TOWARD IN-DEPTH ANALYSIS OF TRAIN PUNCTUALITY DATA

Jouni Paavilainen. Department of Business Information Management and Logistics, Tampere University of Technology, Finland. Email: Jouni.paavilainen@tut.fi

Riikka Salkonen. Department of Business Information Management and Logistics, Tampere University of Technology, Finland. Email: riikka.salkonen@tut.fi

ABSTRACT

Punctuality (on-time performance, schedule adherence) is one of the most important success factors for a railway traffic system. Despite the importance of punctuality, there seems to be a lack of broad understanding when it comes to the formation of punctuality. Within a railway traffic system, delays concatenate easily, that is, a single delay is likely to cause many other delays, so-called secondary delays. To date, most of the studies related to delays have examined only stations and station-like systems. However, especially within a single-tracked infrastructure—as in Finland—a notable portion of secondary delays takes place outside stations. Thus, the examination of delay concatenation should be done by considering the whole network.

This paper first describes the current practices and challenges related to the analysis of train punctuality data for Finnish railways. Railway organizations in Finland record a lot of punctuality-related data, but with the current data and methods, one is able to allocate primary delays only to their causes. Secondary delays are not analyzed. Considering this, it is also impossible to identify which of the primary delays are the most critical ones. Hence, the research question of this paper is: “How can we analyze train punctuality data more efficiently and systematically in order to gain an understanding about the most critical delay concatenation phenomena?”

With actual motion data from Finnish railways, we prove that at least the most explicit delay chains can be identified and mapped, and this information can be used, for example, for allocating delays to their real causes, and developing a more robust timetable. We thus argue that this kind of examination can be made and, more importantly, should be made. We also suggest a data mining method called sequence analysis to automate this process. Sequence analysis, with the other data mining techniques, would provide a significant improvement over traditional statistical techniques when analyzing the train punctuality data.

Keywords: railway traffic system, punctuality, delay propagation, primary delays, secondary delays, statistical methods, data mining, performance measurement

1. INTRODUCTION

This paper describes the current practices and challenges related to the analysis of train punctuality data in Finnish railways, and uses this basis to propose a more in-depth analysis process of the data.

Punctuality (on-time performance) is one of the most important success factors for a railway traffic system (Olsson, Haugland 2004). Unpunctuality reduces the capacity of the system and dilutes the overall service quality, negatively affecting the performance and thus the competitiveness of the whole mode of transport (see, e.g., Albrecht et al. 2008). Despite the importance of punctuality, there seems to be a lack of broad understanding when it comes to the formation of punctuality. The general challenge is that a railway traffic network forms a complex system, making it difficult to gain an understanding of all the important phenomena involved in the system.

Within a railway traffic system, delays concatenate easily, that is, a single delay is likely to cause many other delays, so-called secondary delays. For example, a train arriving or departing late at a station may delay the arrival or departure of other trains (Carey, Carville 2000) as the train still occupies the scheduled route, preventing other trains from passing (crossing) (Yuan, Hansen 2007). A secondary delay can also take place when a train is waiting for transfer passengers from a delayed train. Secondary delays dilute remarkably the overall punctuality of the system, especially in countries with single-tracked infrastructure—like in Finland. Altogether, according to our dataset described later in section 2, secondary causes are responsible for more than 50 percent of the total delay minutes in Finland.

By far only little attention has been given to secondary delays when it comes to the whole traffic network: many of the studies have concentrated on restricted systems, like stations. However, a railway network forms a system, in which everything affects everything else. Hence, it would be important to study how the system works as an ensemble: how secondary delays emerge due to primary delays that could have happened even at the other end of the network. This is especially important when it comes to single-tracked systems, but also a fully double-tracked system forms an ensemble where a perturbation at one station might affect traffic at other stations. Hence, the results and observations of this study can be generalized to be applicable in many other systems.

But how can we gain understanding about the formation of delay concatenation? Nowadays, railway organizations in Finland record a lot of punctuality-related data but do not have the means to analyze and process it into relevant information. This is because neither the structure of the data or the analysis tools used have not supported this kind of examination. With the current data and methods, one is able to allocate primary delays to their causes. Thus, the organizations can, for example, point out the most common primary causes of delays, and point out the most critical track sections, train types, equipment, and so on, to that end. Instead, railway organizations do not have the possibility to define what the total

consequences of primary delays are, that is, what secondary delays emerge because of primary delays. Considering this, it is also impossible to identify which of the primary delays are the most critical ones, causing the most negative effects to the system. Under these circumstances, our hypothesis is that the railway organizations are not able to direct their development activities to the problems that cause the biggest negative effects to the overall system.

In light of the described challenge, we formulate the research question of this paper as follows: "How can we analyze train punctuality data more efficiently and systematically than is done today, in order to gain an understanding about the most critical delay concatenation phenomena?" By answering this question, the study achieves both scientific and pragmatic significance, for this study takes the first step to examine secondary delays within the whole railway traffic system, and suggests a useful process for end-users for analyzing the actual data. With the understanding gained, railway organizations will be able to make more informed decisions when it comes to developing the system.

In section 2, we review the significant literature on railway traffic punctuality. In section 3, we describe our research method and compare it with other methods, as well as introduce our empirical material. After exploring the actual data from Finnish railways (section 4) and presenting the results of this process (section 5), we discuss the potential of a selected data mining method for future analysis of the data in section 6. Conclusion and our propositions for the future research are presented in section 7.

2. LITERATURE REVIEW

The English language literature on railway operations is very large. In this section, we aim to highlight studies that are the most significant when it comes to punctuality and the delay concatenation, and analysis methods of them.

According to Olsson and Haugland (2004), **punctuality** refers to deviations, usually negative, from the timetable: "if a train runs within the accepted deviation from the timetable, it is punctual, otherwise it is not." Thus, punctuality can be defined as the ability to meet the predefined timetable. **Delay** is "a continuous measurement of the deviation from the timetable" (Olsson, Haugland 2004). If a train arriving at or departing from a selected point—e.g., a station—is delayed more than the predefined level, the train is considered to be unpunctual. **Primary delays** refer to delays that are not caused by other trains, but, e.g., technical failures of rolling stock or infrastructure, slower-than-scheduled running speed, prolonged alighting and boarding times of passengers, or bad weather conditions (Yuan, Hansen 2007). In addition, terms such as exogenous delay, original delay, and initial delay are used alternatively and complementary with the term primary delay (see, e.g., Olsson, Haugland 2004, Albrecht et al. 2008, Carey 1999, Carey, Carville 2000, Vromans, Dekker & Kroon 2006, Mattsson 2007). If a delay is due to other trains, the term **secondary delay** is used (Albrecht et al. 2008). The following terms are also used in parallel to secondary delay: consecutive delay, knock-on delay, and reactionary delay (see, e.g., Olsson, Haugland 2004,

Albrecht et al. 2008, Yuan, Hansen 2007, Carey 1999, Vromans, Dekker & Kroon 2006, Gibson, Cooper & Ball 2002). In this paper, we concentrate on these secondary delays.

The **concatenation of delays**—that is, the emergence of secondary delays—has been subject to considerable research at the Technical University of Delft in the Netherlands in recent years (see e.g., 2002, 2004, 2006, and 2007). This research has been mostly conducted by Yuan and Hansen. According to one of their articles (2007), the concatenation of train delays depends significantly on the infrastructure layout and timetable design. Among the systems they have studied, most of the secondary delays seem to emerge at stations, since the crossing and merging of lines and platform tracks that are located at the stations' areas are, in most cases, the bottlenecks in highly used railway networks (Yuan, Hansen 2007). Thus, they have naturally concentrated on stations and station-like systems in their research. That is also the case in many other studies related to delay concatenation, for example the vast research of Carey et al. (see e.g., 2000, 2003, and 2007).

The systems the aforementioned scholars have studied are mainly double-tracked. That, however, is not the case in all countries. For example in Finland, the railway network consists mainly of single lines with passing places along the way. Within these kinds of systems, a large number of secondary delays take place along the route, far away from stations. Consequently, in Finland and similar countries, secondary delays are an even bigger problem than in countries with double-tracked infrastructure.

Also single-track systems have been subject to a considerable research. Lindfeldt (2007) has studied the Swedish railway network, which is mainly single-tracked. He states that single-track systems are very sensitive to disturbances, that is, to suffer from secondary delays. Thus, he proposes an analytical model for analyzing critical crossing situations. Higgings et al. (1997) on their behalf have proposed a model to determine the required number and position of sidings on a single track rail. Their model minimises both the risk of delays and the delays caused by train conflicts.

Castillo et al. (2009) concentrate on the timetabling problem of a single-track railway line. They combine several criteria to optimize timetables within relatively restricted systems. They demonstrate that sensitivity analyses are very useful in identifying critical trains, segments, stations, etc. Also Zhou & Zhong (2007) study the single-track timetabling problem. They propose a mathematical solution to obtain optimized and still feasible schedules. Besides mathematical modelling, also simulation methods are used for examining single-track systems. For example Dingler et al. (2009) have simulated the North American mainly single-tracked system, in order to analyze the effects of traffic heterogeneity to delays and capacity. They claim that this understanding enables more effective planning and efficient rail operations.

In general, analysis methods of railway traffic disturbances are usually divided into three different categories, which are analytical methods, simulation, and statistical methods based on empirical data (see, e.g., Carey 1999, Mattsson 2007). According to Mattson (2007), each method has several good qualities as well as weaknesses, and the method chosen depends

on the purpose of the study and on the definitions that are used in it. The following kinds of tasks are conducted by using one or more of these methods:

- determine how the capacity utilization affects delays or the risk of unreliability,
- analyze the propagation of delays within a specified infrastructure layout,
- map the probabilities and distributions of specified events.

Mattsson (2007) compared **analytical methods** with simulation methods and notes that analytic methods are usually much faster to apply. This feature makes analytical methods useful for strategic planning and design stage. Carey (1999) also used analytical methods in his study, which examines the system behind various delays within a double-track infrastructure. Analytical methods are also used in various other studies (see, e.g., de Kort, Heidergott & Ayhan 2003, Huisman, Boucherie 2001). The analytical methods most frequently used are related to the queue systems and Max-plus algebra (Vromans 2005). But as mentioned, analytical methods also have disadvantages. Mattsson (2007) pointed out that analytical methods lead easily to simplifications and cannot be used in complex problems. A real railway system is a very complex system, and it has many variables, which cannot be taken into account in analytical models.

Mattsson (2007) represented **simulation** as the most common and widely used way to do accurate modeling. In particular, micro-simulation is used in railway traffic system modeling. Simulations have the advantage of accuracy and the ability to solve complex processes. In railways, this means that the traffic situation can be simulated realistically. On the other hand, simulation tools can be laborious to construct and time-consuming. The more realistic results are needed, the more parameters the model needs. The simulation can be used to create the disturbance at a certain area and to study how the delays propagate on the network, causing secondary delays (Vromans 2005).

Statistical methods have been the third method for analyzing the railway system. Due to the weaknesses of other methods, Mattsson (2007) suggested that statistical methods offer the only realistic alternative to model the most unpredictable phenomena, e.g., the emergence of primary delays. Among the studies that apply an empirical statistical method, one rises above the others: Gibson's (2002) groundbreaking research described the cost of secondary delays through the analysis of primary delays. In section 4, we examine our dataset statistically to get an overview of how the system behaves, presenting the results in section 5.

Because of the weaknesses of aforementioned methods, the majority of the studies—also those presented in this section—related to the delay concatenation has concentrated on limited parts of the overall system, or simplifies the system, in order to get it modeled and analyzed. By far only little attention has been given to the whole traffic network, including all the nodes (stations) and arcs (rails between the stations). However, a railway network forms a system, in which everything affects everything else. Thus, it would be important to study how the system works as an ensemble: how secondary delays emerge due to primary delays that could have happened even at the other end of the network. Because of this, in section 6, we propose the use of a selected data mining method for analyzing the whole traffic network.

3. METHOD AND MATERIAL

Data mining

Data mining is defined as "the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database" (Giudici 2003). That is to say, the idea of data mining is that, instead of testing existing hypotheses with the data, one lets the data reveal what is going on in the system (Kudyba 2001). Thus, in data mining, no prior knowledge is needed (Giudici 2003), which does not restrict the user to dealing with assumptions or preconceptions that might not be known beforehand in the first place. Consequently, data mining has a clear advantage over any traditional statistical technique in the cases when the system is complex and the underlying patterns are not trivial. A railway traffic system indeed is this kind of system. Thus, data mining would seem to be a useful way to examine train punctuality data for the following reasons:

- The amount of data is seemingly large (tens of thousands of observations a day),
- considering the number of different trains and possible correlations among them, the data structure is rather complex,
- many of the underlying patterns in such a large data mass are unknown at first.

According to Kudyba (2001), data mining can offer answers for the following kinds of questions:

- Are there statistically dependent variables in the dataset?
- How strong are the abovementioned relationships?
- How will an alteration of a selected variable affect other variables?
- How selected variables will behave in the future?

Exactly these kinds of questions must be asked and answered when one is gaining an understanding of the concatenation of delays: it is all about variables—that is, trains—affecting other variables (trains). Hence, data mining in some form seems to be a respectable approach for analyzing train punctuality data.

Data mining is an abstract umbrella term for the activity described above. The actual analysis process could be done with different alternative and complementary methods, among which are, for example, clustering, segmentation and classification, regression, neural networks, association analysis, and visualization. In section 5, we examine the potential of the sequence analysis method, which is an extension of the popular association analysis method (Kudyba 2001).

The use of data mining methods is not widely studied when it comes to train punctuality data. Peters et al. (2005) have concluded a study that applies neural networks to predict future (secondary) delays, based on the existing (primary) delays. The neural network is a data mining method, whose idea is to learn from the data in order to make predictions about the

future (Kudyba 2001). The authors demonstrate that the neural network is able to learn from current delay constellations and can handle much more complex patterns than the traditional rule-based method, though any practical application is in need of a huge amount of test data (Peters et al. 2005). The focus of the study is on predicting near-future events to support the decision-making of train dispatchers. Hence, the study does not unfold how to map, analyze, and prioritize the current challenges of the system based on long-term data, in order to develop the system.

Vukadinovic et al. (1996) also applied the neural network approach, aimed at mitigating traffic disturbances. The researchers demonstrated that a neural network can learn from train dispatchers' decisions and then simulate the dispatching process. Consequently, it would be possible to create a system that can help the train dispatcher in his or her work. This study concentrated on real-time dispatching. Hence, when it comes to the concatenation of delays, the possibilities of other data mining methods, especially in the analysis of long-term data, should be studied. This paper lays the groundwork for this examination.

Description of the material

Train motion data

In Finnish railways, the motion data of trains is recorded by the system named JUSE. The system is based on control points, which in practice are situated at all stations. The motion data is automatically recorded right before a train's arrival and right after the departure. The accuracy of the measurements is one minute. This motion data is compared to the timetable of the train in the case, and the delay in minutes is logged into JUSE (zero is also recorded, in the case of no delay). If the delay is more than 5 minutes, the train dispatcher on the shift must enter a cause for the delay by using one of the 60 delay codes. In addition, if the train suffers from any extra delay after the first delay, the cause code must be entered (SysOpen Oyj 2003). Since the motion data is recorded before and after the control point—that is, a station, one is able to conclude whether the delay has happened at the station or between two stations.

Table 1 illustrates the form of data in JUSE by presenting the imaginary motion of a train from the city of Helsinki to the city of Tampere, with five intermediate stations.

Table 1. Imaginary motion data of a train from the city of Helsinki to the city of Tampere

	Absolute delay in minutes	Cause for the extra delay	Info	
Helsinki.Origin	**0**			No delay occurs
|				
Pasila.Arr.	**0**			No delay occurs
Pasila.Dep.	**5**	**M1** Passenger congestion		The train suffers a delay of 5 minutes, which is allocated to the cause code M1
|				
Tikkurila.Arr.	**5**			No extra delay occurs; hence, no cause code is logged
Tikkurila.Dep.	**5**			No extra delay occurs; hence, no cause code is logged
|				
Riihimäki.Arr.	**5**			No extra delay occurs; hence, no cause code is logged
Riihimäki.Dep.	**20**	**L1** Waiting for connection transport	**#328**	An extra delay, 15 minutes, is allocated to the cause code L1 (the actual cause train is also specified)
|				
Hämeenlinna.Arr.	**20**			No extra delay occurs; hence, no cause code is logged
Hämeenlinna.Dep.	**20**			No extra delay occurs; hence, no cause code is logged
|				
Toijala.Arr.	**22**	**I1** Bad weather		A 2-minute delay is allocated to the cause code I1
Toijala. Dep.	**22**			No extra delay occurs; hence, no cause code is logged
|				
Tampere.Dest.	**20**			The train gains 2 minutes on the schedule; no cause code is logged

In our example, the train leaves Helsinki with no delay. Similarly, the train arrives at Pasila on schedule. However, the departure from Pasila is recorded as being delayed 5 minutes. The system requires that the train dispatcher enter a cause for the delay. He or she judges that the problem was passenger congestion at the station. After that, the train makes its way to Riihimäki with no delay. However, at Riihimäki, the train has to wait transfer passengers from a connection train, which in turn happens to be delayed. The train leaves Riihimäki 20 minutes behind schedule, and the train dispatcher is required to enter the cause for the delay. Fifteen minutes is allocated to the L1 code, which refers to the wait for the connection transport. With this code, also the actual cause, train #328 is specified. After that, the train suffers still another delay of 2 minutes between Hämeenlinna and Toijala, which is allocated to I1 code (bad weather). Between Toijala and Tampere, the train gains 2 minutes on the schedule, whereupon the total delay at the destination station, Tampere, is 20 minutes.

In this study, our actual dataset has following characteristics:

- The time period is from 1st of September to 31st of September.
- All long-distance trains are included, whereas commuter and cargo trains are excluded
- Motion data is presented only in those control points where the delay (or extra delay) has recorded. That is, though a train is behind the schedule at a control point, the

absolute delay is not presented if it is the same or less than at the previous control point.

- The total number of instances is 5815, which in aggregate corresponds to 33479 delay minutes.

Cause allocation

Whenever the absolute delay exceeds 4 minutes, the train dispatcher must enter the code that tells the actual cause of the delay. After that, every extra delay of at least 1 minute must be coded similarly. In these cases, the system allocates only the extra minutes to the appropriate code, not the absolute delay at that certain point. The code is selected from a list that contains 60 codes in different categories. The categorization is illustrated in Table 2.

Table 2. Cause codes used in Finnish railways (classification into primary/secondary causes and into external/operator/infrastructure causes is made by the authors) (VR-Group Ltd 2009b)

Cause	Codes	Example
Primary causes		
External		
Traffic accidents	O1…O4	O2: Accident (animal)
Passenger service	M1…M6	M1: Passenger congestion
Freight service	T1…T4	T1: Loading/unloading of goods
Other causes	I1…I4	I1: Bad weather (e.g., fog)
Operator		
Personnel	H1…H3	H2: Waiting for personnel
Train setup	J1…J5	J3: Testing of brakes
Locomotives	V1…V4	V2: Technical fault
Rolling stock	K1…K7	K2: Bearing fault
Infrastructure		
Track	R1…R4	R3: Maintenance or construction
Electricity	S1…S4	S2: Technical fault
Other equipment	P1…P7	P3: Switch fault
Secondary causes		
Traffic system	L1…L8	L1: Waiting for connection transport
Totally 60 cause codes		

As seen in Table 2, the 60 codes are divided into 12 categories, each of which contains 3 to 8 codes. In the table, we have classified the codes a bit further. First, we have made a difference between primary and secondary causes. As stated in the first section, primary causes refer to the delays that are not due to other trains, whereas secondary delays are. Furthermore, we have divided the primary delays into three different groups, external, operator, and infrastructure. This classification can be used when allocating delays to the stakeholders responsible.

Official punctuality thresholds

In Finland, a long distance passenger train is considered punctual if it is at most 5 minutes behind schedule. Punctuality in long distance passenger traffic is presented as a percentage of trains that have arrived at their final destination punctually. Accordingly, a commuter train

is considered punctual if it is less than 3 minutes behind schedule. Punctuality in commuter traffic is presented as a percentage of the Helsinki region commuter trains that have departed from their station of departure and arrived at their final destination punctually (The Finnish Transport Agency 2009a). These measures are published every month by the Finnish Transport Agency.

Punctuality analyses for inter-company use

In addition to the official measures, both the rail administrator, the Finnish Transport Agency, and the only operator, VR-Group Ltd, analyze the punctuality data more profoundly for inter-company use. The analyses are done mainly to identify the most critical problems among the system, in order to steer the strategic and operational development activities. These kinds of analyses are done monthly, and contain the following kinds of examinations:

- Number of primary delay minutes divided into different cause classes, rail sections, and train types,
- The main causes for the most critical of the aforementioned delays,
- Rail sections and stations where have been the most secondary delays,
- The major traffic perturbations and their causes, and
- Punctuality of the ten most important trains.

Naturally, the rail administrator is the most interested in the delays caused by the infrastructure; whereas the operator pays attention to the delays caused, for example, by rolling stock (VR-Group Ltd 2009b, The Finnish Transport Agency 2009b).

The main weakness of the current analyses

According to our dataset described earlier in this section, secondary causes (coded with L1…L8) are responsible for more than 50 percent of the total delay minutes. In the current analyses, these causes are handled the same way as primary causes: they are presented as if they were the actual causes for the delays. However, as stated in the introduction section, the secondary delays take place only due to other trains. Thus, it would be very useful to link the secondary delays to the primary delays that caused the other delays: nowadays, 50 percent of the delay minutes are not allocated to their real causes. Our hypothesis is that this could set challenges, for example, prioritizing of development activities; how can one recognize the most critical problems, if half of the consequences of all the problems are left without attention?

4. EXAMINING THE DATA

Setting the challenge

According to previous section, it seems obvious that secondary delays must be linked to the corresponding primary delays. By doing so, we would be able to track delay chains from the secondary delay to the primary delay (which train caused the delay of the examined train)

and vice versa (which other trains were delayed due to the examined train). However, this is considered to be a quite challenging task; the railway traffic system forms a very complex network, and therefore, it is believed that tracking of delay chains is impossible in many situations. It is certainly true that in a single situation it might be impossible to track all the cause-consequence chains by using the current data. However, by examining a large number of similar situations, it should be possible to recognize at least the most critical correlations.

According to the punctuality report of the Finnish Transport Agency, the city of Tampere is one of the most critical areas when it comes to the emergence of secondary delays (The Finnish Transport Agency 2009b). Our dataset from September 2009 supports this perception; more than 10 percent of all the secondary delays took place in Tampere. This is because Tampere is a very congested crossing and transfer station. This can be seen in Figure 1, which illustrates the main part of the Finnish rail network. Based on the recognized problem, we decided to take Tampere as an example for this study.

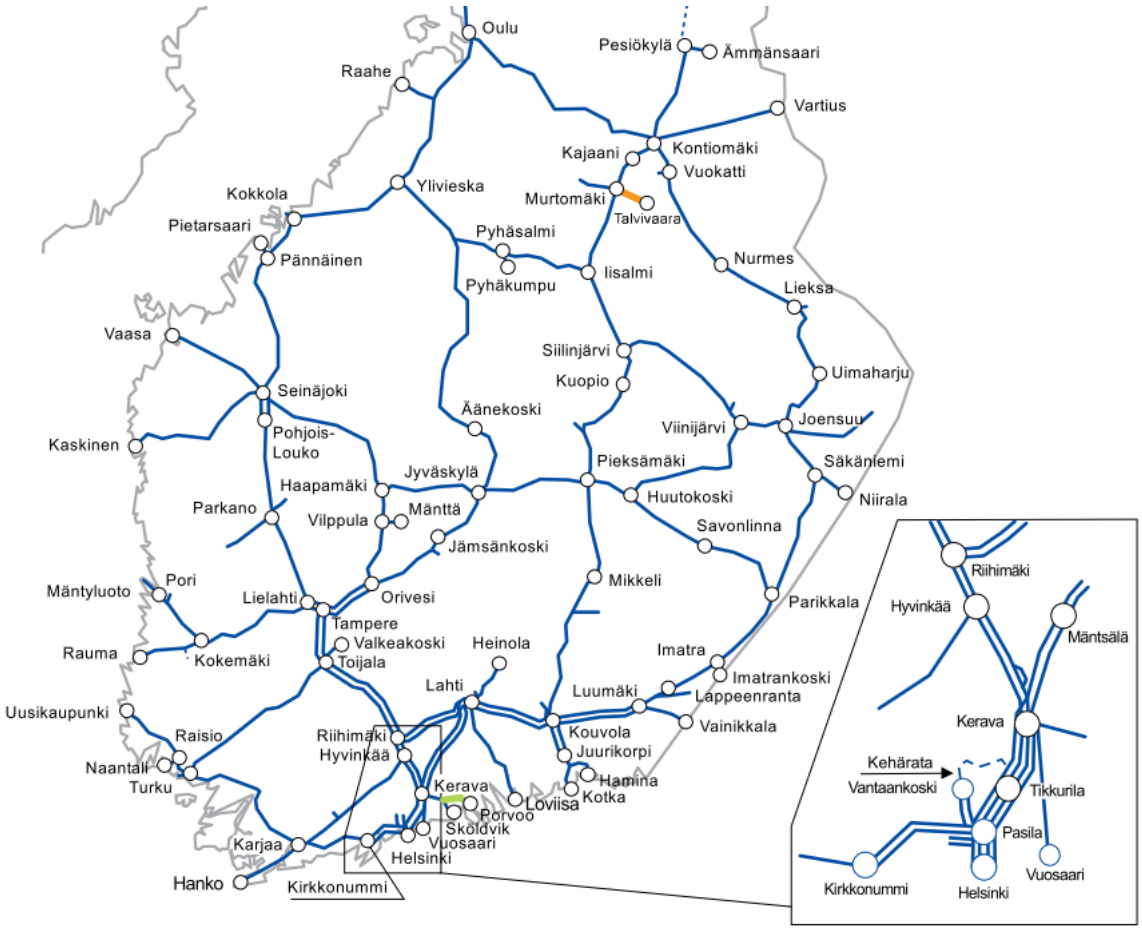

Figure 1. The main part of the Finnish rail network (The Finnish Transport Agency 2010).

Analyzing secondary delays in Tampere

In our example, we explore the dataset in order to find out if it possible to identify and map the most explicit delay chains in Tampere with the current data. If the answer is yes, the

same process could be used in the future when developing the analysis of the data to be more systematic, automatic, and efficient. Thus, with the dataset, we completed the following steps:

1. We filtered away everything else but the instances that fulfilled the following requirements:
   - The (extra) delay is recognized when arriving to or departing from Tampere, or when arriving at the first control point after departure from Tampere.
   - The (extra) delay is coded with L1 (waiting for connection transport) or L2 (passing of trains, siding of trains, or slower train ahead). These are the most explicit secondary delays. They require that the number of the cause train be entered, and besides, they constitute the great majority of all secondary delays, 87 percent of L coded delay minutes in our dataset.

2. We cross-tabulated the filtered data so that the columns of the table presented the trains that suffered from secondary delays, whereas the rows of the table presented the trains that caused the secondary delays. Among both columns and rows, we emphasized the trains that suffered or caused the most delay minutes (see Table 3). With this kind of illustration, we can able to identify the most critical correlations between the trains.

Table 3. The trains that suffered the most from secondary delays in Tampere and the trains that caused the most of these secondary delays.

		Delayed train								
		#49	**#55**	**#91**	**#93**	**#182**	**#475**	**#921**	**#927**	**Total**
Cause train	**#60**				62		76			138
	#88									117
	#90			42		46				114
	#94				114					132
	#916									124
	#922	48								90
	#928		33					48		152
	Total	67	60	118	196	80	80	86	76	**1962**

3. We chose train #93 for further examination as it suffered the most delay minutes (196 minutes) during September. The route of #93 is from the city of Helsinki to the city of Jyväskylä, via Tampere. As can be seen in the table, in Tampere, the majority of the secondary delay minutes—90 percent to be exact—of #93 were caused by two trains, #60 from the city of Oulu to Helsinki and #94 from the city of Pieksämäki to Helsinki. The former caused 62 delay minutes to #93 (coded with L1 in Tampere), whereas the latter is responsible for 114 delay minutes of #93 (coded with L2 between Tampere and Jyväskylä).

4. We took the aforementioned problematic trains into further examination, by using the original data. Thus, we filtered out everything else but these trains and highlighted the causes that have incurred the most of the delay minutes to them. It emerged that 33 percent of delay minutes of #60 were coded with R3 (maintenance or construction of

the track) and took place between the city of Oulu and Tampere. There was also a single animal accident (O2 code), which caused 14 percent of the delay minutes. Similarly, it came out that #94 was most often delayed for the following reasons: R3 between Jyväskylä and Tampere (26 percent of the total delay minutes), L2 in Jyväskylä because of the #927 (26 percent), and L1 in Pieksämäki because of #709 and #710 (14 percent).

5. Further, we examined more closely the trains that caused the aforementioned L1 and L2 delays to #94, that is, #927, #709, and #710; the analysis process was the same as in the previous step. After this, for illustrative purposes, we also examined #111 that seemed to be responsible for the delays of #709.

6. We sketched a diagram (Figure 2) that shows graphically the most explicit reasons for #93 suffering from secondary delays.

Reviewing the outcome

As described in Figure 1, the secondary delays of #93 in Tampere surroundings are most often due to two reasons:

- Waiting for passengers from #60 in Tampere (L1)
- Waiting for the track section between Tampere and Jyväskylä to be released by the delayed #94 (L2).

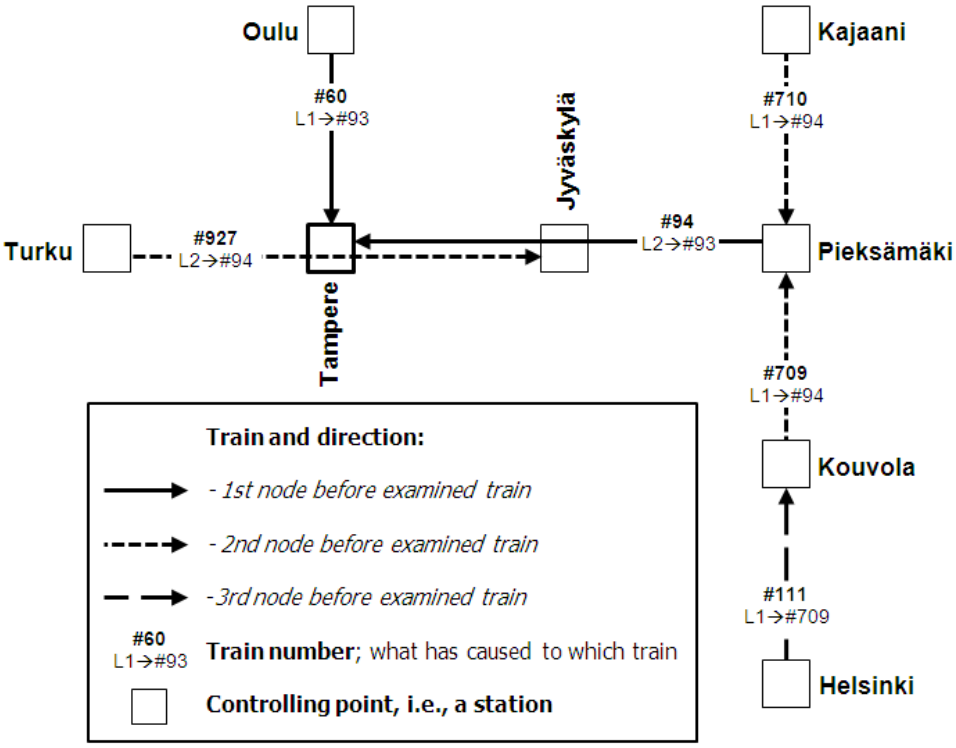

Figure 2. Trains that most affect the examined train, #93

Train #60 is often late because of maintenance or construction of the track (R3) between Oulu and Tampere. Thus, a part of the delays suffered by #94 should be allocated to the

code R3 in order to get the real understanding about the consequences of this specified maintenance or construction work. On the other hand, since the link between the trains has been proved to be critical, the railway organizations could add some more buffer time to the timetable of #60. All in all, this kind of information about the delay concatenation can be used in development activities.

Train #94 seems to be late due to three main reasons:

- R3 between Jyväskylä and Tampere,
- L2 in Jyväskylä because of #927,
- L1 in Pieksämäki because of #709 and #710.

So, a part of the secondary delays suffered by #94 should be allocated to R3, as was the case with #60. Similarly, a part of the delays should be allocated to the causes that affected #927, #709 and #710 to be late, and so on. As can be seen in the diagram, the delay chains can be tracked at least up to the 3rd node before the examined train. When recognizing this kind of critical but complex correlations, the railway organizations could examine whether the timetable structure could be altered to be more robust.

5. RESULTS OF THE DATA EXAMINATION

The main findings of the previous section can be summarized as follows:

- At least the most explicit delay chains can be tracked with the existing data
- Even the most trivial-looking delay can unfold a complex chain of delays
- The delay chains can be tracked at least down to the 3rd node; even with the existing, partly insufficient data
- By tracking the chains, at least the some of the secondary delays could be allocated to the primary causes.

Considering the findings above, the systematic analysis of the train punctuality data could provide answers for the following delay propagation related issues:

- What are the real consequences of a specified primary delay?
- What is the real cause of a specified train to be late?
- What are the most critical links between the trains when it comes to the whole railway network, that is, causing the most cumulative delays?
- Does the system contain too complex and vulnerable links that seem to cause too many cumulative delays? And if it does, is it possible to improve the system to be more robust?

In light of the aforementioned findings we can state that this kind of examination can and should be done, in order to achieve a deeper understanding of the system and delay concatenation. With this understanding, one is able to conclude both short- and long-term development and management activities. In short-term activity, as in real-time train dispatching, the information can offer answers to questions such as how to mitigate the emergence of secondary delays, how to prioritize trains, and how to recognize possible jams.

Correspondingly, in long-term planning, the information can help to develop the structure of the timetable in the form of adding buffer times, removing critical correlations and so on, in order to gain a more robust railway system.

However, the analysis process described in the previous section is quite laborious to do and it is not realistic to do it systematically. Hence, in the next section we discuss how to automate the process with the help of a selected data mining method.

6. DISCUSSION ABOUT THE POSSIBILITIES OF DATA MINING

As discussed in the previous section, doing the described analysis process systematically is too time-consuming without any automation. However, the process is quite simple and straightforward; so, if it were automated somehow, it could be used on a regular basis to get understanding of the state of the system. This automation could be done with data mining methods. Data mining is defined as “the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database” (Giudici 2003). Considering the definition above, data mining would seem to be a useful tool in examining train punctuality data for the following reasons:

- The amount of data is seemingly large (tens of thousands of observations a day),
- Considering the number of different trains and possible correlations among them, the data structure is rather complex,
- Many of the underlying patterns in such a large data mass are unknown at first.

The advantage of data mining over any traditional statistical techniques is that in statistics, there is usually a hypothesis based on some kind of a priori knowledge, and this hypothesis is then tested on the data in order to conclude whether the hypothesis is true or not (Anderson, Sweeney & Williams 2009). In data mining, no prior knowledge is needed (Giudici 2003), which does not restrict the user to dealing with assumptions or preconceptions that might not be known beforehand in the first place.

Currently, as mentioned in section 2, the reports related to train punctuality data examine only the primary delays when it comes to the delay causes. That is, the reports do not examine the dependencies between the trains in the perturbation situations. However, as we demonstrated in section 3, it is possible to reveal at least the most explicit delay chains with a quite simple process, by using the existing data. The problem is that a railway traffic network forms a complex system, making examining all the chains in this manual manner an impossible task. Thus, the process must be somehow automated. One interesting way to do this seems to be a data mining method called sequence analysis.

Sequence analysis is an extension of a popular association rules technique (Kudyba 2001) where the analyst is faced with a task of finding which items occur together the most in a dataset. This technique is often referred to as market basket analysis as it is usually used in order to find out which items customers purchase in stores (Kudyba 2001, Han, Kamber

2006). A number of probabilistic figures can be used to gain more insight into the occurrences of the different combinations (rules) of items in a dataset. Association rules, however, do not take into account the order in which the occurrences, for example, delays, have happened. In order to identify and map the real delay chains, this information is, however, obligatory. Sequence analysis takes into account the temporal order in which the items occur in the dataset (Giudici 2003). By utilizing sequence analysis, it is possible to describe the whole delay chain from the first event (primary delay) to the last (secondary delay that doesn't cause any further delays), or vice versa. By using the aforesaid probabilistic figures, it is possible to identify, e.g., the most frequent delay chains. Hence, it seems that with the sequence analysis technique the analysis process described in the previous section could be automated at least to some extent.

Once the most crucial delay chains have been identified and mapped, the next logical step would be optimizing the system so that these chains are removed or their negative effects are minimized. At this point, Finnish railways have yet to begin delay chains identification, so any optimization activities are purely on a theoretical level, and are heavily dependent on the operator and infrastructure administration management's vision. However, the sequence analysis technique complemented with other data mining approaches could provide answers for the delay concatenation-related questions mentioned in the previous section.

To conclude, the authors propose that different data mining techniques would provide significant improvement over traditional statistical techniques when analyzing the train punctuality data in order to gain an understanding of the overall system. This knowledge would help railway management make better and more informed decisions in their management activities, concerning operational and strategic decisions. Compared to any traditional analysis technique, data mining gives the possibility to examine the whole traffic network, instead of modeling only a part of it.

7. CONCLUSION AND FUTURE RESEARCH

In this paper, we have shown the significance of secondary delays especially in railway traffic systems with a high proportion of single tracks. We discovered the current ways to analyze delays based on the train motion data and found that data mining could offer a novel approach for understanding the concatenation of delays. Then we proved manually that at least the most explicit delay chains can be identified with the existing data. Finally, we discussed the evident potential of the selected data mining method for automating the described process.

In Finland, 890 commuter services and 310 long-distance trains are operated daily, as well as 500 freight trains (VR-Group Ltd 2009a). When providing this much service on a rail network that consists mainly of single lines, the overall system becomes quite prone to perturbations. This reverberates in the number of secondary delays, which in Finland forms more than half of the delay minutes. In addition to the significance, the railway organizations

do not have the means to analyze the delay concatenation process, that is, how the primary delays lead to secondary delays.

With actual motion data from Finnish railways, we were able to prove that at least the most explicit delay chain can be identified and mapped, and this information can be used for allocating the delays to their real causes, developing the timetable to be more robust, and so on. Thus, we suggest that this kind of examination can and should be done, in order to achieve a deeper understanding of the system and delay concatenation. With this understanding, one is able to conclude both short- and long-term development and management activities. However, the analysis process is so laborious that it should be automated somehow. One good approach seems to be a data mining method called sequence analysis, which, with the other data mining techniques, would provide significant improvement over traditional statistical techniques when analyzing the train punctuality data to gain an understanding of the overall system.

Since the main contribution of this paper was to prove that the tracking of delay chains is possible and could be done systematically with selected data mining methods, the results and observations of this study can be generalized to be applicable in many other countries. Hence, it can be said that the study achieves both scientific and pragmatic significance, for this study takes the first step to examine secondary delays within the whole railway traffic system and suggests a useful process to end-users for analyzing the actual data.

In our ongoing research project, we are testing the potential of the selected data mining in practice. We have also planned to link the passenger flows to the examination of the train punctuality. Our hypothesis is that the number of passengers in different trains and routes will alter the significance of a single train.

8. ACKNOWLEDGEMENTS

This study is part of a long-term research program conducted in cooperation with the authors' university and Finland's national rail administration.

REFERENCES

Albrecht, T., Brünger, O., Dahlhaus, E., Goverde, R.M.P., Hansen, I.A., Huisman, D., Jacobs, J., Kroon, L.G., Maróti, G., Martin, U., Pachl, J., Radtke, A., Siefer, T., Wendler, E. & Yuan, J. 2008, *Railway timetable & traffic: analysis, modelling, simulation,* Eurailpress (DVV Rail Media), Hamburg.

Anderson, D.R., Sweeney, D.J. & Williams, T.A. 2009, *Statistics for business and economics,* 10th ed., Thomson South-Western, Mason (OH).

Carey, M. 1999, "Ex ante heuristic measures of schedule reliability", *Transportation Research Part B: Methodological,* vol. 33, no. 7, pp. 473-494.

Carey, M. & Carville, S. 2000, "Testing schedule performance and reliability for train stations", *Journal of the Operational Research Society,* vol. 51, no. 6, pp. 666-682.

Carey, M. & Carville, S. 2003, "Scheduling and platforming trains at busy complex stations", *Transportation Research Part A: Policy and Practice,* vol. 37, no. 3, pp. 195-224.

Carey, M. & Crawford, I. 2007, "Scheduling trains on a network of busy complex stations", *Transportation Research Part B: Methodological,* vol. 41, no. 2, pp. 159-178.

Castillo, E., Gallego, I., Ureña, J.M. & Coronado, J.M. 2009, "Timetabling optimization of a single railway track line with sensitivity analysis", *TOP,* vol. 17, no. 2, pp. 256-287.

de Kort, A.F., Heidergott, B. & Ayhan, H. 2003, "A probabilistic (max, +) approach for determining railway infrastructure capacity", *European Journal of Operational Research,* vol. 148, no. 3, pp. 644-661.

Dingler, M.H., Lai, Y.-. & Barkan, C.P.L. 2009, *Impact of train type heterogeneity on single-track railway capacity*.

Gibson, S., Cooper, G. & Ball, B. 2002, "Developments in transport policy: The evolution of capacity charges on the UK rail network", *Journal of Transport Economics and Policy,* vol. 36, no. 2, pp. 341-354.

Giudici, P. 2003, *Applied data mining: Statistical methods for business and industry,* Wiley, New York.

Han, J. & Kamber, M. 2006, *Data Mining: Concepts and Techniques,* 2nd ed., Elsevier, Amsterdam.

Higgins, A., Kozan, E. & Ferreira, L. 1997, "Modelling the number and location of sidings on a single line railway", *Computers and Operations Research,* vol. 24, no. 3, pp. 209-219.

Huisman, T. & Boucherie, R.J. 2001, "Running times on railway sections with heterogeneous train traffic", *Transportation Research Part B: Methodological,* vol. 35, no. 3, pp. 271-292.

Kudyba, S. 2001, *Data mining and business intelligencea guide to productivity,* Idea Group Publishing, Hershey (PA).

Lindfeldt, Olov. 2007. Quality on single-track railway lines with passenger traffic: Analytical model for evaluation of crossing stations and partial double-tracks. Licentiate thesis. Stockholm: KTH. 49 p. ISSN 1653-445X; ISBN 978-91-85539-27-7.

Mattsson, L. 2007, "Railway capacity and train delay relationships" in *Critical infrastructure: reliability and vulnerability*, Springer, Berlin, Heidelberg, pp. 129-151.

Olsson, N.O.E. & Haugland, H. 2004, "Influencing factors on train punctuality – results from some Norwegian studies", *Transport Policy,* vol. 11, no. 4, pp. 387-397.

Peters, J., Emig, B., Jung, M. & Schmidt, S. 2005, "Prediction of delays in public transportation using neural networks", *International Conference on Computational Intelligence for Modelling, Control and Automation, CIMCA 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, IAWTIC 2005* Institute of Electrical and Electronics Engineers Computer Society, Piscataway, pp. 92.

SysOpen Oyj 2003, *The definition of JUSE active database*, Finland.

The Finnish Transport Agency 2010, *F 8/2009 Description of the rail network 1.1.2010*, The Finnish Transport Agency, Finland.

The Finnish Transport Agency 2009a, *Punctuality in long distance and commuter traffic* [Homepage of the Finnish Transport Agency], [Online]. Available: http://www.rhk.fi/in_english/traffic_and_network_access/punctuality_in_long_distance_and/ [2010].

The Finnish Transport Agency 2009b, *Punctuality of railway traffic in September 200*9, The Finnish Transport Agency, Finland.

VR-Group Ltd 2009a, *High level of punctuali*ty. Available: http://www.vrgroup.fi/index/vastuullista_toimintaa/uusicustomerservice/punctuality.html [2010].

VR-Group Ltd 2009b, *Punctuality of passenger traffic in September 200*9, VR-Group Ltd, Finland.

Vromans, M.J.C.M. 2005, *Reliability of railway systems*, Erasmus University of Rotterdam.

Vromans, M.J.C.M., Dekker, R. & Kroon, L.G. 2006, "Reliability and heterogeneity of railway services", *European Journal of Operational Research,* vol. 172, no. 2, pp. 647-665.

Vukadinovic, K., Teodorovic, D., Pavkovic, G. & Rosic, S. 1996, "A neural network approach to mitigation of vehicle schedule disturbances", *Transportation Planning and Technology,* vol. 20, no. 1, pp. 93-102.

Yuan, J. & Hansen, I.A. 2007, "Optimizing capacity utilization of stations by estimating knock-on train delays", *Transportation Research Part B: Methodological,* vol. 41, no. 2, pp. 202-217.

Yuan, J. 2006, *Stochastic modelling of train delays and delay propagation in stations*, Delft University of Technology.

Yuan, J. 2004, "An analytical model for estimating the propagation of train delays in complex station areas", *A World of Transport, Infrastructure and Logistics, Proceedings of 8th TRAIL Annual Congress*, pp. 1-19

Yuan, J., Goverde, R.M.P. & Hansen, I.A. 2002, "Propagation of train delays in stations", *Computers in Railways VIII,* pp. 975-984

Zhou, X. & Zhong, M. 2007, "Single-track train timetabling with guaranteed optimality: Branch-and-bound algorithms with enhanced lower bounds", *Transportation Research Part B: Methodological,* vol. 41, no. 3, pp. 320-341.