan, J. & Hansen, I., 2007. Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B*, Volume 41, p. 202–217.
- Yuan, J., 2006. Stochastic Modelling of Train Delays and Delay Propagation in Stations, Delft : Delft University.
- Yuan, J., Goverde, R. & Hansen, I., 2002. Propagation of train delays in stations. *Computers in Railways VIII*, WIT Press, pp. 975-984.