Detecting train reroutings with process mining: A Belgian application

Peer-reviewed author version

JANSSENSWILLEN, Gert; DEPAIRE, Benoit & VERBOVEN, Sabine (2018) Detecting train reroutings with process mining: A Belgian application. In: EURO Journal on Transportation and Logistics, 7(1), p. 1-24.

DOI: 10.1007/s13676-017-0105-8
Handle: http://hdl.handle.net/1942/23488

Noname manuscript No.
(will be inserted by the editor)

Detecting train reroutings with process mining

A Belgian application

Gert Janssenswillen · Benoît Depaire · Sabine Verboven

Received: date / Accepted: date

Abstract One of the objectives of railway infrastructure managers is to improve the punctuality of their operations, while satisfying safety requirements and coping with limited capacity. In order to fulfil this objective, capacity planning and monitoring have become an absolute necessity. Railway infrastructure managers possess tremendous amounts of data about the railway operations, which are recorded in so-called train describer systems. In this paper, a set of methods is proposed to guide the analysis of capacity usage based on these data. In particular, train connections are categorized according to the severity of train reroutings as well as the diversity of these reroutings. The applied method is able to highlight areas in the railway network where trains have a higher tendency to diverge from their allocated route. The method is independent from the underlying infrastructure, and can therefore be reused effortlessly on new cases. The analysis provides a starting point to improve the planning of capacity usage and can be used to facilitate the communication between capacity planning at one hand and operations on the other hand. At the same time, it presents an illustration on how process mining can be used for analysis of train describer data.

Keywords Process Mining · Data Analysis · Train Describers · Rail Traffic Monitoring · Train Reroutings

G. Janssenswillen · B. Depaire
Hasselt University, Agoralaan Bldg D, 3590 Diepenbeek, Belgium
E-mail: gert.janssenswillen@uhasselt.be

G. Janssenswillen
Research Foundation Flanders (FWO), Egmontstraat 5, 1000 Brussels, Belgium

S. Verboven
Infrabel, Fonsnylaan 13, 1060 Brussels, Belgium

1 Introduction

Improving the punctuality of railway operations is one of the most important objectives of rail infrastructure managers. In reaching this goal, they are restricted by safety constraints and capacity limitations. As for the latter restriction, optimizing capacity planning and monitoring constitutes a major necessity.

In order to bridge the gap between railway scheduling and execution, it is necessary to analyse to which extent traffic operators make decisions to deviate from the planned capacity allocation. Consequently, it needs to be examined whether these decisions were favourable, thereby possibly pointing at flaws in the railway planning, or not. While a lot of research on train scheduling and realtime rescheduling exist, little literature is available on the ex post analysis of capacity usage. Most research in this area is focussed on train delays and ensuing conflicts, while limited consideration has been given to the evaluation of train rescheduling. Nevertheless, railway infrastructure managers possess tremendous amounts of data about the railway operations, which are recorded in so-called train describer systems. Because of the abundance of data, extracting knowledge from it, is a complicated task.

The contribution of this paper is twofold. Firstly, metrics are proposed to evaluate train scheduling by using train describer data. The metrics allow to identify areas in the train schedule where reroutings are frequent, and will provide guidelines to consequently improve the scheduling of trains. Train describer data recorded by the Belgian railway infrastructure manager Infrabel will be used to illustrate the workings of the suggested metrics. Secondly, the paper illustrates the large potential of process mining techniques to analyse train describer data. Process mining is a relatively young research discipline which aims at the extraction of process-related knowledge from event data (van der Aalst, 2011). The metrics used to quantify deviations are drawn from the field of process mining, which is well suited for the analysis of event data.

The next section will discuss related work. Consequently, a set of metrics will be developed in Section 3, together with a methodology to use them. Finally, in Section 4, the methodology will be illustrated using train describer data recorded by the Belgian railway infrastructure manager Infrabel.

2 Related work

Improving the punctuality of railway operations starts with the development of a robust train schedule. In this area, the work in Törnquist (2006) should be noted. The author provides an overview of 48 techniques for railway scheduling. These techniques were categorized according to the plan perspective, the supported infrastructure, the goal, the level of evaluation and the control strategy. It demonstrates that most attention in literature goes to techniques for tactical scheduling and less to operational scheduling. Moreover, a consider-

able amount of techniques can only be applied on line-infrastructures, and not on more complex and realistic network-infrastructures.

More recently, robust scheduling in a more complex railway infrastructure was investigated in Dewilde (2014). The main focus of this PhD thesis was on the robustness of the complex station area of the North-South connection in Brussels. The author identified the different elements which determined the robustness of a train schedule, and hereupon defined an approach to improve the robustness, by taking into account routing decisions, train sequences and platform allocation.

Once railway schedules are put into service, the performance of the operations needs to be monitored and evaluated. Train conflicts and delay propagations have been studied extensively in literature of transportation and operations research. The Belgian train describer data used in this paper were analysed before with respect to train delays (Cule et al, 2011). Using frequent itemset mining, patterns between train delays were detected. If train A has a delay of x minutes or more, train B will also have a delay with x minutes or more, with a certain confidence y.

The use of train describer data for data analysis, as in this paper, has been done in Kecman and Goverde (2015a) and Kecman and Goverde (2015b). In this work, the authors aimed at the adaptive prediction of train event times, i.e. taking into account not only delay but also predicted route conflicts, braking and acceleration times.

Conte and Schöbel (2008) identified three different types of delay propagation: propagation along the same train, propagation between trains due to required connections, and propagation between trains due to shared use of scarce infrastructure capacity. The last type of delay propagation is better known as *knock-on delays* (Carey and Kwieciski, 1994; Higgins and Kozan, 1998; Yuan and Hansen, 2007). These three types of propagations were analysed through the use of stochastic models.

Related to the work of Conte and Schöbel (2008), Flier et al (2009) present efficient algorithms to detect both resource conflicts and delays from maintained connections, within large scale data sets. Further steps which are proposed are a statistical examination and to extend the approach to global dependencies. The latter could for instance lead to the construction of networks of conflicts between trains.

In the same area, D'Ariano (2008) focussed on real-time dispatching. The objective of this PhD thesis has been to develop a decision support system for realtime management of railway traffic. The resulting tool, called ROMA, *Railway traffic Optimization by Means of Alternative graphs*, assists traffic managers in choosing the *best* trajectory, ordering of trains and the optimal speed of trains. The recommendations done by the system are based on simulations of the resolution of traffic after certain decisions are taken.

In Weeda and Hofstra (2008), the authors advocate that it is important to have feedback from operations to planning, to close the controle loop. In order to achieve this, the performance of the railway operations in the Dutch railway are analysed and this is used as input towards a better planning. In The

Netherlands, train describer data have been used to identify route conflicts. The TNV-conflict tool introduced in Daamen et al (2008) defines a train conflict as the situation in which a train comes within sight distance of a signal which is not open, i.e. obliging the train to slow down or halt. Both conflicts due to scarce capacity and required train transfers were identified, in accordance with the different delay propagations proposed in Conte and Schöbel (2008). The additional tool TNV-statistics has been developed to look into the conflicts with more detail, and to link them together in conflict chains or trees (Goverde and Meng, 2011).

In Sammouri (2014), several data mining techniques are applied on a large set of censor data generated by railway infrastructure and rolling stock. The aim of the analysis is to use temporal sequences of recorded events to predict failure of equipment, as to improve maintenance scheduling. Although not related to routing conflicts, it shows how much can be learned from analysing the great amount of data which is available.

Because of the huge expansion of process event data during the last couple of decades, companies are dealing with the challenge of retrieving useful insights from it, and apply those to gain competitive advantages. By getting a better understanding of business processes and improving them, process mining provides ways to reach this goal (van der Aalst, 2011). The birth of process mining dates back to the end of the previous century (Agrawal et al, 1998; Cook and Wolf, 1999), and focused on the retrieval of process control-flow from event logs containing recorded behaviour. Although the field has become much broader, control flow discovery is the most mature research track within process mining. An overview of existing process discovery algorithms can be found in De Weerdt et al (2012). In order to measure the quality of a discovered process model, different quality dimensions have been defined (Rozinat et al, 2007), i.e. fitness, precision, generalization and simplicity. For each of these dimensions, several metrics have been developed and implemented, of which an overview can be found in vanden Broucke et al (2013).

3 Design and development of data analysis methods

Apart from the TNV-statistics tool and the work in Cule et al (2011); Sammouri (2014), little attention has been directed to the analysis of recorded data. Nevertheless, event data such as train describer data can be used to extract process-related knowledge using process mining. The minimal requirement to pursue process mining is that each *event* can be related to both a *case* and an *activity* (van der Aalst, 2011). A case refers to a particular instance of the process, e.g. a specific train trip. Activities are specific types of events. For instance, activities related to a train trip can be the passing of a signal, or the adjustment of its trajectory. Furthermore, each event should have a timestamp attached to it. Other attributes may be available, which can be related to the event as well as to the case. Typical additional event attributes relate to resources. As such, the passing of a signal may also record which signal

was passed. Case attributes can contain any characteristic information about the process instance. For instance, the type of rolling stock, or the number of carriages.

The main focus of this paper is to analyse how recorded train routes deviate from the planned route. Thus, two routes are needed for each train: the planned route and the actual route. In our analysis, the planned routes refer to the routes which are communicated to the signal area before the entering or departure of the train. Note that hereby, anticipated changes to the capacity allocation, e.g. due to infrastructure works, are neutralized. Both planned and actual routes have been defined at the level of signals. In order to describe the complete path, also the final track segment has been taken into account. This is the track where the train arrives in the destination station or where it leaves the signal area. Considering this track segment is essential since the train might have different routes after passing the last signal. Reroutings on this point of the route include platform changes, and should therefore not be ignored. Formally, we define the actual and planned route as follows. An overview of the terminology used in this paper is provided in Table 1.

Definition 1 (Preliminaries) We define $\mathcal{S}$ as the alphabet of signals and $\mathcal{T}$ as the alphabet of track segments. $\mathcal{S}^*$ is the set of al finite sequences over $\mathcal{S}$.

Definition 2 (Actual route) The actual route of a train i, denoted by σ_i, is defined as a sequence of signals plus the destination track segment of the train within the area. Given an $s \in \mathcal{S}^*$ and a $t \in \mathcal{T}$, we can define σ_i as $< s, t >$.

Definition 3 (Planned route) The planned route of a train i, denoted by π_i, is the allocated route of a train 30 minutes before it enters the signalling area. It consists of a sequence of signals plus the destination track segment of the train within the area. Given an $s \in \mathcal{S}^*$ and a $t \in \mathcal{T}$, we can define π_i as $< s, t >$.

Given the planned and actual route, rerouting can be formally defined as follows:

Definition 4 (Rerouting) A rerouting, or deviation, of a train i is defined as the case where a difference exist between the planned route of a train and the actual route of a train, i.e. $\sigma_i \neq \pi_i$.

Using the process mining tool Disco[1], recorded process behaviour can be easily visualised as a directed graph. Graphs G_1 and G_2 in Figure 1 show the visualisation of two fictitious groups of train trips. The hypothetical underlying infrastructure is shown in Figure 2. In the directed graphs, each node refers to a signal which was passed by one or more trains, and each edge refers to a route from one signal to the next that was taken by at least one train. Both nodes and edges are annotated with the number of running trains, which are also visualised by their colour and width, respectively. The darkest path

[1] `http://fluxicon.com/disco/`

Table 1 Terminology used in this paper.

Terminology	Description
Route	A route of a train is a sequence of signals, when needed supplemented with additional details, such as track segments.
Planned route	The planned route of a train, based on planning and apriori known disruptions.
Actual route	The actual route of a train, based on train describer data.
Rerouting	The case where the actual and planned route of a train differ.
Connection	A connection refers to all train trips from station A to station B, when needed taking into account certain waypoints (in case there are multiple way to go from A to B). This does not take into account whether the train stops in all station or just in major cities.
Relation	A relation refers to all train trips from station A to station B, and vice versa, when needed taking into account certain waypoints (in case there are multiple way to go from A to B). This does not take into account whether the train stops in all station or just in major cities.

throughout the graph, i.e. the most frequent path, corresponds to the planned route in this example. When all trains visualised in a graph have the same planned route, reroutings become readily noticeable.

Based on an exploratory inspection of the recorded train routes and interviews with business experts, two dimensions seemed relevant to quantify train reroutings. Firstly, the *severity* of the reroutings should be measured. This refers to both how many deviations occurred and how long they are. Upon inspection of both graphs, one can see that more reroutings occurred in G_1 compared to G_2. Furthermore, reroutings in G_2 seem to be less severe, as they take up at most two signals. In contrast, in G_1, only about three quarters of the trains passed through signal AD as planned, and only 2 signals of the planned routes were never deviated from. The severity of reroutings will be referred to as the *rerouting severity*.

Secondly, the complexity and structuredness of the graphs are relevant, as they represent how many *different* reroutings have occurred. In Figure 1, the model on the left is clearly less complex than the model on the right. The complexity of the model will be used as a proxy for *rerouting diversity*.

Visual inspection of all planned routes would, however, be a cumbersome task. Therefore, the next paragraphs suggest metrics to quantify the severity and diversity of reroutings and to single out the routes which should be examined more closely.

3.1 Rerouting severity

In order to measure rerouting severity, we draw upon insights of conformance checking within process mining. Given a process model, conformance checking determines whether the events that were recorded can be *replayed* by the process model (van der Aalst, 2011). In van der Aalst et al (2012), the *alignment-based fitness* measure has been defined, which is one of the best-known met-

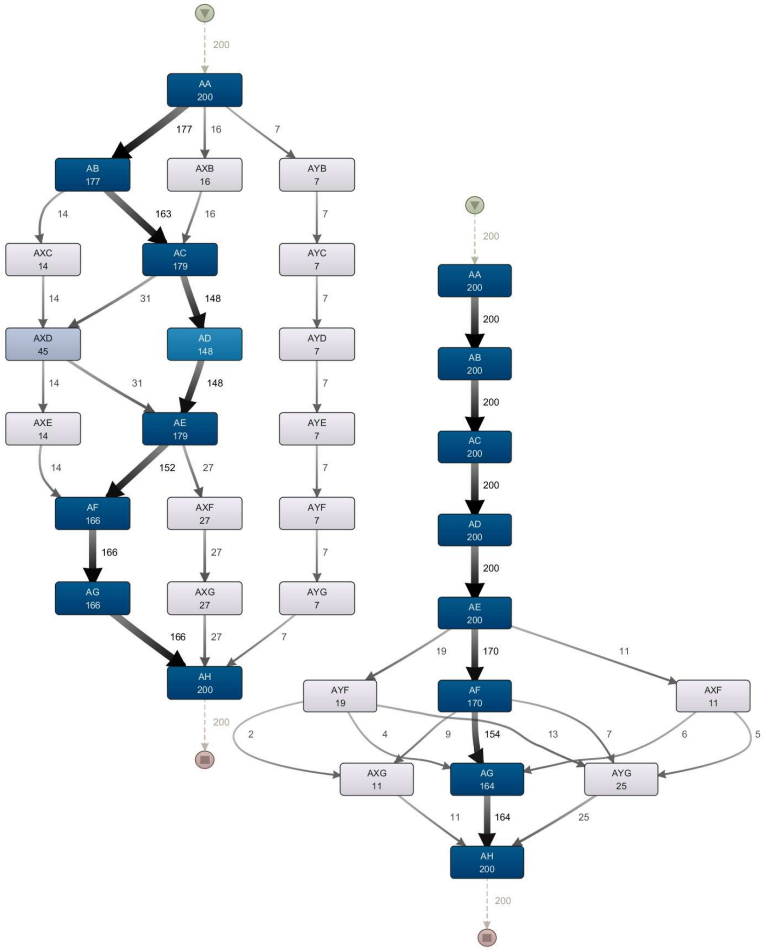

Fig. 1 Fictitious actual routes for two sets of train trips with planned route AA → AB → AC → AD → AE → AF → AG → AH.
Referred to as G_1 (left) and G_2 (right)

rics within conformance checking. In general terms, each case is *aligned* to the most optimally corresponding execution trace of a process model, according to a cost-function. For cases which are allowed by the model, the cost of the alignment is obviously zero. For cases which cannot be replayed by the model, corrections have to be made. A correction can be an insertion of an event, a deletion of an event, or the substitution of an event. Note that multiple alignments can be made, which each have their own cost. Using default values, a single insertion or deletion has a cost of 1, while a substitution is allocated

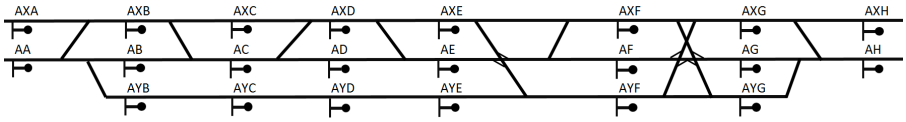

Fig. 2 Example infrastructure.

a cost of 2. The most optimal alignment will be used to compute the overall fitness between the recorded behaviour and the model.

In the context of train deviations, suppose we have a group of k trains which were allocated the same planned route. Let π_L be the planned route of the trains and let $L = \{\sigma_1, ..., \sigma_k\}$ refer to the set of the actual routes of the trains. For train i, given the actual route $\sigma_i \in L$ and π_L, we define $\lambda_{\pi_L}(\sigma_i)$ as the optimal alignment for the actual route σ_i and $\delta(\lambda_{\pi_L}(\sigma_i))$ as the corresponding cost. It should be noted that there are multiple ways to align the actual route with the planned route, i.e. different combinations of insertions and deletions will change the actual route into the planned route. Here, the optimal one is chosen, i.e. with the lowest cost. The fitness for train i is then defined as

$$f(\sigma_i, \pi_L) = 1 - \frac{\delta(\lambda_{\pi_L}(\sigma_i))}{|\sigma_i| + |\pi_L|} \tag{1}$$

where $|\sigma_i|$ and $|\pi_L|$ refer to the length of the actual route and planned route, respectively. In the worst-case scenario, a train followed a completely different sequence of signals throughout the network. To align such a route, all signals that were passed by the train need to be removed, while all signals on the planned route need to be inserted. Consequently, the total cost will equal the nominator, resulting in a fitness value equal to zero. In the optimal case, when $\sigma_i = \pi_L$, then $\delta(\lambda_{\pi_L}(\sigma_i)) = 0$, yielding a fitness value of one. Given the fitness values for all individual trains, the overall fitness value can be computed for a set of train trips L as follows:

$$Fitness\ F(L) = \frac{\sum_{\sigma_i \in L} f(\sigma_i, \pi_L)}{|L|} \tag{2}$$

Table 2 Example of an alignment between actual route σ_1 and planned route π_L.

π_L	AA	$\perp$	AB	AC	$\perp$	AD	$\perp$	AE	AF	AG	AH
σ_1	AA	AXB	$\perp$	AC	AXD	$\perp$	AXE	$\perp$	AF	AG	AH

Table 2 shows a fictitious example alignment between planned route π_L and actual route $\sigma_1 =< AA, AXB, AC, AXD, AXE, AF, AG, AH >$. Three signals of the planned route were not passed and were thus deleted from the route, as indicated with the $\perp$-symbol. Furthermore, three signals were visited, although they didn't belong to the planned route, which results in three insertions. Notice that, in this case, each consecutive pair of one insertion

and one deletion can also be regarded as a substitution, which would yield an equivalent optimal alignment according to the default cost-function. However, in general, it is not obligatory to have one deletion for each insertion, or vice versa. Since the planned route, as well as the actual route, consists of 8 signals, Equation (1) results in a fitness value of 0.625.

The overall fitness values for each of the planned routes will be used as a proxy for the rerouting severity. The lower the fitness value, the more sensitive the route is towards train reroutings. In Figure 1, it was already clear that, weighted by the frequencies, slightly more reroutings occurred in G_1 compared to G_2. Indeed the Fitness-metric for the set of train trips G_1 is 0.8844, while for G_2 it is 0.9175.

An analysis of variance can be performed to see whether deviation severity differs significantly among different groups of trains. These groups can be composed in different ways, depending on the purpose of the analysis: e.g. comparing trains on different itineraries, comparing trains at different times of the day, etc. Pairwise differences between groups and corresponding p-values can then be used to identify which specific groups perform significantly worse or better.

Once the interesting cases have been identified, the reroutings can be scrutinized further. For instance, are there only a limited number of distinct deviations, or are there many different ones? How are they distributed along the route? How many distinct reroutings generally happen at one specific point of the route, on average? In order to answer these questions, the dimension of *rerouting diversity* will be further defined in the next paragraph.

3.2 Rerouting diversity

The aim of this second dimension is to investigate whether trains on a certain route always deviate in a similar manner or have many different reroutings over time. In order to measure diversity, we take a new look at the directed graphs displaying all recorded behaviour, as those shown in Figure 1. The complexity of these models can be used as proxy for the deviation diversity.

Based on the visual inspection of a series of graphs, it was observed that diversity cannot be measured in a single metric. For instance, in the lower part of G_2, about 8 different routes have been observed from signal AE to signal AH. This is remarkably more than the number of different routes observed at any point in G_1. It is therefore said that the reroutings of G_2 are *wider*. This type of diversity will be referred to as *horizontal diversity*. Conversely, deviations in G_1 have occurred in a larger part of the itinerary, i.e. on all signals except for the first and the last. This type of diversity will be referred to as *vertical diversity*. Two different process complexity metrics have been adapted to the specific context of this paper, both taking into account one specific type of diversity. Both metrics are discussed in the following paragraphs.

3.2.1 Horizontal diversity

The *Extended Cyclomatic Metric (ECyM)*, or *cyclomatic complexity* has been defined by Thomas J. McCabe (McCabe, 1976) as a means to estimate the testability and maintainability of software systems. It uses a directed graph as input, consisting of *nodes* and *edges*. Given the number of edges e, the number of nodes n and the number of connected components p, ECyM was defined as

$$ECyM(e, n, p) = e - n + 2p \tag{3}$$

Note that the formula for the cyclomatic complexity differs from the formula for the cyclomatic number, which is equal to $e-n+p$. The cyclomatic number only has a logical interpretation in the context of strongly connected graphs[2]. In contrast, the cyclomatic complexity is primarily directed towards graphs which are not strongly connected, but which have clear start and end points, as in our case. However, the cyclomatic complexity is equal to the cyclomatic number of a graph in which and extra edge was added from the end to the start of every component, in order to make the components strongly connected (Watson et al, 1996). As a result, $e-n+p+p = e-n+2p$. As stated in McCabe (1976), in a strongly connected graph, the cyclomatic number is equal to the maximum number of linearly independent circuits. Consequently, $ECyM$ is meant to quantify *horizontal diversity*.

3.2.2 Vertical diversity

In order to measure *vertical diversity*, Separability (Π) is introduced. Mendling et al (2007) defined the notion of Separability, referring to the number of cut-vertices in a graph. A cut-vertex can be defined as a node which *separates* the graph into two parts when it would be deleted. As such, it provides an estimate of the modularity of a process model. Formally, given a set of actual train routes L,

$$\Pi = |\{s \in \mathcal{S} \mid \forall\, \sigma_i \in L :\ s \in \sigma_i\}| \tag{4}$$

Within the context of train deviations, a cut-vertex s is a signal through which all actual routes $\sigma_i \in L$ have passed. When more cut-vertices are present, it means there is a higher proportion of the planned route which is never deviated on. However, this only holds under the assumption that each signal on the planned route was passed by at least one of the trains. If this does not hold, a cut-vertex can also be a signal through which all trains have passed, although it did not belong the the planned trajectory. Yet, to measure diversity, it is irrelevant whether the cut-vertex belongs to the planned route or not.

Complexity, as measured by the metrics discussed above, tends to increase as the size of the graph increases. This is indeed a desirable property of complexity measures in the context for which they have been defined. However,

[2] A graph is strongly connected if there is a path from each node to any other node.

longer routes will therefore be negatively biased, i.e. obtaining higher complexity scores. To take this into account, both metrics were corrected for the length of the planned route. Furthermore, the complement of the separability metric is taken, so that higher values correspond to a higher diversity, as is the case with $ECyM$.

$$ECyM'(e, n, p) = \frac{e - n + 2p}{|\pi_L|} \tag{5}$$

$$\Pi' = 1 - \frac{|\{s \in \mathcal{S} \mid \forall\, \sigma_i \in L :\ s \in \sigma_i\}|}{|\pi_L|} \tag{6}$$

While values of Π' are generally in the range from 0 to 1, values of $ECyM'$ are not. In theory, a graph with n nodes can have as many as $\frac{n(n-1)}{2}$ edges. In such a case, the nominator of $ECyM'$ would be equal to $\frac{n^2-3n+4}{2}$. Assuming that half of the nodes in the graph are actually on the planned route, $ECyM'$ is equal to $n - 3 + 4/n$. Consequently, the diversity of such a graph would be a nearly linear function of its size, even when the length of the reoute is incorporated. This means that the range of $ECyM'$ is not really limited. However, in reality, not all nodes, i.e. signals, are connected with each other, since the infrastructure is limited. The upper bound of the $ECyM'$ metric is thus dependent on the corresponding infrastructure.

In order to calculate the $ECyM'$ of each graph in Figure 1, the number of nodes and edges needs to be counted. G_1 contains 20 nodes and 25 edges, while G_2 contains 12 nodes and 18 edges. Thus, $ECyM'(G_1) = 0.875$ and $ECyM'(G_2) = 1$. To compute the adjusted separability-measure Π', it can be seen that G_1 contains only 2 cut-vertices, while G_2 contains 6. As a result, $\Pi'(G_1)$ is equal to 0.750 and $\Pi'(G_2)$ is equal to 0.250.

This illustration shows that both measures of diversity take into account different aspects of the reroutings which occurred. The reroutings in G_2 are assessed by Π' to have a much lower diversity, as they only occur at the end of the trajectory. However, G_2 is allocated a higher diversity score by $ECyM'$, as the reroutings in the lower part of the graph are judged to be *broader* then those in G_1. It can indeed be observed that G_1 is more structured, whereas the lower part of G_2 is more dense.

In order to further illustrate the meaning of $ECyM'$ and Π', Figure 3 shows graphs with combinations of low and high values for both metrics. The x-axis depicts the level of horizontal diversity while the y-axis depicts the level of vertical diversity. In the upper right graph, reroutings are wide and well spread along the route, resulting in high values for both the metrics. Meanwhile, in both lower graphs, reroutings are not spread along the whole route, yielding a low value for separability. The graphs in the right part of the table are relatively wide, leading to a high value for the $ECyM'$ metric.

After having identified the instances which are the most sensitive to rerouting, their values for the diversity metrics can be computed. Consequently plotting them on a xy-scatterplot allows the data analyst to map the different instances to the different types of graphs in Figure 3. As such, one can have a

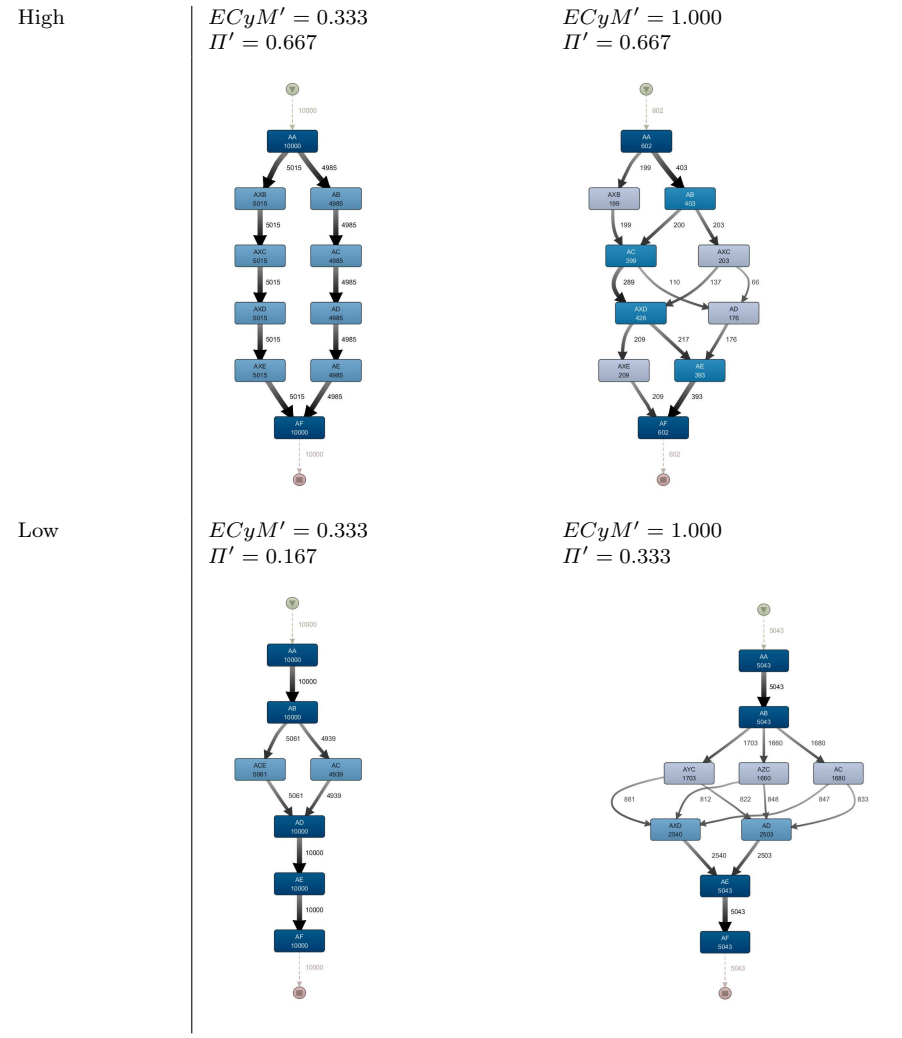

Fig. 3 Typical graphs for low and high values of the diversity metrics.

preliminary idea of how the different graphs look like, without having to look at each of them individually. The analyst can then decide which instances are the most interesting to inspect further.

3.3 Discovering patterns

So far, the methods and metrics proposed are able to both identify which groups of trains are the most sensitive to reroutings and to map different

groups of trains to different types of graphs. Finally, the question might be asked which patterns can be found in the reroutings? In other words, under what circumstances are certain reroutings occurring? For instance, do specific types of deviations always occur at the same time of day?

When the diversity is low, e.g. like in the lower left graph in Figure 3, it is very easy to see which reroutings occur when, since there are only a limited number of distinct reroutings. However, when moving to the upper right graph in Figure 3, distinguishing the different types of reroutings gets more difficult. However, using clustering techniques, reroutings can be grouped into different clusters of similar instances. This can be done using a hierarchical clustering design, in which the distance between two routes is measured using the Sequence Alignment Method (SAM) (Hay et al, 2004). The SAM-measure calculates the difference between two sequences based on the number of insertions and deletions that have to be performed on one sequence, in order to be equal to the other sequence. An hierarchical clustering can be conducted using *average linkage*, where the number of clusters can be decided for each clustering by inspecting the dendogram. The clusters can subsequently be compared to each other along different characteristics: the time of day, day of week, type of rolling stock, etc. This will yield a first understanding as to when and why certain deviations occur. In the next paragraphs, the discussed methods will be illustrated using data from the Belgian train describer system, provided by Infrastructure Manager Infrabel.

4 Application: Belgian train describer data

In the context of the Belgian railway network, the need to optimize capacity usage is amplified by several factors. Firstly, the railway network has a high density, containing many bifurcations within short distances from each other, and it is star-shaped with Brussels at its gravity centre. At a daily basis, 57% of the railway passengers travels to or through Brussels. Secondly, the amount of passengers has risen steadily over the last decades, mounting up to 230 million a year in 2013. Meanwhile, annual punctuality has been decreasing gradually over the last couple of years until 2013. The complexity of the network makes it a non-trivial task to identify the causes of certain delays. As stated in Cule et al (2011), it might not be clear whether a structural delay is the result of ordinary busy traffic or of certain decisions that are made consistently by traffic operators who are unaware of its negative impact.

The data selected for the analyses conducted in this paper were recorded in the signal area of Leuven. With on average 32.247 departing passengers each day (2014), the station of Leuven is the 6^{th} most important railway station in Belgium. Furthermore, the signal area constitutes an important *gateway* to Brussels, and is also responsible for all trains to and from the national airport. The data were recorded during the period from 15 December 2013 to 15 March 2014. The logbook used for the analyses consists of three main categories of events: train movements, user commands and auxiliary functions. The train

movement events were used to reconstruct the actual trajectory of the train. On the other hand, specific message events conveyed the planned trajectory of a train.

A total amount of 5.36 million train describer events were recorded during the three month period. Together, these records describe the history of 75 382 trains trips. On average, circa 950 train trips through the signal area were recorded on a working day, and approximately 600 train trips on a typical weekend day. Given this abundance of available data, there is a clear need for scalable methods in order to produce useful insights about the railway operations. The actual train trajectories covered a total number of 394 different signals and approximately 160 track segments. Note that this area only covers a small percentage of the total railway infrastructure in Belgium. Nevertheless, since the used methods require no a priori knowledge of the infrastructure, scalability of the approach is a clear advantage.

Different types of train movement recordings, related to signals on the one hand and to track segments on the other, were transformed into one standardized format. These events constitute the building blocks of the actual train routes. Table 3 shows the events which are related to train number 1234 on the 10^{th} of January 2014[3]. Each train trip is considered as one *instance* or *case* of the process. Each case is identified by the date and the train number. Each row in Table 3 is an event, which has both a timestamp and a location attached to it. The location may refer to both a signal, which is a combination of letters, or a track segment, which is a number. Recall that only the destination segment is taken into account.

Table 3 Trajectory of train 1234 on January 10th, 2014.

Date	Train number	Timestamp	Location
2014-01-10	1234	6:23:17	AB
2014-01-10	1234	6:24:15	AC
2014-01-10	1234	6:25:49	AD
2014-01-10	1234	6:27:02	100

Next to the actual routes, the planned routes are extracted from specific communication messages. These messages deliver the planned trajectory to the traffic control system as the train approaches the area. Table 4 shows this record for the corresponding train. The message column contains the original encapsulation of the planned route, while the last column shows the route after the extraction and cleaning. This route was send to the signal box at 6:14am, about 10 minutes before the arrival of the train in the area.

The analysis of rerouting severity and diversity can be conducted at different levels of abstraction. The rerouting severity can be calculated at the level of a planned route, at the level of a connection, or at the level of relation. A relation contains all trains between two specific locations, in either direction.

[3] Both train numbers and signals have been anonymised.

Table 4 Example extraction of the planned trajectory.

Date	Train number	Time	Message	Planned trajectory
2014-01-10	1234	6:14:36	2E1234 :AB *>> AC *>> AD 20K *>> 100 AE	AB,AC,AD,100

Each relation can be further divided into two connections, by taking into account the direction of the train. For each connection, one or more planned routes might exists.

The selection of the appropriate abstraction level encompasses a certain trade-off. Focussing on a low level, i.e. planned route, will yield very precise results but there can still be an abundance as many planned routes might exist. Focussing at the higher level of relations will limit the number of instances, but might create the risk that certain problem cases remain hidden. Indeed, when a relation consists of 10 planned routes, of which one has an extremely high severity of reroutings, while the other 9 hardly contain reroutings, the problematic route will probably remain unnoticed in a high-level analysis. A recommended approach would be to start the analysis at a high level, and subsequently lowering the unit of analysis, while at each step discarding the most uninteresting cases from the analysis.

The planned route was extracted for all regular trains, excluding empty train rides, freight trains, and working trains. This resulted in the selection of 58042 train trips. Consequently, these were grouped based on their planned route. For each group, a set of train describer records was constructed containing the actual route. In order to make sure the results of the analysis were reliable, only those groups which contained at least 50 instances were considered. The resulting selection contained 54635, i.e. 94.13%, of all regular trains. Among these trains, 7.75% contained reroutings. For each planned route, a corresponding model was constructed. Both the model and the actual trajectories are the main input to the analysis conducted in the next section.

A total number of 109 different planned routes were considered. They were categorized along 22 relations. The relations considered are listed in Table 5 and schematically visualized in Figure 4. For some pairs of locations multiple relations exists, which are distinguished by certain intermediate points[4]. Note that in this paper only the high-level route is used to distinguish train trips, and not their stops or the type of the train (interregional trains vs intercity, etc.). In the remainder of the analysis, the specific connections are treated anonymously. Next to the 109 different planned routes, 590 different actual routes were found. Thus, for each planned route, on average 5.73 reroutings existed, with a minimum of zero (no reroutings) and a maximum of 29. The

[4] Notice that some of the waypoints indicated in Table 5 are not visually distinguished in Figure 4, since they are very local in nature, most commonly in the dense corridor between the National Airport and Brussels.

length of the routes varied between 2 and 23 signals, with an average of 8 signals.

Table 5 Train connections considered in the analysis.

Relation			Number of trains
National Airport	$\leftrightarrow^{d}$	Brussel	6301
Mechelen	$\leftrightarrow$	Leuven	5980
Luik	$\leftrightarrow^{c}$	Brussel	5562
Aarschot	$\leftrightarrow$	Leuven	5418
Leuven	$\leftrightarrow$	Brussel	4841
Hasselt	$\leftrightarrow$	Brussel	3108
Luik	$\leftrightarrow$	Brussel	2657
Mechelen	$\leftrightarrow$	Brussel	2246
Aarschot	$\leftrightarrow$	Brussel	1982
Landen	$\leftrightarrow^{b}$	Mechelen	1937
Mechelen	$\leftrightarrow^{c}$	National airport	1897
Leuven	$\leftrightarrow^{b}$	Brussel	1747
National airport	$\leftrightarrow^{a}$	Brussel	1301
Luik	$\leftrightarrow$	Landen	1039
Leuven	$\leftrightarrow^{b}$	Brussel	872
Leuven	$\leftrightarrow$	National airport	465
Aarschot	$\leftrightarrow^{a}$	Brussel	461
Waver	$\leftrightarrow$	Leuven	457
Landen	$\leftrightarrow$	Aarschot	370
Landen	$\leftrightarrow$	Brussel	169
Hasselt	$\leftrightarrow$	Landen	117
Mechelen	$\leftrightarrow^{b}$	Leuven	114

[a] Via fast track [b] Via National airport [c] Via high speed line [d] Via default track

Table 6 shows some statistics for the rerouting severity and diversity metrics for the different relations, which are visualised in Figure 5. It can be observed that on average, the rerouting severity of the different relations is quite low, with an average fitness-value of 0.984. By comparing the mean and the median, it can be observed that the distribution is left-skewed, with the mass of the observations in the close vicinity of 1. As such, most relations only contain a limited number of reroutings, while some unfavourable outliers exist.

The values for $ECyM'$ are located in the range from 0.181 to 1.761, with a mean of 0.601. It is not unsurprisingly to find diversity levels to be low for

Table 6 Measures of locality and spread for the deviation severity and diversity measures.

	Deviation severity	Deviation diversity	
	Fitness	$ECyM'$	Π'
Min	0.878	0.181	0.085
Mean	0.984	0.601	0.527
Median	0.994	0.508	0.631
Max	0.999	1.761	0.869
Std. Dev.	0.032	0.405	0.279

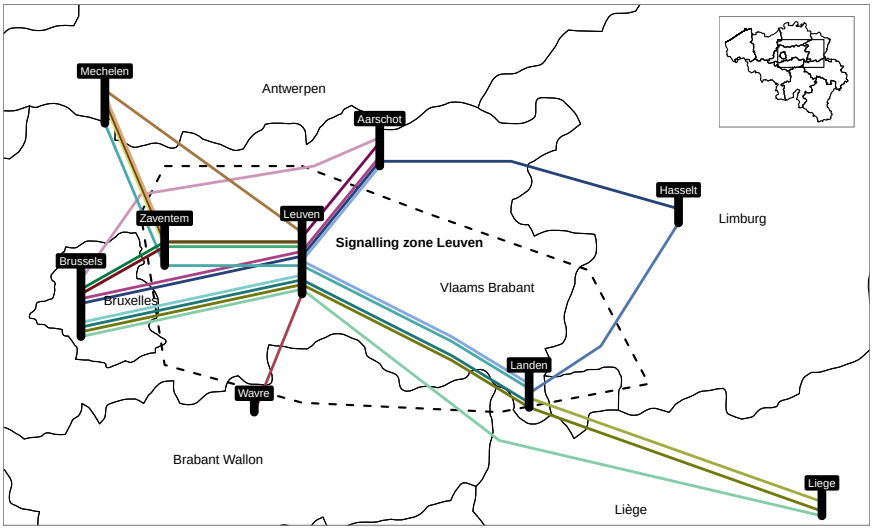

Fig. 4 Schematic overview of considered train relations.

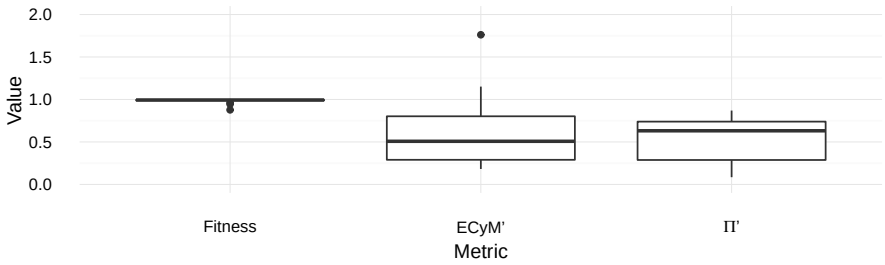

Fig. 5 Boxplots showing the distribution of the metrics.

the majority of the connections, as they are dependent on the extent that reroutings have occurred on these connections. The values for separability are distributed between 0.085 and 0.869. On average, 52.7% of the planned signals is deviated from by at least one train.

The pairwise correlation coefficients between the different metrics are shown in Table 7. It can be seen that, like expected, a negative correlation is found between $fitness$ on the one hand, and $ECyM'$ and Π' on the other hand. As such, when fitness decreases, i.e. rerouting severity increases, the rerouting diversity increases. However, the correlation between $fitness$ and Π' was not found to be statistically significant. Finally, both measures for diversity were found to be significantly positively correlated, which seems legitimate.

Table 7 Pairwise correlations between deviations severity and diversity measures.

	Fitness	$ECyM'$	Π'
Fitness			
$ECyM'$	-0.526***		
Π'	-0.191	0.614***	

*** $p < 0.001$

4.1 Rerouting severity

In order to identify relations with a remarkable severity of reroutings, an analysis of variance can be done for the fitness values, to analyse differences between group means. This means that all trains are grouped according to their relation, and for each train the fitness is computed,using Equation (1). However, two of the underlying assumptions for ANOVA were not satisfied (Iversen and Norpoth, 1987): (1) the dependent variable (i.e. fitness) is not normally distributed within each group and (2) the population variances of the fitness values within each group are not equal. For these reasons, the Kruskal-Wallis test, a non-parametric alternative, was used (Kruskal and Wallis, 1952). Since this test is rank-based, it disregards the magnitude of the differences in fitness. The Kruskal-Wallis test has theoretically less power than the parametric ANOVA when the ANOVA's assumptions are met. However, this is not necessarily true when they do not hold (Demšar, 2006).

The test was able to reject the null hypothesis that there were no differences in rerouting severity among the different relations at a 0.001 significance level. Consequently, a post-hoc Nemenyi test was conducted (Nemenyi, 1963), of which the pairwise results are visualized in a heatmap in Figure 6. The bar chart on the right shows the deviation severity for each relation, ordered from best to worst. The matrix on the left demonstrates whether pairs of relations are significantly different from each other in terms of rerouting severity. A pair of relations with a red cell has a statistically significant difference in rerouting severity at the 0.001 significance level. All the pairs with a green cell are not found to be significantly different with regards to the rerouting severity. It can be concluded that relation 2 has a far higher severity to deviations than all the other connections, followed by connection 3 and 8. These connection are thus identified as the main problem cases requiring further analysis.

4.2 Rerouting diversity

Figure 7 shows a scatter plot based on the two diversity metrics $ECyM'$ and Π'. The horizontal and vertical line display the mean of both metrics. The size of the points refers to the rerouting severity; bigger points having a higher severity. Comparing this plot to Figure 3 gives an overall idea of how the graphs containing the actual behaviour within each relation look like. It can thus be observed that graphs like the one in the lower right of Figure 3 do not

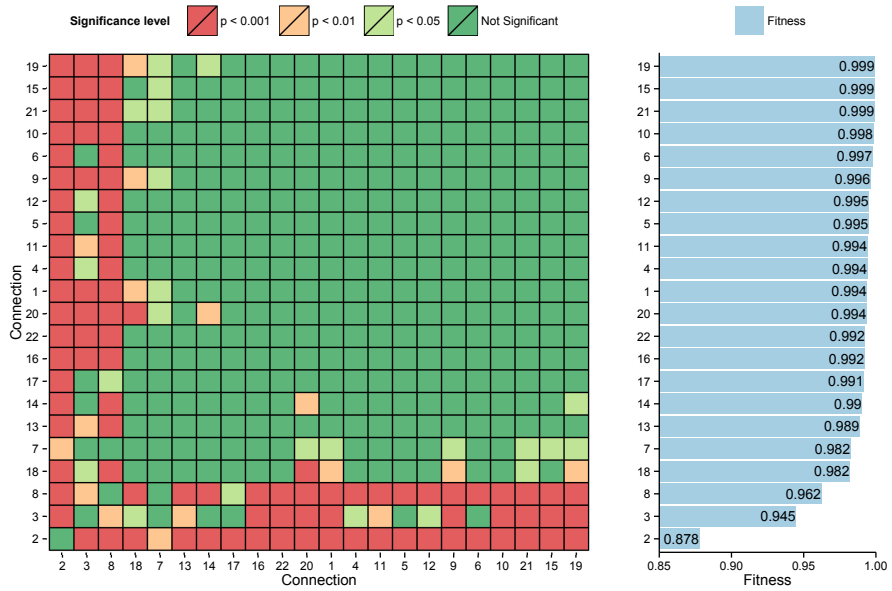

Fig. 6 Heatmap of post hoc Nemenyi test.

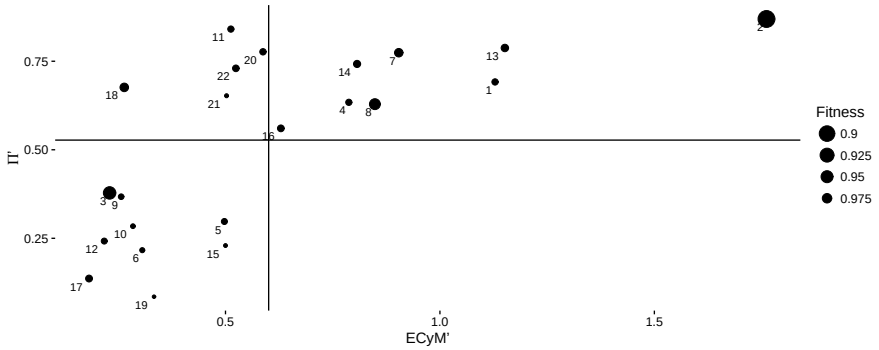

Fig. 7 Scatterplot of rerouting diversity metrics.

seem to occur. Furthermore, the low diversity of reroutings along relation 3 is remarkable in this figure, as it is the second most sensitive to reroutings. As such this will provide a very interesting case, as the low diversity indicates the existence of a limited set of deviations which occur very often. In the remainder of this section, these results will be drilled-down further.

As pointed out before, relations are composed of two connections, one in each direction. In Figure 8 the diversity values are shown for each connection within the selected relations. Connections are distinguished with the letters *A* and *B*. This shows that the two diversity metrics are not always in agree-

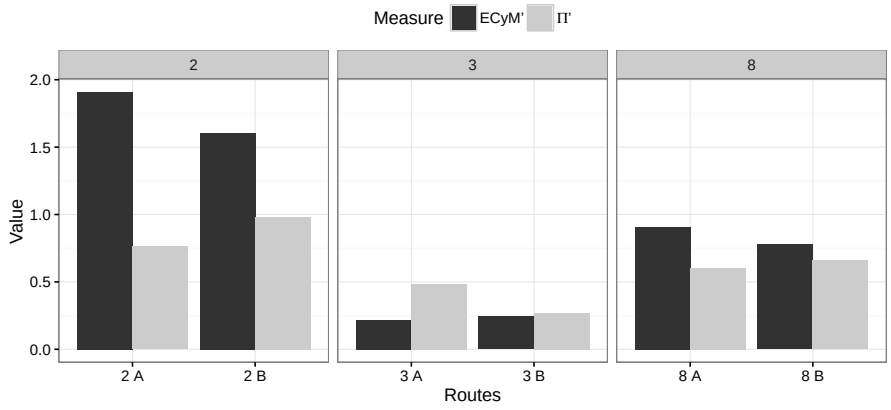

Fig. 8 Diversity of selected relations.

ment with each other within a relation, especially relation connection 2. E.g. Connection 2A has a higher diversity than 2B according to $ECyM'$, but lower according to Π'. As such, reroutings on 2A are expected to differ more in their *width* but are slightly more concentrated along the route, compared to 2B. Furthermore, it can be seen that both connections of relation 3 have a relatively low diversity.

Figure 9 contains two graphs of actual routes, one belonging to each direction of relation 2, each having a similar level of rerouting severity. The graph on the left, belonging to direction A is indeed wider than the graph belonging to direction B. However, the right graph displays reroutings on every signal, while in the left graph the first three signals are never deviated from. It is clear that both graphs fall into the upper right category of Figure 3, having both a relatively high $ECyM'$ and Π'.

In the right graph, some reroutings appear to be relatively systematic. For example OYD > MYD > CYE is taken 8.5% of the cases. Definitely, an in depth analysis should be performed to reveal when and why this rerouting occurs.

Analogously, Figure 10 shows graphs of two routes belonging to connection 3, one in both directions. As was apparent from Figure 7, relation 3 has a very low diversity of reroutings. Indeed, it can be observed that in both directions only one single rerouting has occurred, albeit relatively often. It therefore corresponds to the lower left category in Figure 3. The above-average rerouting severity in accordance with a low deviation diversity yields some interesting inquiries: are there any patterns in the occurrence of this deviation? Why does it occur so often? And were the occurrences beneficial for the operations?

The first question can be easily answered by looking at the data. For instance, it could be observed in the data that about 70% of the reroutings in the left graph in Figure 10 took place at six in the morning. The reason for the deviation can be discovered in different ways. Firstly, one could focus on

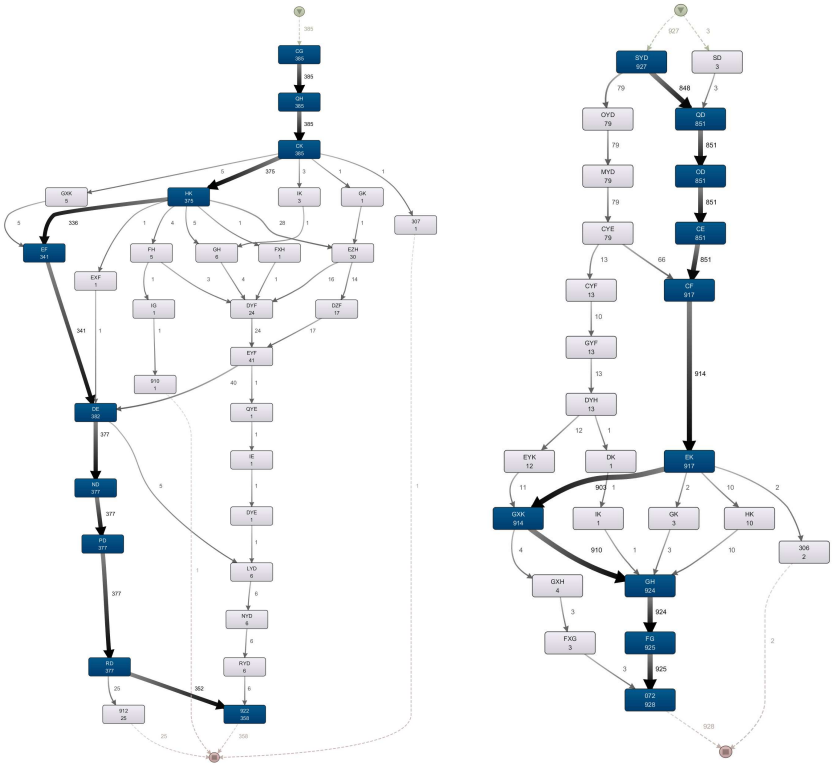

Fig. 9 Actual routes on relation 2 along direction A (left) and direction B (right). Only the most frequent planned route for each direction is selected.

the detailed infrastructure at the location of the rerouting and *simulate* the movement of the trains in this area at the time of rerouting. As such, replaying history can give insights about why certain decisions were taken. Secondly, observations and interviews at the signal box can be clarifying.

The last question, whether the rerouting was beneficial for the overall performance of the network, is much more harder to assess. It involves the linking of reroutings with each other and with impacts on performance measures, such as train punctuality.

Finally, a closer look will be given to relation 8. Just as relation 3, it has an above-average severity to rerouting. While the diversity of reroutings was still rather low, there did not seem to be only a single rerouting. For instance, along one of the planned routes, still 10 different deviations occurred. Nevertheless, relating rerouting to specific characteristics of both train and time can still be meaningful. In order to do so, all rerouting along the planned routes underlying the relation were clustered.

On the routes of connection *8A*, four different clusters were found. For simplicity's sake, the precise composition of the clusters is abstracted from. The

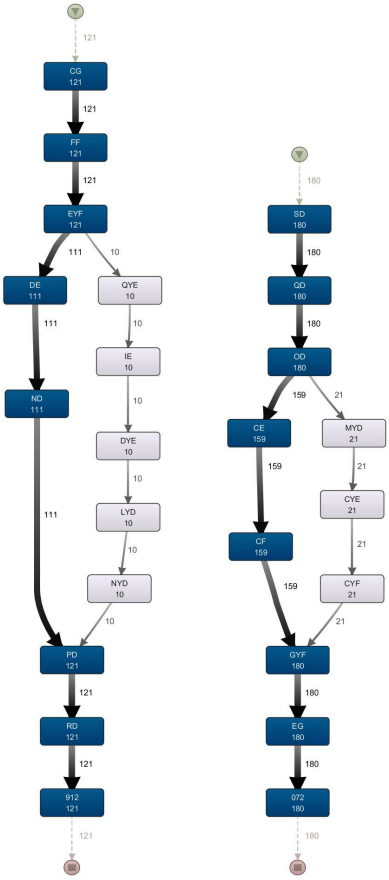

Fig. 10 Actual routes on relation 3 along direction A (left) and direction B (right). Only the most frequent planned route for each direction is selected.

distribution of the clusters over the timespan of a day is shown in Figure 11. It can be observed that reroutings belonging to cluster 3 are more likely to occur in the evening, while reroutings from cluster 0 are more likely to occur in the early morning. It could be further investigated why these reroutings occurred a their specific moments, by replaying history and interviewing business experts, and how they influenced the network operations.

5 Conclusions and Further Research

This paper proposed and illustrated a set of metrics and methods which can be used as a guide for an exploratory analysis of train reroutings, using train describer data. The techniques suggested are able to highlight interesting cases

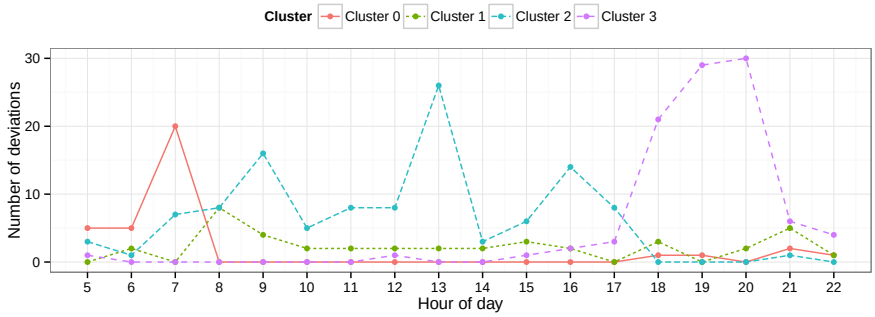

Fig. 11 Distribution of clusters of reroutings on connection 8A over the timespan of a day.

and to point out various paths to conduct further analysis. To this end, measures used in process mining and process modelling were applied to quantify the severity of train reroutings, entitled *reroutings severity*, as well as the variation of the reroutings which occurred, referred to as *reroutings diversity*. The analysis was centred around different train relations. A Kruskal-Wallis test was able to detect differences in the severity to reroutings among the different train relations. Subsequently, inspection of the most remarkable connections validated the correct assessment of the proposed metrics.

The results of these analyses provide a basis for potential improvements of the capacity allocation. Nonetheless, closer investigation by business experts is needed in order to decide whether the reroutings have been beneficial for the overall performance or not. As a first step in understanding the detected reroutings, a cluster analysis has been suggested. By clustering similar deviations into different groups, patterns can be found in their occurrences.

The main advantage of the techniques used in this paper is that they are independent of the underlying railway infrastructure. As the infrastructure is not required as input, the techniques can be easily reused on new cases. Moreover, this allows the metrics to be used on every sort of infrastructure, whereas many existing algorithms are typically limited to a certain set of infrastructure characteristics.

Notwithstanding their proper functioning, some improvements to the metrics can still be made. One would be to allocate costs to the different signals, as a means to make certain reroutings more severe than others. These costs can be determined based on expert knowledge, thereby implicitly requiring input about the infrastructure. Alternatively, costs can be determined based on the data. For instance, signals which are located in an area with a lot of traffic might get a higher cost attributed to it, as reroutings in these areas might have more far-reaching consequences.

Another improvement might be needed in order to accommodate the $ECyM$ metric with a proper scale. In order to scale the metric between 0 and 1, an upper bound needs to be calculated. This upper bound can be de-

termined by looking at the infrastructure, i.e. what would be the maximum number of nodes n and edges e when all possible reroutings would have occurred. When the information on the infrastructure is not provided, these numbers can be estimated by looking at all the behaviour which has occurred, on the condition that data is recorded over an sufficient amount of time.

References

van der Aalst WMP (2011) Process mining: discovery, conformance and enhancement of business processes. Springer, Heidelberg

van der Aalst WMP, Adriansyah A, van Dongen B (2012) Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(2):182–192

Agrawal R, Gunopulos D, Leymann F (1998) Mining process models from workflow logs. In: Schek HJ, Saltor F, Ramos I, Alonso G (eds) Advances in Database Technology - EDBT '98, Springer-Verlag Berlin Heidelberg, vol 1377, pp 467–483

vanden Broucke SKLM, De Weerdt J, Vanthienen J, Baesens B (2013) A Comprehensive Benchmarking Framework (CoBeFra) for conformance analysis between procedural process models and event logs in ProM. In: Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on, IEEE, pp 254–261

Carey M, Kwieciski A (1994) Stochastic approximation to the effects of headways on knock-on delays of trains. Transportation Research Part B: Methodological 28(4):251–267

Conte C, Schöbel A (2008) Identifying dependencies among delays. PhD thesis, Niedersächsische Staats-und Universitätsbibliothek Göttingen, Germany

Cook JE, Wolf AL (1999) Software process validation: quantitatively measuring the correspondence of a process to a model. ACM Transactions on Software Engineering and Methodology (TOSEM) 8(2):147–176

Cule B, Goethals B, Tassenoy S, Verboven S (2011) Mining train delays. In: Advances in Intelligent Data Analysis X, Springer, pp 113–124

Daamen W, Goverde RMP, Hansen IA (2008) Non-discriminatory automatic registration of knock-on train delays. Networks and Spatial Economics 9(1):47–61

D'Ariano A (2008) Improving real-time train dispatching: models, algorithms and applications. T2008/6, Netherlands TRAIL Research School

De Weerdt J, De Backer M, Vanthienen J, Baesens B (2012) A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs. Information Systems 37(7):654–676

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7:1–30

Dewilde T (2014) Improving the robustness of a railway system in large and complex station areas. PhD thesis, KULeuven, Belgium

Flier H, Gelashvili R, Graffagnino T, Nunkesser M (2009) Mining railway delay dependencies in large-scale real-world delay data. In: Robust and online large-scale optimization, Springer, pp 354–368

Goverde RMP, Meng L (2011) Advanced monitoring and management information of railway operations. Journal of Rail Transport Planning & Management 1(2):69–79

Hay B, Wets G, Vanhoof K (2004) Mining navigation patterns using a sequence alignment method. Knowledge and information systems 6(2):150–163

Higgins A, Kozan E (1998) Modeling train delays in urban networks. Transportation Science 32(4):346–357

Iversen GR, Norpoth H (1987) Analysis of variance. 1, Sage

Kecman P, Goverde RMP (2015a) Online data-driven adaptive prediction of train event times. IEEE Transactions on Intelligent Transportation Systems 16(1):465–474

Kecman P, Goverde RMP (2015b) Predictive modelling of running and dwell times in railway traffic. Public Transport 7(3):295–319

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 47(260):583–621

McCabe TJ (1976) A complexity measure. Software Engineering, IEEE Transactions on 2(4):308–320

Mendling J, Neumann G, Van Der Aalst W (2007) Understanding the occurrence of errors in process models based on metrics. In: On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS, Springer, pp 113–130

Nemenyi PB (1963) Distribution-free multiple comparisons. PhD thesis, Princeton University

Rozinat A, De Medeiros AKA, Günther CW, Weijters AJMM, van der Aalst WMP (2007) Towards an evaluation framework for process mining algorithms. Beta, Research School for Operations Management and Logistics

Sammouri W (2014) Data mining of temporal sequences for the prediction of infrequent failure events: application on floating train data for predictive maintenance. PhD thesis, Université Paris-Est

Törnquist J (2006) Computer-based decision support for railway traffic scheduling and dispatching: A review of models and algorithms. In: 5th Workshop on Algorithmic Methods and Models for Optimization of Railways, p 659

Watson AH, McCabe TJ, Wallace DR (1996) Structured testing: A testing methodology using the cyclomatic complexity metric. NIST special Publication 500(235):1–114

Weeda VA, Hofstra KS (2008) Performance analysis: improving the dutch railway service. J Allen et al, Proceedings Computers in Railways XI pp 463–471

Yuan J, Hansen IA (2007) Optimizing capacity utilization of stations by estimating knock-on train delays. Transportation Research Part B: Methodological 41(2):202–217