Validity of railway microscopic simulations under the microscope

Two case studies

Roungas, Bill; Meijer, Sebastiaan; Verbraeck, Alexander

DOI
10.1504/IJSSE.2018.094562

Publication date
2018

Document Version
Final published version

Published in
International Journal of System of Systems Engineering

Validity of Railway Microscopic Simulations Under the Microscope: Two Case Studies

Bill Roungas*

Department of Multi Actor Systems,
Delft University of Technology,
Delft, The Netherlands
E-mail: v.roungas@tudelft.nl
*Corresponding author

Sebastiaan Meijer

Department of Health Systems Engineering,
KTH Royal Institute of Technology,
Stockholm, Sweden
E-mail: sebastiaan.meijer@sth.kth.se

Alexander Verbraeck

Department of Multi Actor Systems,
Delft University of Technology,
Delft, The Netherlands
E-mail: a.verbraeck@tudelft.nl

Abstract: Simulations are the core of every railway system. Changes in the timetable and the infrastructure, or even in the internal processes of a railway company should be, and usually are, first tested through simulations. Given their significance and potential impact, simulations should be primarily validated; validation ensures - at least to some extent - that the returned results are credible and can be used for the intended purpose. This study is a detailed report on two case studies from the railway sector. The aim of this paper is to identify critical factors that can advance or hinder the validity and the effective usage of simulation models.

Keywords: microscopic simulations, validation, railways, case study, system of systems

Biographical notes: Bill Roungas is a PhD researcher in the Department of Multi Actor Systems at Delft University of Technology. His research focuses on the design, validation, and knowledge management of simulations and serious games.

Sebastiaan Meijer is Professor at KTH Royal Institute of Technology, Department of Health Systems Engineering, and at Delft University of Technology, Faculty of Technology, Policy and Management. He is specialised in gaming simulation and other interactive methods to involve the operational level of organisations in innovation processes. He is interested in the theory of design of complex adaptive systems and the backbones of society.

Alexander Verbraeck is Professor at Delft University of Technology, Department of Multi Actor Systems and part-time professor in supply chain management

at the R.H. Smith School of Business of the University of Maryland. His research focuses on modelling and simulation, especially in heavily distributed environments and using real-time data. Examples of research on these types of simulations are real-time decision making, interactive gaming using simulations, and the use of 3D virtual and augmented reality environments in simulations. The major application domain for research is logistics and transportation.

1 Introduction

A nation-wide railway network is not just the sum of its individual components, but should, instead, be seen as a Complex Adaptive System (CAS) (Rinaldi et al. 2001). Upon analysing such a system, it becomes evident that it consists of multiple socio-technical systems (Trist & Bamforth 1951) (trains with drivers, infrastructure with maintenance engineers and train traffic controllers etc.) and several independent actors (passengers, politicians etc.), which adapt their behaviour based on internal and external stimuli and form complex and emergent relationships with each other. At the same time, a railway network can also be seen as a System of Systems (SoS) despite the lack of a definitive definition of SoS. It has been shown that railway networks demonstrate the behaviour of what experts describe as SoS. According to De Laurentis' (2005) work, they seem to adequately satisfy the distinguishing traits of SoS: operational & managerial independence, geographic distribution, evolutionary behaviour, emergent behaviour, to name a few.

On the other hand, by definition, simulations are the imitation of the operations of a real-world process or system over time (Banks et al. 1984), and as such they are an abstraction, or simplification, of the respective process or system. Despite their abstractive nature, simulations are perhaps the best way that systems characterized as SoS can be understood and tested in an affordable, risk-free, and ethical way (Zeigler & Sarjoughian 2012). These three terms, i.e. affordable, risk-free, and ethical, are the *holy grail* of most railway companies. Even a small change in a railway infrastructure can cost several millions. Moreover, it bears significant risks both in terms of construction (e.g. wrong materials, mistakenly positioned switch etc.) and operation (e.g. not alleviating the load on the network, interfere with the normal operations etc.), whereas their mitigation further increases the cost. Finally, since such a system is used by hundred of thousands or even millions of people on a daily basis, the extent to which a railway company exhausts all possible solutions, in order to provide the best possible service, becomes an ethical issue. This is a small example of the complexity of just one decision. But not only changes in the physical infrastructure need extensive testing. Changes in the timetable also need testing, in order to ascertain that the railway resources (infrastructure, rolling stock etc.) successfully accommodate these changes, that any unexpected situation are dealt with in the best possible way, and that the probability the service will become unavailable is minimized.

In effect, the use of simulation in the planning and operations of railways has become increasingly popular. The said popularity has not passed unnoticed by the Dutch railway task organization ProRail, which several years ago started building simulations both internally and through the use of third-party packages, and has todate developed a wide range of simulations, extending from microscopic (Yuan & Hansen 2007) and macroscopic (Middelkoop & Bouwman 2001) to gaming simulations (Meijer 2012, 2015). Microscopic railway simulations simulate every aspect of the system in a detailed manner; the train's

motion at any given moment is determined according to dynamic equations and every aspect of the infrastructure is taken into account (Asuka & Komaya 1996). On the other hand, macroscopic railway simulations simulate only the arrival and departure times of trains, and the general characteristics of the infrastructure (Asuka & Komaya 1996). Finally, gaming simulations can be either microscopic or macroscopic and what differentiates them is the human input. Each of these simulations has a different purpose. Whilst macroscopic simulations are time efficient, microscopic simulations are more precise and thus preferred for networks with high speed trains and high density of train traffic Asuka & Komaya (1996). Gaming simulations are used when human input is necessary.

Several of these simulation packages are quite similar both in terms of input and output data, and in terms of their intended purpose. An example of such similarity is FRISO and OpenTrack, which the authors were assigned to validate, and thus form the two case studies examined in the present paper. FRISO is ProRail's in-house simulation environment (Middelkoop & Loeve 2006) whereas OpenTrack is a well-established program developed at the Swiss Federal Institute of Technology's Institute for Transportation Planning and Systems (ETH IVT) (Nash & Huerlimann 2004). Since they are both microscopic simulation environments, FRISO and OpenTrack have the potential to, and depending on the model usually do, simulate the railway network in a detailed manner; both have the ability to depict the network down to a switch level. In view of the pointed similarities, a comparison of one simulation package over the other seems rather inevitable. In this respect, the authors' initial hypothesis is that let alone their similarities, FRISO and OpenTrack are different, thus suitable for different usages. This is a hypothesis that will be accepted or rejected, once the comparison of the two software packages has been concluded.

In this study, two different instantiations of models on both packages, which led in the development of customized tools for their validation, are presented. Through the analysis and the comparison of the two models, and the development of the respective tools for the ensuing validation, this study aims at identifying critical factors that influence the success of simulation models. Therefore, the intention of the comparison is not to decide which simulation package is more valid but to demonstrate the steps that were followed during their validation. Particularly, the comparison of the two packages aims at pinpointing the common practices during the validation of a simulation model. Subsequently, the analysis of these common practices would be of great interest, since they could be considered good candidates for factors that can critically influence the validation study of a simulation model and, as a result, the model itself.

An important distinction should be made between the validation methodologies applied in each one of FRISO's and OpenTrack's models, and the methodology used to accomplish the aim of this paper. The former are two methodologies for validating punctuality and the train driving behaviour, respectively. The latter is built on the steps of this paper, which by definition also includes the two aforementioned methodologies. The purpose of this study is to identify the similarities and differences between the two packages and their subsequent models, pinpoint the impediments during the validation study, and in turn, present the most striking results and the identified critical success factors.

In Section 2, FRISO's and OpenTrack's architecture is described. In Section 3 and Section 4 the detailed models for FRISO and OpenTrack are presented along with their respective validation studies. In Section 5, the results of both models are compared to each other and a conclusion is drawn in regard to the initial aim of the paper. Finally, in Section 6, future steps that can improve the research on this field are outlined and final remarks are made.

2 Simulations' Architecture

In this section, the architecture of both simulation packages is described. The authors did not have access to any of the packages' source code nor to the internal structure of the models; they were only assigned to validate the models. Therefore, the architecture analysis in both cases is limited to the input the simulations require and the output they produce.

The overall scheme of both packages seems to have a remarkable resemblance. Arguably, this is a product of both packages being microscopic simulations, thus requiring more or less the same input and producing the same output. As such, both FRISO and OpenTrack require three basic components as input:

- Timetable, which for an experimental study can be an hourly pattern timetable whereas for a more operationally oriented study it should be as close as possible to the actual timetable.

- Rolling Stock, which includes all the different trains along with their technical characteristics.

- Infrastructure, which for a microscopic simulation means not just the railway tracks but also switches, signals, and any other detail that influences train operations.

The visualization features of both packages are quite similar in that they both offer animation and interactive capabilities. Finally, the output is equally similar for both, with OpenTrack having a few more features when it comes to graphs and tables. Particularly, OpenTrack offers two additional options: a. rail occupation statistics followed by occupation diagrams, and b. a train power and energy consumption output. A depiction of OpenTrack's main elements is shown in Figure 1. What differentiates FRISO from OpenTrack, and it is the main reason ProRail built it in the first place, is its ability to be incorporated in a High Level Architecture (HLA) scheme (Kuhl et al. 1999), which allows the interaction with other computer simulations regardless of the computing platforms, and provides for a wide application in gaming simulations (Middelkoop et al. 2012).

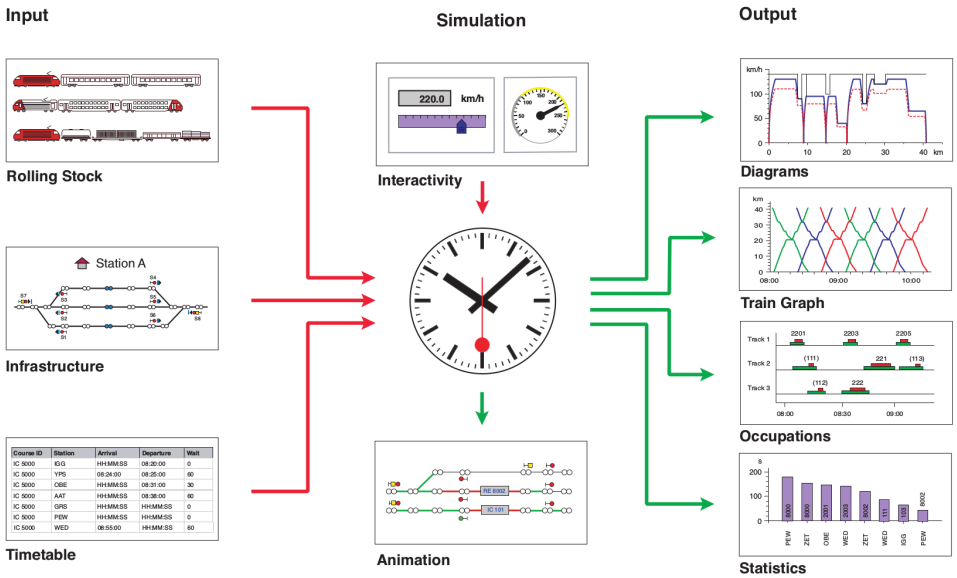


Figure 1: Main elements of OpenTrack (Nash & Huerlimann 2004).

In the next two sections, the input, output, and the validation study of two models instantiated in FRISO and OpenTrack are described.

3 The FRISO Model

In this section, the model used in FRISO (hereinafter references to FRISO and OpenTrack will be references to their respective models, unless otherwise stated) and its subsequent validation is described in detail. The model was built in 2014 and it simulates the train operations in one of the most heavily utilized sections (Amsterdam Centraal - Utrecht Centraal) of one of the largest corridors in the Netherlands (A2), during the whole month of June 2013. The intended use of the model was to examine the punctuality of the timetable with the particular focus being the Amsterdam and Utrecht central stations. The input elements of the model were the:

- Timetable, which was the theoretical timetable for the month of June 2013. The term theoretical is used in order to denote that the actual timetable slightly changes every day, so as to accommodate urgent and unplanned events, e.g. an unplanned freight train, heavy weather conditions etc.
- Rolling Stock, which was the exact rolling stock operating between Amsterdam and Utrecht central stations in June 2013. This included the major train series 120 (Operator: DB, Type: ICE), 3000 (Operator: NS, Type: Intercity), 4000 (Operator: NS, Type: Sprinter), 7400 (Operator: NS, Type: Sprinter), 800 (Operator: NS, Type: Intercity), and the minor train series 47700, 48700, 77400.
- Infrastructure, which was the exact infrastructure (tracks, traffic signals, switches etc.) between Amsterdam and Utrecht central stations in June 2013.

Moreover, several variations of the Gamma, Normal, and Negative Exponential distributions were used for the delays in arrivals.

The analysis of the output of the model required extensive data cleaning for both the model output data and the operational data. In more detail, the data cleaning comprised of queries that:

- Deleted several columns from both datasets that were not relevant for the study (like columns with specific codes used in planning),
- Deleted rows containing regions and train series that were not common on both datasets (ensuring that both datasets were of the exact same region and containing the same train series), and
- Deleted from the realization data all rows with unrecorded arrival or departure time.

The data cleaning resulted in a reduction of the number of variables (i.e. columns) for both datasets and of their overall size down to half, which significantly reduced the execution times of the queries. Finally, renaming certain columns became necessary, in order to allow the customized validation tool to automatically calculate the various statistical tests and produce the necessary graphs.

As a whole, the model shows remarkable precision by being only two seconds off in estimating the average delay in arrivals. Even at a station level, for all stations, the difference in the average arrival delay between the model and reality is less than 30 seconds (Table 1). Despite this rigour in a macroscopic and in a station level, there are three striking observations:

Train	All	120	3000	4000	47700	48700	7400	77400	800
Delay	-2.04	19.82	24.53	-20.66	-99.89	-101.98	-9.73	-65.31	27.4

Station	All	Ut	Mas	Dvd	Asa	Asdm	Asdma	Asd
Delay	-2.04	17.05	20.97	-19.99	9.94	1.9	-1.3	-3.34

Table 1 Difference in delays in arrivals at a train and at a station level between FRISO and reality. **Abbreviations:** Ut: Utrecht Centraal, Mas: Maarssen, Dvd: Duivendrecht, Asa: Amsterdam Amstel, Asdm: Amsterdam Muiderpoort, Asdma: Amsterdam Muiderpoort aansluiting (passing point), Asd: Amsterdam Centraal

1. At a train level, the behaviour of the model does not appear to be consistent. The major train series (120, 3000, 4000, 7400, and 800) exhibit relatively good behaviour (difference between reality and model is less then 30 seconds), whereas the minor train series (47700 and 48700) seem to experience big delays throughout the whole route from Amsterdam to Utrecht and vice versa. This is due to the minor series accounting for less than 5% of the total traffic in this particular route, resulting in modellers focusing mainly on the major train series. The train series 77400 is not an actual independent series but the 7400 series with an addition of a 7 in front, in order to indicate that the train performs shunting movements, i.e. parking or sorting the rolling stock. Therefore, given the significant influence of the major train series in the model, and the rather insignificant impact of the minor train series, any negative effect resulting from the latter is diluted.

2. Due to the nature of FRISO's output data, i.e., the availability of the arrival and departure times alone, the only visualization that could show the richness of the data is a histogram of delays, either at a station (Table 2) or at a train level. The histograms were built by creating 20-second intervals (bins) of the delays occurred in the model and in reality. The large number of observations gave histograms a fine granularity and allowed for concrete conclusions. Hence, the second striking observation is how delays in arrivals are distributed. Both in Amsterdam and in Utrecht central stations, the operational delays seem to follow a right skewed distribution, while in the model, the delays seem to follow a bimodal distribution, although slightly less sharp for Amsterdam central station. However, since the positive delays (i.e., delays greater than zero), for both cities - and especially in Amsterdam - appear to be distributed similarly in the model and in reality, experts were confident that this discrepancy was not enough to invalidate the model.

3. The third striking observation is that the more a train drives away from Amsterdam central station (Asd), which is the epicentre of the model, the more the precision of the model decreases; and as the train reaches Utrecht central station (Ut), which is the second most important station in the model, the precision of the model slightly rises again. This observation can only be valid if the negative delays (early arrivals) are disregarded. Unless an early arrival causes an indirect delay to another train, which was not observed in this particular case, then it does not cause any negative effect to the system, thus it can be disregarded. Table 3 serves as a heatmap, in which the difference in delays between reality and the model are marked as follows:

   - Under 15 seconds with green.
   - Between 15 and 30 seconds with orange.
   - More than 30 seconds with red.

   The stations' sequence in the table is the same a train follows from Amsterdam to Utrecht and vice versa. The colour distinction used in Table 3 is meant to help observers identify patterns in the delays. Indeed, upon carefully examining Table 3, it becomes easier to notice that the model experiences a ripple effect. The more a train diverges from the main area of focus in the model, i.e., primarily Amsterdam central station and secondarily Utrecht central station, the less the factors that influence the predictability of the model are taken into account and calculated with precision. Nevertheless, similarly to the minor train series, experts concluded that, due to the fact that those cities are outside the main area of focus, their overall impact in the model is insignificant. Hence, this observation was also not enough to invalidate the model.

In this section, FRISO, and its subsequent validation, were described and the most striking observations were presented. All these striking observations are deemed by experts to constitute insufficient evidence for invalidating the model because:

1. The train series accounting for more than 95% of the traffic were modelled with high precision.

2. Disregarding the early arrivals in Friso, which did not appear to cause direct or indirect delays to other trains, resulted in transforming the bimodal distribution of delays in a right skewed distribution very similar to the one of the operational data.

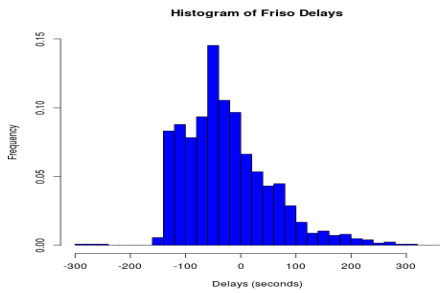
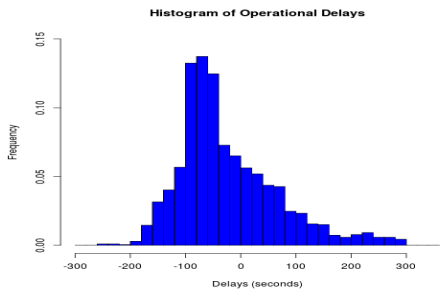
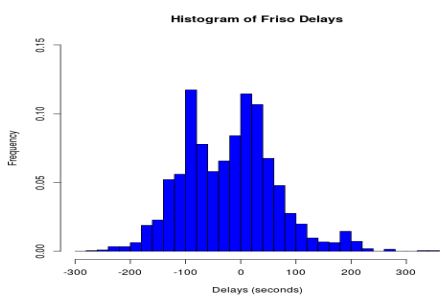
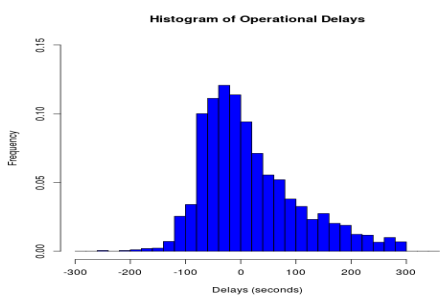
Model	Reality
Amsterdam	
	
Average: -25.38 sec. Standard Deviation: 78.56 # of observations: 1252 Max: 315 sec. Min: -282 sec.	Average: -22.93 sec. Standard Deviation: 90.21 # of observations: 2061 Max: 298 sec. Min: -242 sec.
Utrecht	
	
Average: -19.07 sec. Standard Deviation: 83.9 # of observations: 2070 Max: 345 sec. Min: -263 sec.	Average: 17.6 sec. Standard Deviation: 88.62 # of observations: 3496 Max: 299 sec. Min: -253 sec.

Table 2 Simulated and real delays in Amsterdam and Utrecht central station.

3. Any stations other than Amsterdam and Utrecht central stations were not the focus in this model, hence any inconsistency between the Friso and operational data in these stations was not taken strongly into account.

As a result, experts considered the model to be valid for its intended purpose, which is to test the punctuality in major train stations like the ones in Amsterdam and Utrecht. In addition to the three aforementioned reasons, the model exhibits striking resilience especially in the two stations of focus (Amsterdam and Utrecht). This is evident from the fact that despite the relatively large differences in the off-focus stations, the model adapts and covers a significant amount of these differences, either by accelerating or decelerating. A more in depth analysis and a comparison of FRISO with OpenTrack, which is described in Section 4, takes place in Section 5.

		Stations						
		Ut	**Mas**	**Dvd**	**Asa**	**Asdm**	**Asdma**	**Asd**
Trains	**120**	10.89	24.66	40.86	22.83	15.16	4.31	4.24
	3000	15.12	47.6	38.9	9.24	0	-1.27	-1.21
	4000	-	-	-32.32	-2.43	-0.79	-11.16	-14.73
	47700	29.36	-106.31	-121.72	-52.82	-42.23	-52.15	15.64
	48700	3.16	-209.71	-66.05	-55.83	-41.99	-58.06	-0.95
	7400	8.14	-3.9	-24.22	-0.8	-38.1	-11.26	-14.94
	77400	-	-	16.6	36.2	17.21	-8.86	-22.78
	800	34.57	44.31	40.43	18.17	11.25	9.19	10.64

Table 3 Heatmap with difference in delays in arrivals between FRISO and reality.
Abbreviations: Same as in Table 1

4 The OpenTrack Model

In this section, OpenTrack and its subsequent validation is described in detail. The model was built in 2016 and it simulates the train operations in a heavily utilized section (Eindhoven Centraal - Utrecht Centraal) of one of the largest corridors in the Netherlands (A2), according to the newly designed timetable, which was to start in January 2017. The input elements of the model were the:

- Timetable, which was the newly developed timetable intended to be put in use in January 2017.
- Rolling Stock, which was part of the rolling stock operating between Eindhoven and Utrecht central stations according to the new timetable. This included four train series: 800 (Operator: NS, Type: Intercity), 3500 (Operator: NS, Type: Intercity), 6000 (Operator: NS, Type: Sprinter), and 9600 (Operator: NS, Type: Sprinter).
- Infrastructure, which was the planned infrastructure (tracks, traffic signals, switches etc.) between Eindhoven and Utrecht central stations for January 2017.

Similarly to FRISO, the analysis of OpenTrack's output required data cleaning for both the model and the operational data. But unlike FRISO, with OpenTrack the data cleaning was more about transformation of the data in comparable units (model) and correction or deletion of GPS data that were either distorted or off certain limits (reality).

The initial dataset included four train series running through one of the major corridors of the Netherlands (A2), namely from Utrecht to Eindhoven. Unlike FRISO, in which the intention of the study was to test punctuality, with OpenTrack the purpose was to test the conflicts occurring throughout the timetable by examining the train driving behaviour. Despite this intention, due to the nature of its data, OpenTrack could also provide for a punctuality test by extracting the arrival time of trains. Hence, OpenTrack's validation study was divided in two independent studies: validation of the driving behaviour and validation of the punctuality. Since the models in Friso and OpenTrack depict different instantiations of the railway system, it could be argued that they cannot be compared. Ideally, there should have been a model for each simulation package simulating the exact same scenario, thus allowing the two models to be directly and indisputably compared. Nevertheless, in a commercial setting this is hardly ever possible due to time and budget

restrictions. Companies cannot usually afford such additional costs and, further, do not have the luxury of time to build and run experiments of the same scenario on multiple platforms for testing purposes. This results in situations like the one described in this paper, where a comparison is performed between the different models by incorporating experts' knowledge about the system. In other words, a comparison in such a situation can become fruitful when along with the results of each model, experts provide insights based on their experience that provide context, which in turn mitigates the risk arising from the different instantiations.

4.1 Driving behaviour

The driving behaviour modelled in OpenTrack depends on five parameters: the acceleration rate, the minimum speed, the maximum speed, the breaking (deceleration) rate, and a performance coefficient, and it should be separated into two different categories, namely the actual driving behaviour and the breaking behaviour. The driving behaviour is determined by the acceleration rate, and the minimum and the maximum speeds after all of these have been multiplied for adjustment by the performance coefficient. The breaking behaviour is determined by the breaking rate, which is also adjusted by being multiplied with the performance coefficient. The acceleration rate, the minimum speed, the maximum speed, and the breaking rate are predetermined by the Dutch railway operator ProRail and they are fixed. Therefore, any variation on the driving and breaking behaviour depends on the performance coefficient. The performance coefficient is determined by the modellers through trial and error based on observation from past operational data, in order for the train behaviour to be as realistic as possible. For this particular model, the performance coefficient fluctuates from 97.5% to 100.5% depending on the train type, i.e., fast train (intercity) or slow train (sprinter).

A visualization of the driving and breaking behaviour is shown in Figure 2. The graphs were construed in order to, initially, have all the available information visualized, allowing for the deduction of the necessary information afterwards. The x-axis shows time in minutes and the y-axis shows the starting and stopping station, as well as all intermediate stations trains pass through without stopping. The graphs include four different kinds of line graphs:

- All operational data from February 2017 depicted in a low opacity black colour line, which create a grey shadow that can reveal patterns.

- The three or four percentile lines (10^{th}, 50^{th} (median), 90^{th}, 95^{th} percentile) depending on the number of observations, depicted in a blue colour line. The 95^{th} percentile line is shown only in cases where the sample size of the operational data is more than 100. The percentiles were calculated by creating 10-second intervals (0 to 10 sec., 10 to 20 sec. etc.). Then, all available data points from the GPS data that belonged at each interval were gathered and the position of all trains within each interval was linearly extrapolated to the floor (e.g. getting the position of a train 11 seconds after starting would mean that its position would be linearly extrapolated at 10 seconds, getting the position of a train 23 seconds after starting would mean that its position would be linearly extrapolated at 20 seconds and so forth). Finally, the list of data points was sorted in an ascending order, and the 10^{th}, 50^{th}, 90^{th}, and 95^{th} percentiles were chosen.

- OpenTrack's data obtained in March 2016 depicted in a red colour line.

- OpenTrack's data obtained in March 2016 increased by one minute is depicted in a dotted black colour line. The use of this line aimed at planning and visualizing potential conflicts with trains based on the minimum running time (minimum difference between two consecutive trains that are on the same track) allowed by ProRail.

The breaking behaviour of the model, as shown in all three subfigures in Figure 2, is modelled with extreme precision when compared to the median (50^{th} percentile). On the other hand, regarding the actual driving behaviour, in some cases (Figure 2a) the train drives almost on the exact path of the median (50^{th} percentile), which is the desired driving behaviour, but on some other occasions it either drives along the 90^{th} percentile (Figure 2b), which is the slowest 10%, or exhibits some sort of irrational behaviour (Figure 2c), which upon further examination is caused due to a conflict with another train.

As a result, this fluctuation in the driving behaviour brings up the following question: To what extent can a single driving profile adjusted only by a coefficient simulate a realistic driving behaviour? In order to answer this question, the purpose of the simulation study, and consequently the extent to which this model is valid, should be taken into account. This particular simulation study was intended to test the newly developed timetable against conflicts between trains rather than the driving behaviour per se. Modelling the driving behaviour was a means to an end.

The conflict identified in Figure 2c was observed in 11.1% of the cases in the model and approximately in 10-15% of the cases in reality, depending on the train series. By all means, not every conflict resulted in exactly the same behaviour, neither in reality nor in the model. Nevertheless, the model managed not only to anticipate the possibility of conflicts but also to approximate the probability of the occurrence of these conflicts.

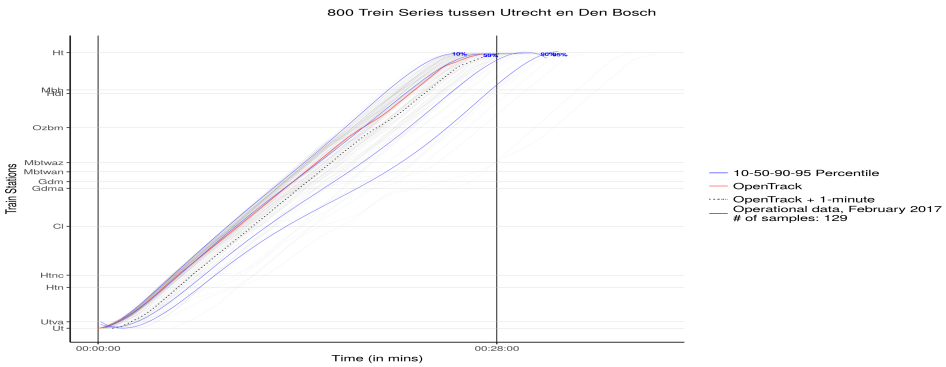
4.2 Punctuality

The richness of OpenTrack's output also provided for a punctuality test, similar to the one performed on FRISO, which allows for a direct comparison of the two models. Unlike FRISO, in which the focus of the punctuality test revolved around the central stations of Amsterdam and Utrecht, the focus in OpenTrack's punctuality test was in the central stations of Eindhoven and Utrecht. OpenTrack provided an output file with all the delays, and the operational delays were calculated by extracting the last value of the GPS data from each sample.

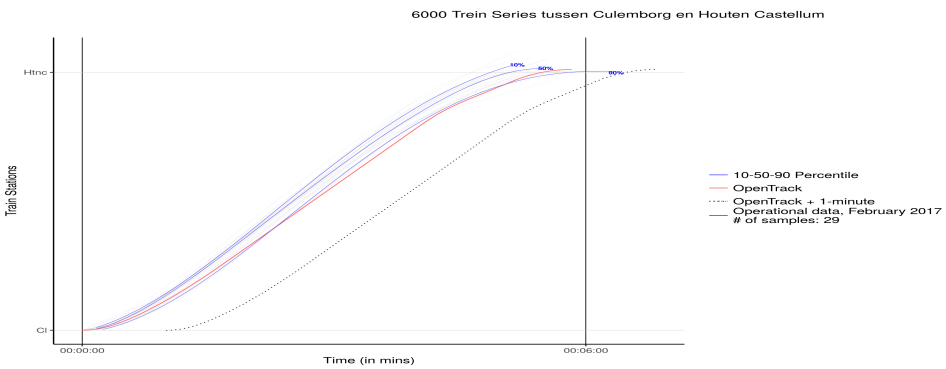
The histograms of delays from the model and reality are shown in Table 4 for Eindhoven and Utrecht central stations. While the average difference in the delays between the model and reality for both cities is approximately 10 seconds, which indicates a good estimation of punctuality, the delays are distributed completely differently between the model and reality.

4.3 Conclusion on OpenTrack

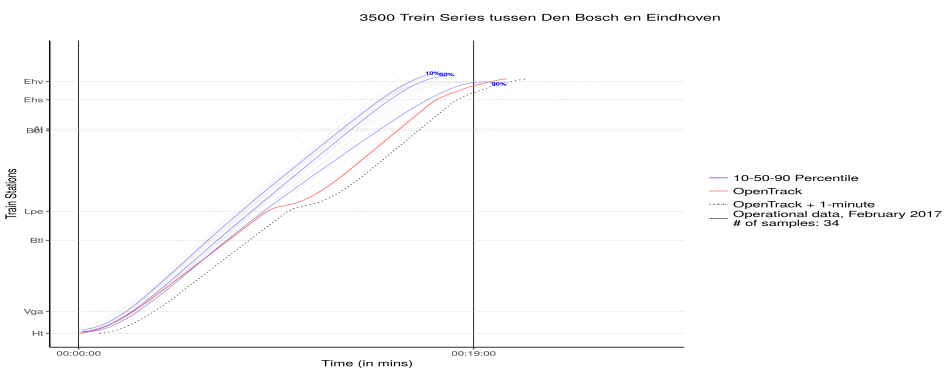
This section focused primarily on OpenTrack and, subsequently, on its validation parameters. The validity of the model was tested with regards to the train driving behaviour and the punctuality of the contemplated timetable. Experts considered the model to be valid for the purpose of simulating the driving behaviour and for identifying, in turn, conflicts between trains. On the contrary, the difference in the distributions of delays has been so significant that the model cannot be considered valid for a punctuality study at a microscopic level.



(a) Train series 800 between Utrecht central station and Den Bosch



(b) Train series 6000 between Culemborg and Houten Castellum



(c) Train series 3500 between Den Bosch and Eindhoven

Figure 2: Driving and breaking behaviour of OpenTrack.

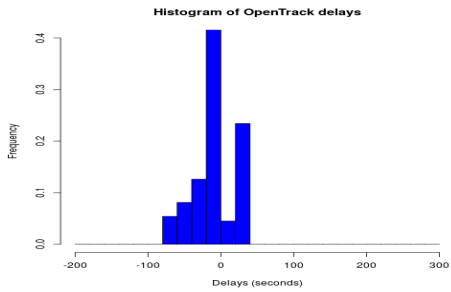
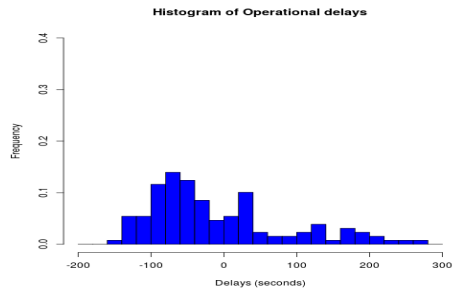
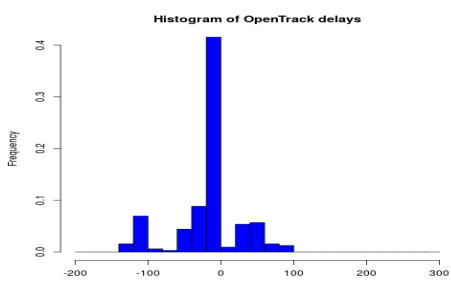
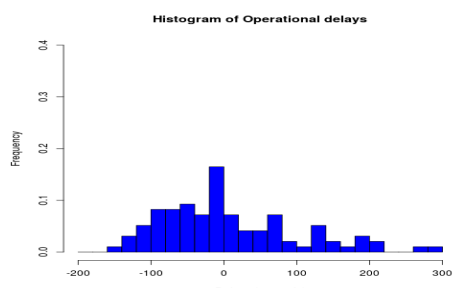
Model	Reality
Eindhoven	
	
Average: -3.73 sec. Standard Deviation: 27.39 # of observations: 111 Max: 36 sec. Min: -68 sec.	Average: -12.85 sec. Standard Deviation: 94.71 # of observations: 129 Max: 262 sec. Min: -159 sec.
Utrecht	
	
Average: -8.74 sec. Standard Deviation: 41.58 # of observations: 316 Max: 96 sec. Min: -136 sec.	Average: 2.85 sec. Standard Deviation: 93.99 # of observations: 97 Max: 282 sec. Min: -149 sec.

Table 4 Simulated and real delays in Eindhoven and Utrecht central station.

5 Discussion

The analysis in Section 3 and Section 4 showed the advantages and disadvantages of FRISO and OpenTrack, which were both deemed by experts to be valid for their intended purpose. As mentioned in Section 1, a comparison between similar simulation packages and their subsequent models is inevitable, but in this particular case, doing so would be a mistake. Comparing FRISO with OpenTrack would not be fruitful due to multiple reasons. First and foremost, despite both being microscopic simulation packages, the intended purpose of each model is different (FRISO: punctuality, OpenTrack: conflict detection).

Moreover, the datasets used in the models and the parameters tweaked within the simulation software have significant differences. The dataset used in FRISO has only the arrival and departure times of trains, whereas in OpenTrack, the exact location of each train is available every two seconds. Additionally, FRISO was built based on the existing infrastructure, rolling stock, and timetable, for which delays were known. On the contrary, OpenTrack was built based on the existing infrastructure and rolling stock, but based on a future timetable for which delays were not known and could only be assumed according to the modellers' knowledge. Finally, the two datasets used in the models were almost four years apart, focused on different cities (with a small overlap in Utrecht), and were based completely different timetables.

Therefore, there are several critical factors, identified in this study, for a simulation model to be successful, both in a conceptual or design level as well as in a practical or analytic level. With regards to the conceptual or design level, these critical factors are the following:

- Whereas it is very difficult and perhaps of no use for all stakeholders to know in detail how the model works, they should have an understanding of the intended purpose of the model, in order to build it, validate it, and eventually use it effectively. Public transportation in general, including the railway sector in particular, is a multi-disciplinary field, and as such it should be assumed that not all involved stakeholders have similar background, either professional or educational. Hence, information should be carefully and appropriately disseminated among the different stakeholders (Balci 1990).

- Whereas it is common knowledge (not only among validation experts) that there is no such thing as *absolute validity* (Martis 2006), it is often overlooked, even by experts. In other words, imperfect models can still be valid. An example of such case was demonstrated in Section 3, where FRISO showed a few discrepancies between the model and reality but it was nevertheless considered valid by experts for its intended purpose.

- Similar to the previous point, the degree to which the model deviates from reality but still remains within acceptable limits is neither certain nor the same for all models (Balci 2004). In other words, a discrepancy observed in one model might not invalidate it, whereas, if it is observed in another model it might do. An example of such case was demonstrated in both Section 3 and Section 4, where in the former the differences on how delays were distributed did not invalidate FRISO, while in the latter, OpenTrack was invalidated.

- One model can be valid for one purpose but invalid for another (Balci 1990). An example of such case was demonstrated in Section 4, where OpenTrack was deemed valid for conflict detection but invalid for testing the punctuality of the timetable. If OpenTrack had been used only to test the punctuality of the timetable, it would have been invalidated and thrown away, resulting in some sort of a Type II error (false negative).

- Similar to the previous point, the methodological approach on validation changes in accordance with the purpose of the model (Balci 2003). An example of such case was demonstrated in Section 4, where one methodology was used to validate the driving

behaviour and the conflict detection of the model, as opposed to a different methodology used to validate the model's ability to accurately assess the punctuality of the timetable.

With regards to the practical or analytic level of simulation models, and specifically the validation of simulation models, these critical factors are the following:

- In public transportation, the validation of a simulation model can focus on a geographical level (for the railway sector that means *stations*), an example of which is shown in Table 2, on a vehicle level (for the railway sector that means *trains*), an example of which is shown in Figure 2, or on a mixture of those two, an example of which is shown in Table 3. Therefore, the appropriate tools should be used to validate a model depending on its focus.
- Given the problem at hand, e.g. testing a new timetable or assessing the punctuality of the current timetable etc., companies and researchers should carefully craft the methodology that would result in a successful validation. This methodology includes the selection of the most appropriate validation methods (Roungas et al. 2017) and the acceptability criteria, as well as the requisite tools for the implementation of the methods in reference and the presentation of the validation results to the experts, who are, in turn, the ones to assess the validity of a model or models.
- During the data analysis stage, one should prefer to initially plot all available information - e.g. Figure 2, including all the operational samples in a graph; depending on the subsequent needs the information may be reduced and more simplistic graphs or tables will be created incorporating only the necessary data under examination. The other way around bears a significant risk of neglecting pieces of information in the beginning, which would later on be proven crucial.

6 Conclusion & Future Work

Simulations are conditio sine qua non in the planning and operations of the railway sector and their success, or failure, depends on several critical factors. In this paper, two simulation models, which were instantiated in two different simulation packages, and their respective validation studies, were presented. While the models per se might not be of great interest to the transportation community as a whole, the lessons learned from their validation certainly are. Some of these lessons are old, yet still applicable, and some of them are new. The lack of absolute validity - even if old - is always applicable and yet neglected many times, resulting in an endless pursuit of the *perfect* model. On the other hand, the visualization of multivariate datasets is an emerging field in which studies often have some new insight to offer. Moreover, what is also new are the methodologies needed to analyse systems with multi-disciplinary nature. These methodologies, along with the system at hand, evolve and adapt, or should at least do so. In turn, the role of simulation models is to give life to those methodologies in an affordable, risk-free, and ethical way.

As a conclusion, the critical success factors identified in this study can be summarized as follows:

- Information and knowledge dissemination should be conducted by taking into account the diversity of the stakeholders, who are involved in such complex systems. As a result, companies should either develop or adjust, as the case may be, their knowledge

management protocols, so as to accommodate the difference in the educational and professional background of all the involved actors.

- The validation of a simulation model is heavily determined by its intended purpose. Hence, the intended purpose of the model dictates the methodology that has to be followed, in order for the validation study to be fruitful and avoid Type I & II errors.
- The analysis and the visualization of data should initially include all the available information, which analysts can then reduce, abstract, or simplify in view of fulfilling their respective goals.

With regards to future work, a more in depth analysis of the algorithms and source code of simulation models which are nonetheless validated, yet not necessarily the ones analysed in this study, could give further insight on what makes simulation models of Complex Adaptive Systems & System of Systems successful. Moreover, the complexity of such systems restricts a modeller from building a simulation model which will encompass, in detail, the totality of the components that comprise them. Therefore, the extrapolation of the aspects of complex systems, which are of paramount importance to the intended purpose of the simulation model, can raise awareness of how models of such systems should be built in order to be valid.

Acknowledgement

This research is supported and funded by ProRail; the Dutch governmental task organization that takes care of maintenance and extensions of the national railway network infrastructure, of allocating rail capacity, and of traffic control.

References

Asuka, M. & Komaya, K. (1996), 'A simulation method for rail traffic using microscopic and macroscopic models', *WIT Transactions on The Built Environment* **18**, 287–296.

Balci, O. (1990), Guidelines for successful simulation studies (tutorial session), *in* O. Balci, ed., 'Proceedings of the 22nd Conference on Winter Simulations', IEEE Press, New Orleans, Louisiana, USA, pp. 25–32.

Balci, O. (2003), Verification, validation, and certification of modeling and simulation applications, *in* S. Chick, P. J. Sánchez, D. Ferrin & D. J. Morrice, eds, 'Proceedings of the 35th Conference on Winter Simulation', Winter Simulation Conference, New Orleans, Louisiana, USA, pp. 150–158.

Balci, O. (2004), Quality assessment, verification, and validation of modeling and simulation applications, *in* R. G. Ingalls, M. D. Rossetti, J. S. Smith & B. A. Peters, eds, 'Proceedings of the 36th Conference on Winter simulation', Association for Computing Machinery, Washington, D.C., USA, pp. 122–129.

Banks, J., Carson, J., Nelson, B. L. & Nicol, D. (1984), *Discrete-event system simulation (Fourth edition)*, Pearson Education India.

DeLaurentis, D. A. (2005), Understanding transportation as a system-of-systems design problem, *in* '43rd AIAA Aerospace Sciences Meeting and Exhibit', American Institute of Aeronautics and Astronautics, Reno, Nevada, USA, pp. 1–14.

Kuhl, F., Weatherly, R. & Dahmann, J. (1999), *Creating computer simulation systems: An introduction to the high level architecture*, Prentice Hall PTR, Upper Saddle River, NJ, USA.

Martis, M. S. (2006), 'Validation of simulation based models: A theoretical outlook', *Electronic Journal of Business Research Methods* **4**(1), 39–46.

Meijer, S. (2012), 'Introducing gaming simulation in the Dutch railways', *Procedia - Social and Behavioral Sciences* **48**, 41–51.

Meijer, S. (2015), 'The power of sponges: Comparing high-tech and low-tech gaming for innovation', *Simulation & Gaming* **46**(5), 512–535.

Middelkoop, D. A. & Bouwman, M. (2001), Simone: Large scale train network simulations, *in* Peters. B. A., J. S. Smith, D. J. Medeiros & M. W. Rohrer, eds, 'Proceedings of the 33nd Conference on Winter Simulation', ACM, Arlington, Virginia, USA, pp. 1042–1047.

Middelkoop, D. A. & Loeve, L. (2006), 'Simulation of traffic management with FRISO', *WIT Transactions on the Built Environment* **88**, 501–509.

Middelkoop, D., Meijer, S., Steneker, J., Sehic, E. & Mazzarello, M. (2012), Simulation backbone for gaming simulation in railways: A case study, *in* C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose & A. Uhrmacher, eds, 'Proceedings of the 44th Conference on Winter Simulation', IEEE, Berlin, pp. 3262–3274.

Nash, A. & Huerlimann, D. (2004), 'Railroad simulation using OpenTrack', *WIT Transactions on The Built Environment* **74**, 45–54.

Rinaldi, S. M., Peerenboom, J. P. & Kelly, T. K. (2001), 'Identifying, understanding, and analyzing critical infrastructure interdependencies', *IEEE Control Systems* **21**(6), 11–25.

Roungas, B., Meijer, S. & Verbraeck, A. (2017), A framework for simulation validation & verification method selection, *in* 'SIMUL 2017: The Ninth International Conference on Advances in System Simulation', Athens, Greece, pp. 35–40.

Trist, E. L. & Bamforth, K. W. (1951), 'Some social and psychological consequences of the longwall method of coal-getting', *Human Relations* **4**(1), 3–38.

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

Zeigler, B. P. & Sarjoughian, H. S. (2012), *Guide to modeling and simulation of systems of systems*, Springer London Heidelberg New York Dordrecht.