Sound evaluation of simulation results

Matthias Becker [a 1], Thorsten Büker [a],
Eike Hennig [a], Felix Kogel [a]
[a] VIA Consulting & Development GmbH
Römerstr. 50, 52064 Aachen, Germany
[1] E-mail: m.becker@via-con.de, Phone: +49 (241) 463 662-26

Abstract
Simulation is one of the powerful means within the toolset of railway operations research. In contrast to timetabling and to queuing theory, it supports a precise representation of interdependencies and has thus a large field of application. Since in today's railway operation many timetable concepts and even big investment-decisions are based on studies conducted with simulation tools, a focus should be set to the sound evaluation of simulation results, too. Nevertheless, the aggregation, validation and interpretation of simulation (raw) data can barely be found in literature. This fundamental task is subject of this paper.

A simulation consists of the following steps: model design, parametrisation and calibration, simulation, processing of raw data, interpretation and visualisation of results. First, various input parameters are manipulated and simulation results are manually evaluated in a simple closed-loop principle. As each simulation is subject to outliers, runs affected by dubious conflict solutions have to be identified and excluded automatically. In most cases, a special focus is on the comparison of different scenarios and the necessity of establishing comparability by forming intersections between the simulation runs. The remaining subset of simulation runs per scenario can be considered (statistically) representative, as soon as the key figure of each scenario series converges. Finally, the raw data can be processed for the evaluation of simulation results. Results of simulations are mostly complex but by producing results for different target groups the complexity has to be reduced without losing important details or provoking misinterpretation. For this reason, it is necessary to choose key figures which comprehensively represent the simulation results.

Keywords
simulation, evaluation, calibration, intersectioning, interpretation

1 Introduction

There are many different procedures to analyse railway operations. All of these approaches have different objectives. By some of them, it is possible to analyse real-time operational data to evaluate the current performance of a railway system, while others focus the calculation of capacity and operational quality by means of queuing theory or simulation. Simulation is one of the powerful means within the toolset of railway operations research. In contrast to pure timetabling and to queuing theory, it supports a precise representation of interdependencies and has therefore a large field of operation. Some of the benefits of simulation are:

1. Illustration of complex systems (infrastructure, timetable and operational procedure)
2. Cost-effective and fast analysis of different crucial questions

3. No need for real time tests on existing infrastructures

With the use of railway simulation tools, it is possible to analyse various different scenarios and evaluate the resulting effects. As an example, the scenarios may differ by infrastructure design (microscopic track layout), by command and control system (CCS) or by timetable-concepts. The results of a simulation run is a huge amount of data. These data mainly consist of planned and actual arrival-, departure- and passage-times of all trains (in all stations) within the simulation model. Afterwards all representative information (punctuality/delay) has to be gathered out. Since in today's railway operation many timetable concepts and even big investment-decisions are based on studies conducted with simulation tools, a focus should be set to the sound evaluation of simulation results, too.

In R&D tradition, there has been substantial work in the development of simulation tools of different nature. While either the simulation algorithm or the simulation evaluation is addressed within a variety of publications, the execution of studies relies on an important interim step: aggregation, validation and interpretation of simulation (raw) data. Barely no literature can be found. This fundamental task is subject of this paper. It is structured as follows: Paragraph 2 describes the motivation for this paper. In paragraph 3 we focus on the requirements for conducting a simulation and describe the possible key figures one can get from railway simulations. Afterwards chapter 4 covers the aggregation and interpretation of simulation raw data. Finally yet importantly, we conclude this paper in paragraph 5.

2 Motivation

There is long-lasting series of research on simulation of railway operation and only some exemplary publication can be listed (Penglin (2000), Gröger (2002), Gray (2013), Jensen (2014), Ochiai (2014), Lindfeldt (2015)). Some microscopic simulation tools, such as RailSys and LUKS, provide an explicit conflict detection and solution in a synchronous and/or asynchronous manner (Weymann (2008)). Recently optimisation components are applied within conflict solution (Weymann (2015)), too.

All publication mentioned above have in common, that they describe either the simulation algorithm or the evaluation of results (with a clear focus on the first aspect). Nonetheless, to achieve reliable outcomes an important interim step may not be discarded: aggregation, validation and interpretation of simulation (raw) data. Standard literature like (Hansen (2008)) leaves out this aspect, too. To close the gap, we try to give some insights within this paper.

Subsequently, the wording is related to simulations following the Monte-Carlo principles: Per scenario, a series of simulation runs is carried out in a deterministic manner. The delays per train and location are returned per run. Results of all runs form a sample that is evaluated by stochastic means to aggregate key figures related to the scenario. For simulation approaches, which rely on direct manipulation of distribution functions instead of Monte-Carlo principles, such as (Büker (2012)), the majority of subsequent considerations is also valid.

3 Requirements

Subject of any simulation are the timetable plus the underlying infrastructure. To achieve a realistic representation of in-field operation, various input parameters serve the calibration of the simulation model:

- Simulation results depend on the magnitude of primary delays being “spread” into the simulation model. Usually, cumulative distribution functions describe the random primary delays, which are sampled to lists of realisations. Per simulation run, a list of realisations is used. (If various scenarios are under investigation, a superset of random variables has to be used to guarantee comparability.)
- Those primary delays cause delays, which may result in secondary delays due to delay propagation.
- The conflict-solution component (e.g. two-train approach or linear optimisation) aims to reduce the magnitude of secondary delays under the regime of a target function. Any conflict-solution component has to be configured by train- and route-dependent priorities as well as malus coefficients to ensure a behaviour close to real world.

3.1 Simulation Model

In order to produce valid simulation results, it is necessary to rely on resilient input data. The allocated primary delays, the available stopping and running time supplements, the dimensioning of the investigation area as well as the settling time are highly relevant. The simulation model has to be fine-tuned to such an extent, that the key figures are sufficiently accurate to draw conclusions either by comparison or in an absolute manner.

Primary Delays

The compensation of delays is probably the greatest challenge of railway operation. In order to represent real world disturbances different disturbance variables are considered in a simulation model:

- Primary delays at entry into the investigation area,
- Primary delays at commercial or operational stops
- Continuous running time extension

Ideally, primary delays at entry and commercial stops can be derived from operational data. If this is not possible – and this is the common situation – one has to make use of standard delays which are ideally differentiated according to type of train and utilization rate.

Supplements

Supplements enable a train to recover from possible delays and to approach the reference trajectory again. The success of this intention significantly depends on the available stopping and running time supplements. As the stopping time supplement is the share of stopping time that is not used for door operation, passenger exchange and dispatching time, it is very important to define this minimum stopping time with caution so that the stopping time supplement is not overestimated.

The running time can be differentiated into a technical minimum running time and into additional running time supplements, which are allocated either for timetable robustness or during timetable construction as part of the conflict solution. A delayed train is able to make use of its supplements with regard to interdependencies with other trains.

Investigation Area
It is necessary to border the investigation area sufficiently large so that partly far-reaching interdependencies can be evaluated within the simulation. In a first approximation, it is useful to limit the investigation area at least at the next larger main railway station. Furthermore, it is recommended to extend the investigation area if there are turnaround station in the closer proximity. As disturbance lists are always inflexible after calculated once, the assumed delays at entry cannot be reduced in scenarios with a better overall operational quality even if trains might enter the investigation area more punctual as a reaction of the more punctual system itself. By the integration of turnaround stations this disadvantage can be reduced, as the arrival delay is propagated to the next train run (minus stopping time reserves) and the number of fix entries is minimized.

Stable State of the System
Besides the geographical definition of the investigation area, it is also mandatory to define a time window to be analysed and as well to determine the necessary lead time in order to guarantee a stable state of operation during the examination time window. A tight lead time provokes that the operating programme has not yet completely started so that the simulation results overestimate the operational quality.

Turnaround and Passenger Changing Connections
In turnaround stations it is a must that consecutive train runs are linked so that the operation with one single train is considered. The simulation tool ensures that the following train run can only start after arrival of the first train and the following time demand for the turnaround, which is usually configurable. If purposeful, dependencies due to staff and passenger transfer times can be defined, too.

3.2 Preparation of the Model for the Simulation

A major driver to perform simulation studies is the adaptation of infrastructure (e. g. reduction, extension, changes to command and control technology). In most cases, the infrastructure model is prepared in a semi-manual manner. Afterwards the timetable is compiled in conjunction with the infrastructure model. Scheduling as well as first series of simulation serve to validate the basic model and to detect and fix modelling errors. (Daily practice underpins that even simulation on infrastructure models for productive train-path allocation requires error elimination, as merely such data is maintained which is necessary to schedule regular train paths.) This validation requires spending a close look to the results of the early simulations instead of blind trust into simulation outcomes. Once all errors have been corrected, the model calibration may be launched.

3.3 Key Figures

The primary output of a simulation is a huge amount of raw data, which have to be analysed, aggregated and interpreted in order to draw meaningful conclusions. Standard key figures are described in various publications. For this reason, key figures are only that shortly defined as it is necessary for the later paragraphs. Standard key figures are:

- Average lateness per train
- Average lateness per delayed train
- Average additional lateness per train

- Number of late trains
- Percentage of late trains
- Punctuality of trains
- Propagation of delays between selected stations

Most key figures cannot only be calculated for simulation results but also be derived from measurements in real world operation. The absolute delays as well as the punctuality, the development of delay over a train run and the travel-time quotients for operation can be calculated based on operational data. Simulation tools offer an additional key figure called infrastructure-related hindrances that usually cannot be derived from measured data as the dependencies cannot be reconstructed. The evaluation of operational raw data is already analysed (Graffagnino (2012)).

Table 1 gives an overview on the consecutively described key figures.

Table 1 Key figures in reality and simulation

Key figure	**Evaluable in reality**	**Evaluable in simulation**
Delay (+ related key figures)	x	x
Punctuality	x	x
Travel-time Quotient Operation	(x)	x
Infrastructure-related hindrances	-	x

Absolute Delay and Development of Delay

The base key figure of any simulation is the absolute delay of a train at each occupation element. Nearly all further key figures are based on the absolute delay. The difference of at least two absolute delays describes the development of delay over a section of a train run. The development of delays is a relative consideration, which helps identifying bottlenecks within the infrastructure that are places of a high delay propagation.

Punctuality

When an absolute delay is compared with a quality target that limits the acceptable delay by aid of a threshold, it is possible to attribute each train run to be punctual or not. The relative share of punctual trains results in the key figure punctuality. This threshold can be defined for each country, infrastructure manager or even system.

Vice versa it is possible to determine a delay that is not exceeded by a defined amount. Typical thresholds are the quantiles as well as the 95 percent probability that excludes five percent of the worst trains.

Travel-time Quotient

The travel-time quotient is represented by two running times. It is possible to differentiate between the travel-time quotient for timetables and for operation. In this paper only the travel-time quotient for operation (TTQ Operation) is relevant and further described. The TTQ Operation describes the quotient of simulated running time of a train and its scheduled running time. Consequently, a TTQ Operation smaller than one expresses a situation where a train realises a shorter running time than planned on the one hand. Usually this phenomenon can be observed, when a train is initially delayed, but has running and/or stopping time supplements that can be used for delay reduction. A TTQ Operation larger than one describes on the other hand that the planned running time is not sufficient so that the train gains delays. Whereas a quotient smaller than one is unambiguously interpretable a quotient slightly larger or equal one may result of insufficient supplements or even a punctual train that has no need to run faster than planned.

Infrastructure-related Hindrances
Infrastructure-related hindrances reveal infrastructure elements that produce or propagate delays. These hindrances are accumulated over all events and their duration. Hindrances usually occur at signals, turnouts and stopping positions as result of a parallel demand of more than one train. The hindrances can be visualized within the track diagram and indicate the bottlenecks within a network.

4 Methodology

This paragraph describes the process from the calibration of a model and preparation of raw data to the interpretation and preparation of the results. A special focus is on the comparison of different scenarios and the necessity of establishing comparability by the identification of outliers and forming intersections between the simulation runs.

4.1 Calibration of the Simulation

In a simple closed-loop principle, various input parameters as well as settings are manipulated and simulation results are manually evaluated. The major mean of validation are time-distance graphs after simulation as they provide the best visualisation of a simulated operation with focus on the simulation specific conflict solution. In this step, calibration happens to the expectations of the user, who needs to have specific knowledge on railway operation in general and to the specific situation. Real-world key figures may serve as secondary reference, only, if available for the specific situation at all. To ensure an overall comparability of outcomes, the calibration principles should thus be as standardised as possible – as well throughout setting up various models as throughout working by different users.

It is in the responsibility of the user to adapt the settings of the simulation tool or the whole model in order to define a proper solution space for the conflict solution. Concerning the infrastructure model it can be useful for instance to remove opposite track movements from the solution space or reduce the costs for alternative track occupations at the same platforms where it is practicable and useful in reality. Furthermore, it can be recommendable to adapt the priorities of a train family, if a train is much discriminated by conflict solution otherwise. This may happen for instance, if a freight train shares its infrastructure with highly prioritized long-distance trains and is unrealistically long directed into sidings due to the target function of conflict solution. From the perspective of operation, it has to be clarified if the simulation may use additional operational stops ahead of junctions in order to reduce the length overtaking sections and enable an earlier departure of the trains from the previous station. This calibration reduces the number of unrealistic conflict solutions and prioritizes real-world conflict solutions even if they might be worse than computed conflict solutions.

4.2 Individual Evaluation of each Simulation Run

After calibrating the model, simulations are ideally performed in a mostly automatic setup. In most tasks, studies do not cover just one simulation series for one (calibrated) model but are of comparative nature. For instance, the optimum combination of infrastructure and timetable shall be found and proven. Subsequently we name a combination of input parameters a scenario. If either timetable or infrastructure vary between the scenarios, disturbances and configuration should be as constant as possible between the scenarios. For each scenario, a series of runs is simulated. To assure comparability of figures between the scenarios, any evaluation has to follow certain principles. Figure 1 provides a first insight into the process of the preparation of raw data.

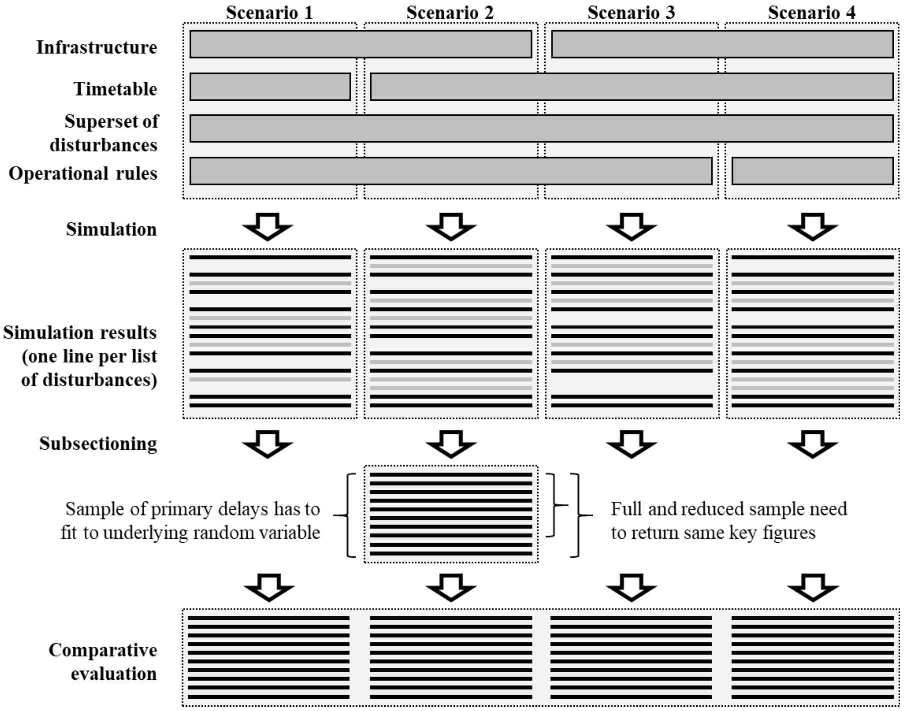
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Infrastructure
Timetable
Superset of disturbances
Operational rules
Simulation
Simulation results (one line per list of disturbances)
Subsectioning
Sample of primary delays has to fit to underlying random variable
Full and reduced sample need to return same key figures
Comparative evaluation

Figure 1 Simplified flow-chart of comparative analysis

Even in case of diligent calibration of input parameters, there are simulation runs whose results differ from actual behaviour severely. This happens as the conflict-solution component, as any human dispatcher, either does not find the optimum solution or even misbehaves. This mostly results from the reduced set of options for conflict solution compared to reality (e. g. partly cancellation of service, discordance of stops). The dubiety of such simulation runs in general correlates to a mix of:

- High primary delays
- Ambitious timetable concepts (only few supplements and low buffer times)
- Complicated/limited infrastructure (e. g. many crossings, single track)

An example of a set of simulation runs including those dubious ones is visualized in Figure 2.

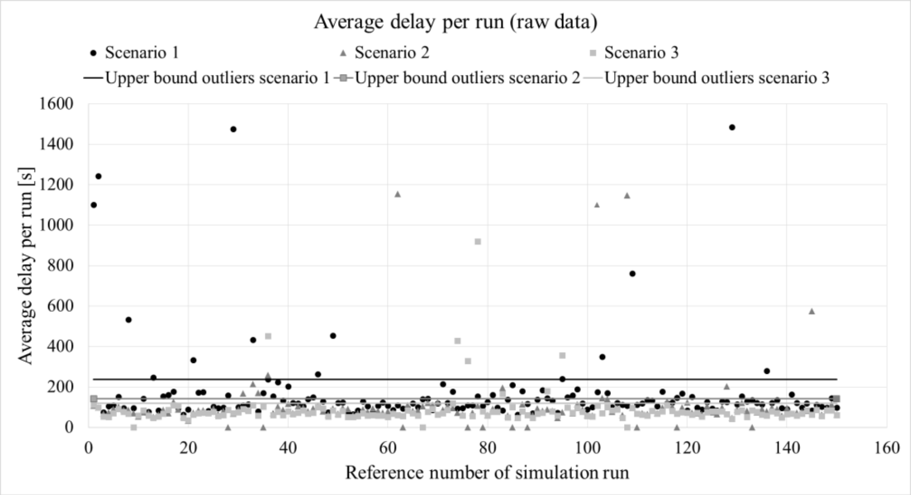

Figure 2 Exemplary set of raw data

As simulations are executed mostly automatic, those runs out of the series being affected by dubious conflict solutions have to be identified. This identification has to happen either on the level of the whole run or on the level of the train. Again, identification has to follow standardised principles to guarantee comparability. A useful key figure for identifying outliers is the average delay in all stations. This key figure is exemplary visualized in Figure 2 for each run for each scenario. For instance, it is obvious that simulation run number 129 of scenario 1 with an average delay of 1485 seconds is an outlier (upper right corner). This statement is supported by the fact that the average delay of scenario 2 and 3 is only 46 and 42. Statistically these outliers can be excluded for each scenario by an upper bound, which can be defined as the 1.5-fold of the range between the 25 and 75 percent quantile on top of the 75 percent quantile. This upper bound is also visualized in Figure 2 by a solid line.

If results of various scenarios shall be compared, a scenario-comprehensive intersection of runs with similar disturbance sets has to be created, firstly. The elimination of outliers and intersecting the remaining simulation runs of each scenario afterwards sometimes reduces the number of remaining evaluable simulation runs considerably. Figure 3 shows the remaining simulation runs after the previously described steps. In this case, the number of evaluable simulation runs is reduced from 150 to 102 simulation runs.

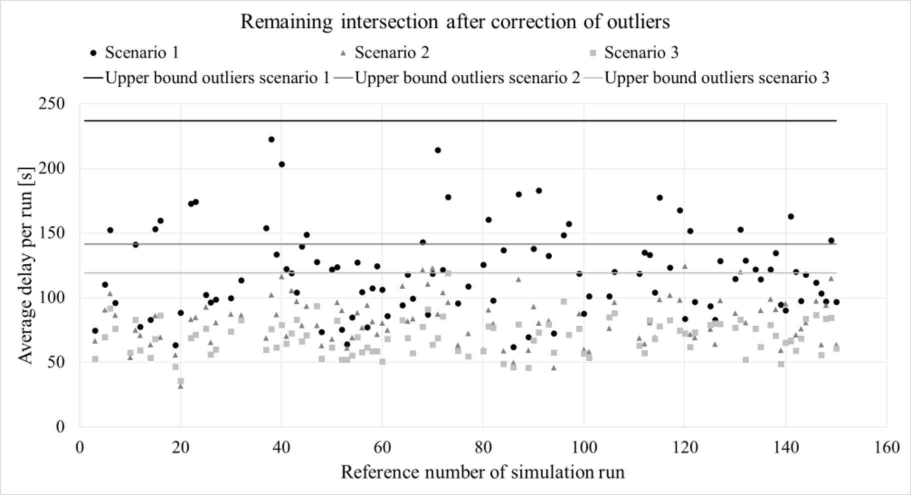

Figure 3 Remaining simulation runs after correction of outliers and intersecting

The related remaining subset of simulation runs per scenario can be considered (statistically) representative, as soon as the key figure of each scenario series converges. This subset is analysed per scenario:

- Firstly, the resulting disturbances have to fit to the distribution function of primary delays
- Secondly, key figures have to converge. For that purpose, the key figure average delay per run of each scenario is averaged over an increasing amount of simulation runs. In case they differ by less than an acceptable epsilon, results are considered to be statistically sound.

Figure 4 shows the effect of converging simulation results over the quantity of simulation runs. It is neccessay that this key figure is statistically distributed as simulated and not sorted according to size. The crosshatched lines represent the upper and lower boundary in relation to the average of all simulation runs plus/minus an epsilon. The epsilon represents the accepted dispersion of the results related to the expected value of the entire sample and is defined as 1.5 %. In this example scenario 1 converges when evaluating at least 60 simulation runs, sencario 2 after 74 and scenario 3 after 76 simulation runs. As all three scenarios shall be compared at least 76 (identical) runs have to be compared. Of course it is necessary to have a sufficient quantity of simulation runs in order to reliably determine the convergence.

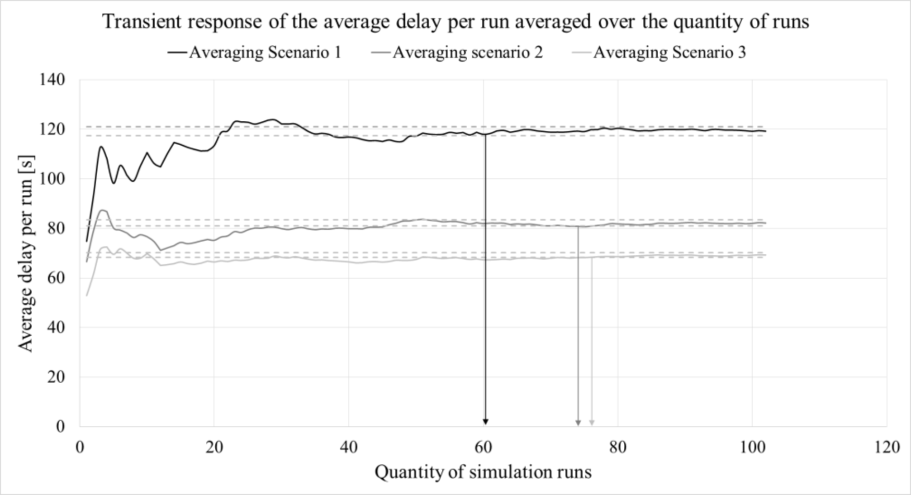

Figure 4 Exemplary convergence of the results of a simulation over the quantity of runs

After the subset of runs per series for evaluation has been identified, further checks can be performed on the layer of trains. For instance, such trains can be excluded from evaluation, which would be subject to cancellation in in-field operation. To ensure a standardised application, real-world delay threshold can be taken into account in this step.

4.3 Evaluation of Simulation Raw Data

After preparation of raw data by eliminating either whole runs or specific trains, the actual evaluation starts and key figures are aggregated. Within the presentation of results it is necessary to produce meaningful key figures for the single scenarios, which represent a general statement. For this, the previously described key figures have to be aggregated in a sound manner. It has proven reliable to interpret any key figure within its context and the return to the roots (infrastructure and timetable) to validate its statement. Results of simulations are mostly complex but by producing results for management level the complexity has to be reduced without losing important details or provoking misinterpretation. For this reason, some key figures are more able to express and to simplify the results than the others. The central issue of the evaluation of simulation results is whether scenario A or scenario B has a better operational quality. This question can be broken down by the evaluation of a model of two trains. Results of a detailed consideration of two identical trains in two different scenarios can be aggregated and for instance be averaged again for a holistic evaluation of all train runs. There are more or less four cases that describe the relation of two trains from the perspective of operational quality. The following figures underline that the expression of one isolated key figure is not necessarily in line with the overall evaluation of a scenario. Furthermore, the suitability of a key figure also depends on the target criterion to be optimized. In the standard case the operational quality shall be improved which means that the sum of delays shall be minimized. Figure 5 illustrates the delay of an identical train in two different scenarios.

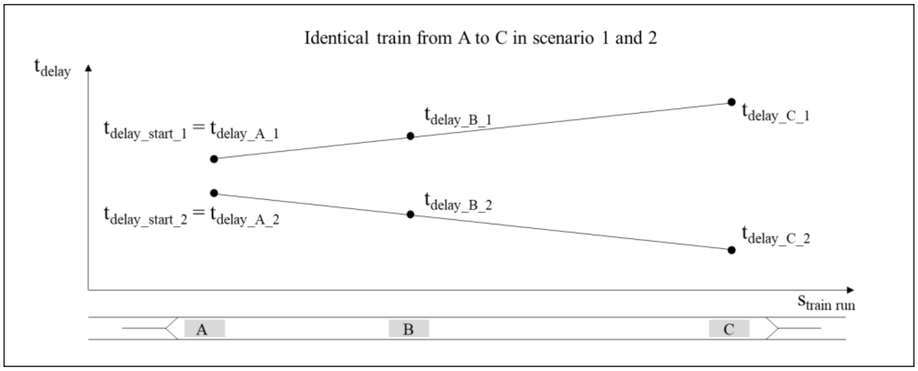

Figure 5 Case 1: Evaluation of the operational quality for a comparison of two scenarios

In the second scenario the train has continuously less delay than in the first scenario. It is obvious that scenario 2 has a better operational quality. This simple analysis is supported by the key figures introduced in paragraph 3.3 and qualitatively summarized in Table 2.

Table 2 Key figures and their statement in case 1

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} > t_{delay\_development\_2}$	Scenario 2
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} > t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 2
TTQ Operation	$t_{simulated\_running\_time\_1} > t_{simulated\_running\_time\_2}$	Scenario 2
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} > t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

Case 2 is an example for a situation where the reduction of the delay of a starting train may result from a longer turnaround time or a more punctual arrival from the previous ride whereas the development of delays remains identical as there were no measures met on the line. A related representation of this constellation is visualized in Figure 6.

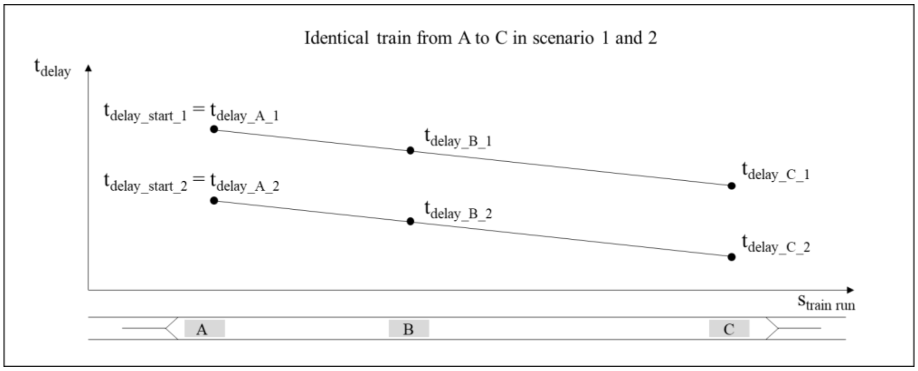

Figure 6 Case 2: Evaluation of the operational quality for a comparison of two scenarios

As the train has continuously less delay in scenario 2 it is undeniable that this scenario is preferable to scenario 1. Table 3 summarizes that the absolute delay in the first and in the last stop as well as the average arrival delay at each stop attest scenario 2 a better operational quality. At the same time the key figure "development of delay" and "TTQ Operation" are not able to detect the better scenario or even lead to wrong conclusions because these key figures only consider the development and not the absolute delay. In this case, a parallel consideration of absolute and relative delay as a combination of absolute delay at start and the development of delay could help to correctly interpret the situation.

Table 3 Key figures and their statement in case 2

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} = t_{delay\_development\_2}$	none
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} >$ $t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 2
TTQ Operation	$t_{simulated\_running\_time\_1} = t_{simulated\_running\_time\_2}$	none
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} >$ $t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

The third possible case is still evaluable by visual checking. In this case the delay of the starting train in scenario 2 is significantly reduced but therefore the delay increases over the course of the train. In reality, this may happen if the turnaround is improved like in case 2 but there is more traffic on the line so that more delays are propagated. The according developments of delay are illustrated in Figure 7.

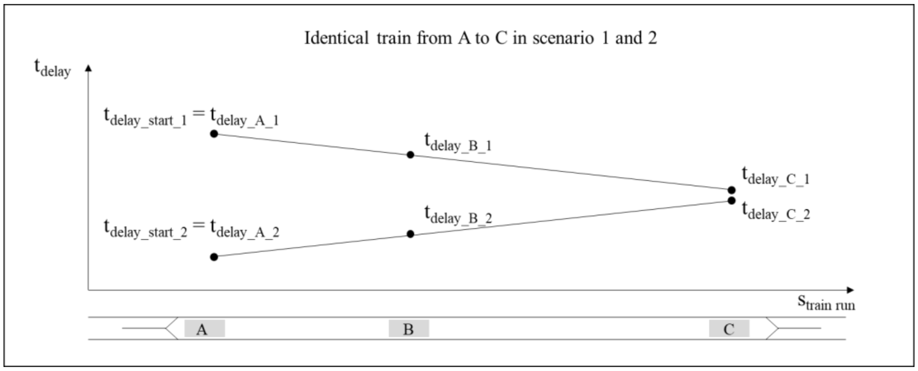

Figure 7 Case 3: Evaluation of the operational quality for a comparison of two scenarios

It is still easily evaluable that the train has a better operational quality in scenario 2 as the amount of delays is continuously smaller than in scenario 1. In this case, the standard key figures have greater difficulty in determining the better scenario than in case 1. The development of delay and the "TTQ Operation" are misleading as they only consider the relative change of delay but do not have any reference to the absolute delay. Given this it seems helpful to combine the development of delay with the absolute delay at start in order to reference to an absolute value. This value corresponds with the absolute delay at the end (compare Table 4).

Table 4 Key figures and their statement in case 3

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} < t_{delay\_development\_2}$	Scenario 1
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} >$ $t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 2
TTQ Operation	$t_{simulated\_running\_time\_1} < t_{simulated\_running\_time\_2}$	Scenario 1
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} > t_{delay\_A\_2} +$ $t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

Case 4 demonstrates that even the combination of absolute delay and the development of delays reaches its limits as soon as the functions of delay intersect. This situation is a blend of cases 1 and 2 and visualized in Figure 8. This case may appear "constructed" but reality shows that a punctual departure and a punctual operation over the train run are completely decoupled and may appear in all possible combinations. From the perspective of the editor of the simulation it is not directly visible in which of the four cases the trains behave between different scenarios. Additionally the cases are mixed within a comparison of scenarios so that it is not easily identifiable which case dominates in which comparison of scenarios.

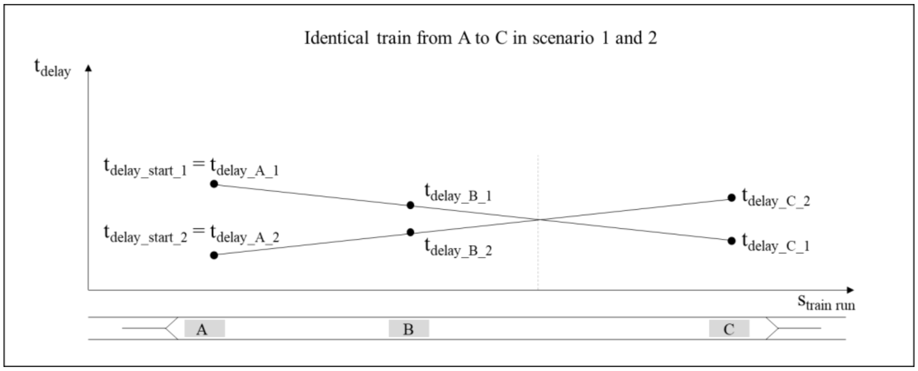

Figure 8 Case 4: Evaluation of the operational quality for a comparison of two scenarios

This setup leads to even more different statements of the key figures whereas it is not possible to evaluate this case only by taking a closer look. Three of five key figures in Table 5 indicate that scenario 1 has a better quality than scenario 2 because the train reduces its delay by making use of its supplements and arrives with less delay in its terminus. Nevertheless, the key figure average delay arrival indicates that scenario 2 has a better operational quality.

Table 5 Key figures and their statement in case 4

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} < t_{delay\_development\_2}$	Scenario 1
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} <$ $t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 1
TTQ Operation	$t_{simulated\_running\_time\_1} < t_{simulated\_running\_time\_2}$	Scenario 1
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} >$ $t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

The reason for this statement is reasonable when comparing the most important aim for each passenger namely the delay at his last stop. In Figure 9 the arrival delays of each station are accumulated. Indeed the sum of delays is smaller in scenario 2 than in scenario 1. As the number of stations is equal in both exemplary scenarios, the average behaves identical. As the number of stops may differ between different scenarios, it is recommended to evaluate the average delay at arrival instead of the sum of delay at arrival. In the comparison of all four cases the average delay arrival is the only key figures which continuously represents the results of the simulation correctly. That does not necessarily mean that the other key figures are unsuitable for representing the results of a simulation but they have to be stated carefully along with explanatory remarks so that a misinterpretation can be avoided.

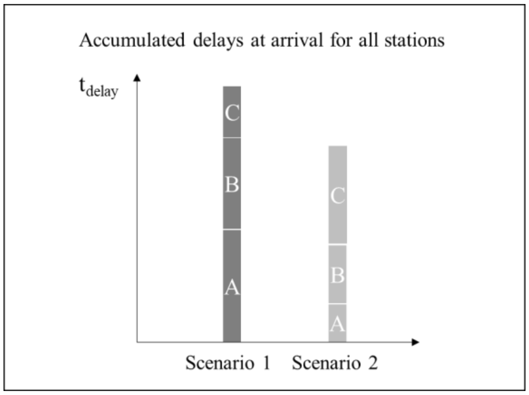

Figure 9 Comparison of accumulated delays at arrival for all stations

Furthermore, we strongly recommend analysing the deviation of the results so that the impact of outliers can be estimated. In case of a high number of outliers it can be helpful to make use of the median instead of the average. Additionally it is recommended to name quantile values in order to give a reference on the distribution of the single events. It is possible to visualize the results of the simulation as the development of delays including quantiles for each train over its train run. The visualisation of quantiles is very descriptive in diagrams as the effect of outliers can be eye-catchingly identified. In an interval timetable, the trains of a train family can be averaged as a compromise of number of diagrams and loss of information. It is also attractive to aggregate the development of delays for corridors. This aggregation has to be chosen very carefully as many effects that affect only one train family are merged into one diagram. By this, it can happen for instance that the punctuality unexpectedly rises at a junction within the corridor. In this case the delays are reduced within a corridor because the merging trains are very punctual. For this reason simulation results should not be aggregated for corridors when the share of trains changes over the considered section.

A further key figure, which is not yet discussed, is punctuality. The punctuality is probably the most famous key figure in relation to railway but it can be less meaningful than the previously discussed key figures. In most cases, the punctuality is measured in the terminus station of a train run and therefore derived from the absolute delay at the end of a train run. The same ad- and disadvantages of this previously discussed key figure remain valid for the key figure punctuality. Additionally it can easily happen that the distribution of delays changes dramatically but the punctuality remains constant. For instance, it is not unlikely that a scenario significantly reduces the number of high delays but as long as the delays are not reduced below the punctuality’s threshold, they are not visible for the key figure punctuality. The same effect can happen for small delays below the threshold. For this reason, the key figure punctuality can be used for further argumentation but never as a standalone key figure to describe the results of a simulation or its operational quality.

Even more caution is required when the travel-time quotient for operation (TTQ Operation) is used for summarizing results of a simulation. In first line this key figure describes whether running time supplements can be used for the reduction of delays or not. On the first glance, a TTQ Operation smaller one suggests a better operational quality compared to a scenario with a TTQ Operation larger or equal one because delays can be reduced. On the other hand, there is no need to reduce delays in scenario where the operational quality is comparatively high. Hence, the TTQ Operation has to be interpreted in context with a more meaningful key figure. As well, the target groups of the key figure TTQ Operation rather consists of timetable schedulers than of managers. A scheduler may adapt the timetable if the simulation reveals that running time supplements cannot be used for the reduction of delays. In a management level this key figure leads to wrong conclusions as the result of the supplements namely the variation of delay and punctuality are sufficient. Table 6 summarizes the advantages and disadvantages of the previously discussed key figures. For the target groups, “S” denotes scheduler, “E” denotes editor and “M” denotes management.

Table 6 Advantages and disadvantages of key figures

Key figure	**Strengths**	**Addressee**			**Significance**	**Interpretability**	**Aggregateability**
		S	**E**	**M**			
Absolute delay	general optimization		X	X	++	++	+++
Delay development	disclosure of bottlenecks		X	X	++	++	+++
Graphical delay	detailed optimization		X	X	+++	+++	++
Punctuality	optimization for threshold			X	+++	++	+++
Average delay at arrival	holistic interpretation		X	X	+++	+++	+++
TTQ Operation	evaluation of reserves	X	X		++	+	+++
Infrastructure-related hindrances	disclosure of bottlenecks		X	X	+++	+++	+++

5 Conclusion

Conflict solutions of today’s simulation tools are continuously approaching real world’s operation. In order to draw the right conclusions of simulations it is necessary to evaluate and compare them correctly. Remaining weaknesses of simulation tools have to be detected and eliminated in the preparation of simulation raw data so that they do not affect the overall statement. Filtering out outlier simulation runs supports the evaluation of reliable and durable simulation runs. Secondly, it is important not to compare apples and oranges, which means that excluded simulation runs of one scenario have to be excluded in all other scenarios, too. The previous steps may lead to a massive reduction of evaluable simulation runs. For that reason, it is mandatory to proof the convergence of the results of each scenario for excluding the danger that the quantity of considered simulation runs affects the results. In the end it is important to interpret results of the simulation in sound manner and derive key figures which represent the results of the simulation. What might sound trivial is a rather complex problem as most key figures can be misinterpreted when considered solitary. Graphical courses of delay support pale key figures and concentrate facts and circumstances.

6 References

Büker, Th., Seybold, B., 2012. *Stochastic modelling of delay propagation in large networks.* In: Journal of Rail Transport Planning & Management 2 (2012), pp. 34-50.

Gray, J., 2013. *Rail simulation and the analysis of capacity metrics*. In: Australasian Transport Research Forum 2013 Proceedings 2 - 4 October 2013, Brisbane.

Gröger, Th., 2002. *Simulation der Fahrplanerstellung auf der Basis eines hierarchischen Trassenmanagements und Nachweis der Stabilität der Betriebsabwicklung.* Dissertation at RWTH Aachen.

Graffagnino, T., 2012. *Ensuring timetable stability with train traffic data (Comprail),* SBB AG, Switzerland

Hansen, I.; Pachl, J., 2008. *Railway Timetable & Traffic: Analysis - Modelling – Simulation*, Eurailpress (2008).

Jensen, L.; Landex, A.; Nielsen, O. *Evaluation of robustness indicators using railway operation simulation.* In: Computer in Railways XIV (2014), pp. 329-339.

Lindfeldt, A., 2015. *Validation of a simulation model for capacity evaluation of double-track railway lines*. In: Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis (RailTokyo2015), Tokyo, Japan.

Ochiai, Y., Nishimura1, J., Tomii, N., 2014. *Punctuality analysis by the microscopic simulation and visualization of web-based train information system data*. In: Comprail (2014).

Penglin, Z., Schnieder, E., 2000. *Modelling and Performance evaluation of railway traffic under stochastic disturbances.* In: IFAC Control in Transportation Systems, Brunswick, 2000.

Weymann, F., Kuckelberg, A., Wendler, E., 2008. *Coupling of synchronous and asynchronous methods in the simulation railway operation.* In: Proceedings of the 10th International Conference on Application of Advanced Technologies in Transportation (AATT 2008).

Weymann, F., Nießen, N., 2015. *Optimisation processes to assist with fine compilation of timetables.* In: ETR International Edition 1 (2015) 1, pp. 24-27.