Use of Availability Concepts in the Railway System

BIRRE NYSTRÖM

Luleå University of Technology, SE 97187 Luleå, Sweden

(Received on December 21, 2007)

Abstract: This paper explores and describes indicators related to availability in the railway system. Examples of indicators include train punctuality and condition of track. The paper presents losses to the stakeholders, measured by the indicators. Indicators in Banverket (Swedish Rail Administration) and railway literature are presented. Indicators used in the electric power industry are also presented, as these offer good analogies from which to develop additional railway indicators. The indicators found in the literature, but not in Banverket, include travel time variation, slack and wagons. Indicators found by analogy to the power industry concern passengers, traffic work not delivered and how to subtract the effects of adverse weather. A classification of availability indicators is also suggested.

Keywords: availability indicators, capacity, railway punctuality, electric power industry, maintenance

1. Introduction

Train traffic in the European Union is increasing [1], i.e. more railway network capacity is being used. Simultaneously, there is a wish to improve punctuality [2], while more trains on the tracks might cause delays to spread more easily. This might be counteracted by increasing capacity by upgrading existing railway lines or building new ones, and/or increasing railway availability; in other words, by improving the dependability of existing lines. As several European railways are in a state of transition from one single governmental body responsible for the railway to several actors, public and private. These circumstances call for good ways of communicating between railway stakeholders regarding capacity and availability. The analysis of railway capacity has a long tradition, dating back at least to 1850 [3], while the interest in availability is of more recent origin.

In this paper, availability concepts that might be of use to railway stakeholders are illustrated. The paper begins with a description of the stakeholders in the Swedish deregulated railway. Thereafter, different railway capacity and availability concepts are described. Then, possible losses due to unavailability are described and indicators reflecting these losses found in Banverket (the Swedish Rail Administration) and in railway literature are given. Thereafter, the electric power industry is discussed and power industry indicators given, as well as possible analogies to the railway system. By means of a questionnaire, personnel at a centralised train traffic control office were asked to judge the relevance to their work tasks of the prospective indicators (found in Table 1, Table 2 and Table 4). The indicators that the largest number of respondents judged to be relevant are indicated by a ‘☺’ in the tables, and the indicators that the largest number of respondents considered irrelevant or non-understandable, are indicated by a ‘☹’ [4].
The paper finishes with a classification of indicators related to availability, and conclusions with some suggestions for further research.

2. Railway Organisation

Important stakeholders of the Swedish railway are shown in Figure 1. The trains might be run by any of the Train Operating Companies (TOCs). Banverket is the Infrastructure Manager (IM), creates the timetables and is in charge for train traffic control (TTC). The infrastructure maintenance is performed by any of the Maintenance Contractors (MCs). Outside the train traffic production are the customers and public subsidisers.

The train traffic controllers base their work on the timetable and get input from, for example, train drivers, infrastructure personnel and the train positioning system. The timetable establishes a trade-off between different TOCs, as well as between traffic and track works, for example preventive maintenance. When non-planned events occur, the controllers make trade-offs, utilising different recovery strategies, such as turning trains prior to their timetabled terminus and delaying trains in order to lower total delays [5].

![Figure 1: Railway organisation and important related stakeholders in Sweden.](image)

3. Railway Capacity and Availability

The capacity of a railway line is affected by the timetable, train traffic control, signalling, top speed, stations and rolling stock. The ability of train traffic controllers affects the railway capacity when there are deviations. The capacity of railway terminals is affected by the gates, tracks, as well as loading, unloading and storage facilities [6].

Generally, capacity is defined as the ability of an item to meet a service demand of given quantitative characteristics under given internal conditions. Internal conditions might refer to e.g. a specific combination of functioning and not functioning sub-items [7]. UIC (International Union of Railways) defines theoretical capacity of a railway infrastructure as the number of trains able to run over it per unit of time, with the trains permanently running with minimum headways between them, i.e., the traffic is as dense as it can be and there are no delays. As running at minimum headway means that there are no margins to prevent disruptions from resulting in secondary delays (i.e. trains are delayed because of other trains hindering them), the theoretical capacity is not obtainable in practice. UIC defines practical capacity of an infrastructure as the number of trains able
to run over it per unit of time with the degree of operating quality which corresponds statistically to the level desired, excluding major disruptions [8]. A more succinct way to define (practical) capacity of an infrastructure is as the ability to forward trains with acceptable punctuality [9]. Note that both these definitions of practical capacity require detailing acceptable punctuality.

3.1 Capacity and the Timetable

Consider different timetables for a part of a railway line, given by the time-space diagrams of Figure 2, where six trains travel upwards in each diagram. In (a), the trains have two different speeds, while in (b), they all have the same speed, which gives a shorter cycle time, i.e. the capacity is higher. Rearranging the trains of (a), e.g., as in (c), also increases the capacity of the railway line. The speeds of trains affect the braking distance, and thus the headway, meaning that higher speed does not always give higher capacity. To summarise, the timetable has to be considered when discussing capacity.

Figure 2: Three timetables, each showing six trains. The timetables are heterogeneous (a), homogeneous (b) and with similar train speeds bundled (c). Time is on the x-axis and place is on the y-axis, following Swedish convention.

One way to decrease the impact of the current timetable on the capacity calculations of a railway line is to theoretically compress the trains, i.e. to place the trains at minimum headways, as in Figure 3b. This is the method prescribed by the UIC 406 standard [10]. The output is the share of time which the track is occupied by a train between the considered places (about 50% for the timetable in Figure 3). UIC 406 preserves the order of trains when compressing, while Banverket’s method rearranges the trains (similar to changing from Figure 2a to Figure 2c). Hence, the methods might give different figures.

Figure 3: Timetable compression for calculation of capacity utilisation.

Another way is to reformulate the concept of timetable using the demand for train traffic, rather than a specific time-space diagram. This idea is elaborated in [11], in which a traffic pattern is defined as a set of train movements \( \{x_1, x_2, \ldots, x_n\} \) and a partial order on these, i.e. there are relations of the type \( x_i \geq x_j + \text{time} \). The capacity of a railway network is then taken to mean the inverse of the time between two consecutive traffic patterns. In this way, one avoids calculation of capacity using timetables that were constructed considering railway network bottlenecks that have been removed. Practical capacity
utilisation might be higher than the theoretically calculated capacity utilisation, because trains are delayed (the space in-between a non-delayed and a delayed train is not possible to use for other trains) and because train routes are set early (that is, the space in front of a train is reserved well in advance). A case study at The Hague HS shows a considerable difference between theoretical (50%) and practical (70%) capacity utilisation [12].

To be able to cater for deviations, timetables include allowances, i.e. extra time in excess of the theoretical minimum travel time. Allowances might be allocated along the line or at stations. Margins are the extra time between two consecutive trains [13], i.e. the headway is longer than theoretically needed. Theoretically, allowances and margins lower the capacity of the network. In practice, they are added to reduce delays.

3.3 A Definition of Availability

A general definition of availability (performance) is: the ability of an item to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval, assuming that the required external resources are provided [7]. An item might be anything from a tiny component to an entire system, as long as it can be considered separately. A definition of railway availability is:

\[
\text{Availability} = \frac{\text{Obtained capacity}}{\text{Planned capacity}}
\]

and is calculated or a point in time or over a time interval. The Planned capacity might be set to the theoretical capacity, the practical capacity, or lower. The concepts are illustrated in Figure 4. The Obtained capacity consists of required capacity and spare capacity.

![Figure 4: Availability might be defined as Obtained capacity / Planned capacity.](image)

Spare capacity might absorb variations from day to day or a future traffic increase [14]. As seen earlier, no commonly acknowledged definition of the capacity of a railway line exists. The same is true for availability.

4. Performance of the Railway

Figure 5 shows how performance of the railway might be considered. The performance experienced by the passengers is the right-most arrow in Figure 5. The control affects the infrastructure and rolling stock, but not the passengers. Passengers affect performance, or seen the other way around, losses. An example is how fast they board the train. The performance of the infrastructure, the performance of rolling stock (including personnel) and the performance perceived by the passengers, might all be named ‘performance’.
Note that one cannot calculate the total availability as the product of the availabilities of the infrastructure, rolling stock and passengers. This is because they might be available at different times and an item might be considered available, despite its receiving no input and therefore not being able to perform [15].

5. Indicators Related to Unavailability in Banverket and in the Railway Literature

Losses related to unavailability (i.e. lack of availability), encountered in Banverket and in the railway literature, are listed in Table 1 and Table 2, respectively. Most of the indicators can be detailed further and, directly or indirectly, affect all of the stakeholders. Table 1 and Table 2, show that indicators encountered in the railway literature, but not in Banverket, deal with travel time variation, slack (more time in the timetable) and wagons.

5.1 Availability Concepts in Banverket

Although having no formalised availability concept, Banverket employees use the term availability and there are related indicators in its strategic plan [21]. Some of the indicators, as signified in Table 1, are readily associated to a delay code in Banverket’s delay attribution system, TFÖR [16]. The indicators reflect availability, although the requirements in Table 1 are formulated in other words also. Table 1 shows that the indicators range from changes in the Network statement (which, for a certain timetable period, describes the infrastructure, specifies order of capacity allocation and requirements on TOCs [17]) to punctuality and the condition of the railway infrastructure.

The indicators in the first and second rows of Table 1 consider punctuality. The punctuality indicators reflect the effect of unavailability, for example the ‘Fraction of arrivals within 5 minutes from timetabled time’ (data is in the TFÖR system). The ‘Number of unplanned speed restrictions’ tells that availability of track is lowered, which might cause trains to become unpunctual. Timeliness, i.e. the extent to which a journey or transport occurs when desired, is reflected by e.g. the ‘Fraction of train path applications granted’, which tells to what extent the demands for track time of the TOCs are fulfilled by the timetable. In other words, to what extent the supply of train paths covers demand. However, a low fraction of applications granted does not necessarily means that there is a high capacity utilisation – it might mean that the TOCs, taken together, demand train paths that are not possible. The ‘Number of time exceeded for track work’ is, to some extent, found in TFÖR, in which occasions when trains get delayed hereby are recorded.

For Disabled time of the railway infrastructure, the second last row of Table 1, it is possible to get only approximate data from Banverket’s infrastructure failure database. However, the symptoms and causes of failures, are detailed, which makes it possible, to some extent, to subtract various externalities so as to obtain more comparable indicators. External causes of unavailability might include bad weather and sabotage.
5.2 Availability Concepts in the Railway Literature

Indicators reported in the literature or constructed by the author are listed in Table 2. The indicators range from timeliness to information on cancelled track works.

The indicator on the first row of Table 2, Number of non-delayed passengers / Number of delayed passengers (in [18] called ‘Service reliability’) has the disadvantages of being unbound and being sensitive to changes to the threshold when a train is to be considered as delayed, which may make it difficult to make historical comparisons. Dwell delay, i.e. delay due to extra passengers embarking, is an effect from train delays, but also passenger behaviour and passenger information.

The amount of slack time relative to total travel time is an entity encountered in papers on railway operations research, e.g. [19]. This inspires the author to define availability as:

\[ A = 1 - \frac{\text{Slack time}}{\text{Total travel time}} \]

for a specific train, for a specific TOC as well as for all TOCs. This is an indicator of track time losses and hence also of lack of timeliness. The indicator ‘Fraction of planned but not used track work time’ tells how often the maintenance organisation does not use the assigned time. It is cumbersome to calculate, since actual track work time is currently recorded only on paper. The indicator should be put in relation to the time exceeded for track work (Table 1). Fraction of not accurately registered train data gives the unavailability of one kind of information needed to control train traffic appropriately. The problems associated with this may be remedied by denying departure to trains not registered.

Krueger [20] gives a formula for the delay volume as a function of railway line layout, maintenance, traffic composition and traffic volume, in his study on the Canadian National Railway. Simulation involving characteristics of a line (e.g. how uniformly sidings are distributed), traffic (e.g. heterogeneity) and maintenance (e.g. track work time) were used to construct a delay estimate. Key parameters, such as the ones exemplified, make up each of the factors in the approximation:

\[ \text{Delay volume} = \text{Factor line} \times \text{Factor traffic} \times \text{Factor maintenance} \times e^{k} \times \text{Traffic volume} \]

Now specify the delay volume (i.e. acceptable punctuality). The Factor traffic might then be regarded as the availability (to train traffic) of the investigated line, and is then given by the equation, when the line, maintenance regime and traffic volume are known. The author notes that this way of defining availability is analogous to how capacity is defined in [9]. One way to measure heterogeneity is given in [22]. In [22], the Sum of Shortest Headway Reciprocals (SSHR) is defined as

\[ SSHR = \sum_{i=1}^{n} \left( h_{i}^{-1} \right) \]

with \( h_{i} \) the shortest headway between trains i and i+1, train n is followed by train 1.

The quotient Actual delays / Simulated delays might be used as an indicator. It shows how well availability is allocated among items of the same type, e.g. turnouts. One way to calculate it is as follows. Assign each turnout the same probability of functioning as the average turnout. Simulate the train traffic along the railway line. A small quotient then indicates a good allocation of availability, i.e. availability is where it is most needed.
Table 1: Losses associated with unavailability, found within Banverket. In the left-most columns, the stakeholders to whom the indicators are most relevant, are indicated by ‘x’. The starred indicators were not explicitly defined by Banverket. The relevance to train traffic control personnel is indicated by ☺ or ☹. The rows with thick vertical borders are further described in the text.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Indicator</th>
<th>Loss scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punctuality</td>
<td>Fraction of departures within _ min; fraction of arrivals within _ min; sum of arrival delay minutes (TFÖR)</td>
<td>Delay may incur costs; hindering of other trains; lost connections; worried passengers.</td>
</tr>
<tr>
<td>Punctuality</td>
<td># unplanned speed restrictions [21]</td>
<td>Delays.</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Fraction of train path applications granted [21]</td>
<td>Demand is not fulfilled.</td>
</tr>
<tr>
<td>Timeliness</td>
<td># changes to the Network statement [21]</td>
<td>Planned traffic might not be carried out.</td>
</tr>
<tr>
<td>Passengers / goods on embarking / loading place</td>
<td># late departure from freight terminal (TFÖR)</td>
<td>Delays; more work for yard workers; time losses due to extra transfer or change of transport.</td>
</tr>
<tr>
<td>Clearance gauge</td>
<td>Fractions of different clearance gauges*; # not reported changes to encroachments when a track work has been carried out*</td>
<td>Too small clearance gauges restrict traffic; too large might mean costs to IM; unreported encroachments might cause accidents; costs for finding clearance information.</td>
</tr>
<tr>
<td>Maintenance slot</td>
<td># time exceeded for track work (TFÖR)</td>
<td>Too much or too little time has been planned.</td>
</tr>
<tr>
<td>Availability</td>
<td># faults divided on type of asset and cause code [21]</td>
<td>Failure causes unavailability, which causes repair costs.</td>
</tr>
<tr>
<td>Availability</td>
<td># disturbances resulting in train disturbances [21]</td>
<td>Failure causes unavailability, which causes train delay.</td>
</tr>
<tr>
<td>Availability</td>
<td>Disabled time (several related indicators are in [21])</td>
<td>Unavailability.</td>
</tr>
<tr>
<td>Condition, Availability</td>
<td># inspection remarks broken down by urgency [21]</td>
<td>Infrastructure in too bad condition means expenses for corrective maintenance for IM, losses of comfort for TOC, etc. Too much preventive maintenance means expenses for IM and lower availability for TOC.</td>
</tr>
</tbody>
</table>
5.2.1 Multi-state and Multi-dimensional Availability

The last row of Table 2 concerns multidimensional availability and gives dimensions in which availability might be lowered. When an item can be in more states than an up-state or a down-state, it is said to be able to attain multiple states of availability. When, e.g., a light bulb fails, it changes to a complete down-state. Still, this puts the railway network in only a partial down-state, as trains still might be allowed to pass the signal at reduced speed. The EN 50126:1999 Railway Standard [23] defines System Availability, A, as:

\[ A = \frac{\text{Mean Uptime}}{\text{Mean Uptime} + \text{Mean Downtime}}, \]

which seems to require items to be in either a down-state or an up-state (0 or 1 system). However, EN 50126 also categorises failures in different classes:

- **Significant** (immobilising failure): Prevents train movements or causes a delay and/or cost over specified thresholds
- **Major** (service failure): Must be rectified to achieve specified performance, do not cause delay or cost over thresholds specified
- **Minor**: Does not prevent a system to achieve its specified performance

This categorises some train-stopping failures causing delays just longer than the threshold, e.g. ice in a turnout as Significant, while chronic speed-restricting failures causing delays just under the threshold, e.g. some signalling failures, are categorised only as Major.

6. Availability Concepts in Other Industries

A few industries are studied in order to be compared to the railway: the military mission, computing, electric power, house painting and manufacturing. The industries might be classified according to their relation between production and consumption, with respect to time and space, as in Table 3. For the purpose of this paper, useful analogies may be drawn from the electric power industry. Therefore, it is described in greater detail.

6.1 Electric Power

The electric power industry has moved from integrated utilities into companies generating the power and companies distributing the power. An electric power network is built in a hierarchy of power grids, with transformers between grids of different voltages. Interesting to customers are the interruptions (i.e. loss of electric power), not the outages (i.e. faulty state of an item) in higher voltage grids or in power generation, because grids closer to the customers often are redundantly fed from several higher voltage grids [24].

Power ‘System reliability’ might be divided into ‘System adequacy’, meaning to what extent the energy supplied is enough to cover demand, and ‘System security’, meaning the power system’s ability to deal with disturbances inside itself [29].

‘Time used factor’ is the fraction of total time a power plant produces power, regardless of the effect generated, called ‘on-stream availability’ by Aven [25] and the Z-016 oil and gas industry standard [26]. The continuous nature of oil & gas transports motivates the use of ‘Throughput availability’, defined as the expected throughput as a fraction of the demand. But, as the production of railways is not continuous and 100% throughput availability might be the case even in the case of much unpunctuality, this is not a suitable indicator for railway traffic, at least not for longer periods of time, as it is only affected by cancelled trains. The ‘Time availability factor’ is defined as the fraction of total time the plant produces power or is on standby (i.e. ready to produce power) [27].

The severities of interruptions are calculated as ‘Power disconnected’ (kW) and ‘Energy not supplied’ (kWh). In order to estimate the energy not supplied, Swedish Vattenfall uses e.g. the day’s average temperature [28]. It is unclear how ‘Energy not
Use of Availability Concepts in the Railway System

Table 2: Losses associated with unavailability, found in railway literature or constructed by the author (starred). In the left-most columns the stakeholders to whom the indicators are most relevant are indicated by ‘x’. The relevance to train traffic control personnel is indicated by ☺ or ☹. The rows with thick vertical borders are further described in the text.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Indicator</th>
<th>Loss scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punctuality</td>
<td># non-delayed passengers / # delayed passengers</td>
<td>Delayed passengers.</td>
</tr>
<tr>
<td>Punctuality</td>
<td>Travel time variation.</td>
<td>Delays; lost connections; increased waiting; less train traffic fits into the railway network.</td>
</tr>
<tr>
<td>Punctuality</td>
<td># last minute platform change; # dwell delay, i.e. delay due to extra passengers embarking</td>
<td>Departure delay; crowding; bad traffic information.</td>
</tr>
<tr>
<td>Availability</td>
<td>Fraction of trains that have planned standard for all wagons.</td>
<td>Too few or wrong wagons, which causes delays and/or lower comfort.</td>
</tr>
<tr>
<td>Timeliness</td>
<td>$A = 1 - \frac{\text{Slack time}}{\text{Total travel time}}$</td>
<td>Longer travel time.</td>
</tr>
<tr>
<td>Maintenance information</td>
<td>Fraction of planned but not used track work time*</td>
<td>Worse condition of track; lower capacity available for trains.</td>
</tr>
<tr>
<td>Traffic/train information</td>
<td>Fraction of not accurately registered train data*</td>
<td>Lower registered than factual train weight, renders the IM lower charges and trains might get stuck on slopes.</td>
</tr>
<tr>
<td>Availability</td>
<td>Formula for availability given punctuality, maintenance, line and traffic volume</td>
<td>Suboptimal use with respect to current traffic, maintenance and assets.</td>
</tr>
<tr>
<td>Punctuality</td>
<td>Actual delays / Simulated delays*</td>
<td>Availability is not allocated to the best places possible in the network.</td>
</tr>
<tr>
<td>Availability</td>
<td>Multi-dimensional availability: loading, embarking, allowed trains and speeds</td>
<td>Lower performance than expected.</td>
</tr>
</tbody>
</table>

Table 3: Different products have different time and space relations between production and consumption.
The most commonly used availability-related indicators all calculate averages with respect to the customers. These include System Average Interruption Frequency Index (SAIFI), System Average Interruption Duration Index (SAIDI), Customer Average Interruption Duration Index (CAIDI) and Average Service Availability Index (ASAI). The calculation of the indicators is straightforward. SAIFI is calculated as the number of interruptions per time unit per customer. SAIDI is the average length of interruptions for customers served during a year. CAIDI concerns the lengths of the interruptions for the affected customers (customers who suffer several interruptions is counted as multiple customers). ASAI is the proportion of time when power is available. Another indicator is Customer Average Interruption Frequency Index (CAIFI), defined as the average number of interruptions for the affected customers during a certain period of time. Customer Total Average Interruption Index (CTAIDI) is similar to CAIDI and is the average length of interruptions for affected customers (customers who suffer multiple interruptions are counted as one customer). Average System Interruption Frequency Index (ASIFI) is calculated as interrupted effect divided by connected effect. In the case of large differences in customer size, it gives a more accurate picture than SAIFI does. Average System Interruption Duration Index (ASIDI) is calculated analogously and includes the duration of the interruption [29], [30]. Yet another indicator is the recovery ability, given for example as the number of customers reconnected per day during a large interruption.

SAIFI, SAIDI and CAIDI give different rankings of the causes of interruptions in Canada, e.g. SAIFI points out the cause ‘Defective equipment’ to be the worst, while SAIDI gives ‘Loss of supply’ and CAIDI ‘Adverse weather’ [31]. This illustrates that a tacit valuation is made by the selection of indicator. From the customer’s point of view, it is better to experience many interruptions at short time intervals than the same number of interruptions separated by days or weeks, which affects the suitability of indicators [31]. Large interruptions, caused by e.g. extreme weather, might be treated separately when evaluating. The IEEE P1366 standard [32] defined ‘major reliability events’ as 10% of the customers being affected in a 24-hour period, while IEEE 1366-2003 [30] employs daily SAIDI to statistically define the ‘major event days’ by calculating a SAIDI threshold.

For the Swedish power market, the Network Performance Assessment Model [24] calculates the regulation price. The regulated maximum price is computed by constructing a virtual, ideal network.

6.2 Availability Concepts in Other Industries Summarised

A characteristic of the power industry is that power generators and grids on different voltages are all needed to deliver energy to the customer, but redundancies in generation and higher voltage grids allow for unavailability, without affecting the delivery. Availability, or rather unavailability, might be measured as the fraction downtime, effect lost or amount of energy not delivered, the latter being somewhat difficult to estimate. How to consider the availability of a distribution grid that is not producing (e.g. due to a breakdown in a higher voltage grid) is also an issue. One needs to be aware that different indicators might give different rankings of causes of failure, as well as how one chooses to assess planned interruptions, compared to unplanned interruptions. How to get fair indicators despite external impact is also an issue. One way to measure performance is to compare a virtual network to the real network, regarding interruptions.

For the military mission, the performance of, e.g. an aircraft might be measured by its System effectiveness, in [33] defined as the product of three probabilities:

\[ \text{System effectiveness} = \text{Operational readiness} \cdot \text{Mission reliability} \cdot \text{Design adequacy} \]
‘Operational readiness’ is the probability that the item has begun operating within the time stated. ‘Mission reliability’ is the probability that the item will continue to operate, given that it did so at the beginning of the mission. ‘Design adequacy’ is the probability that the item finishes its mission, given that it operates within its design specifications, thus the similarity to power ‘system adequacy’ is understood.

For the computer industry, faulty specifications are a major concern, probably as the environment and users’ requirements change fast. This makes an issue of a quantification of maintainability (including performance during updates). Concepts are in [34] and [35].

For manufacturing, operations to produce, e.g., a torch might include turning, painting, assembling and packaging, possibly with buffers in-between. The finished product is stored and variations in demand are thereby handled at the cost of the capital tied up and the risk of the product not being in demand in the future. The manufacturing speed has an upper limit, but often, machines are not run at top speed, because e.g., cutting tools might wear uneconomically fast. So, defining a specific top speed might be hard to justify. The speed of the operations might not tell production speed, as faulty pieces may be reworked.

In the presented industries, it is seen that availability might alternatively be viewed as a frequency, probability or as the product of three probabilities; start mission, complete mission and design allows success. None of the encountered indicators answer how a scattered uptime should be described in terms of availability, compared to a continuous uptime. It is preferable to measure the benefit (e.g. energy) lost during downtime rather than downtime hours. The production profile varies between different industries; from producing when needed, as in the military and in power, to producing according to a timetable, as in the railway system, to producing constantly at a maximum speed, as in manufacturing. Therefore, availability, defined as a quotient between capacities, means different things in different industries, and the figures should therefore not be compared.

7. Indicators Related to Unavailability in the Electric Power Industry and their Analogies to the Railway

Analogies can be drawn among the industries discussed in Section 6, for example computing, and the railway. Finding relevant analogies in the military and house painting was difficult. Regarding manufacturing, analogies were found, but as its products are possible to store, it is difficult to find analogies to its customers (corresponding to passengers). Unlike manufacturing, it is not possible to reject the train transport product at a final check and rework it. Electric power has a similar stakeholder structure to that of the railway; power distributor (corresponds to the IM), power generator (TOC) and power consumer (passenger). A difference is that electric power production does not follow a timetable, as demand varies.

Electric power indicators that might supplement the existing railway indicators are listed in Table 4, from which it is seen that many indicators are relevant to passengers. The drawn analogies include power (kW) equals speed (km/h) and energy (kWh) equals travelled distance (km). The railway indicators constructed might be further specified, e.g., speed might be measured as km/h, passenger-km/h or tonne-km/h.

On the third row of Table 4, the SAIFI analogy, ‘Number of trains hindered’, is relevant to passengers, but also to TOCs, TTC and IM, as they might want to know the number of trains hindered by, e.g. turnout failures. The indicator MAIFI_E reflects the fact that outages often cause one another to happen in a rapid sequence. An analogue indicator in the railway system might consider gradually failing items that finally become totally unavailable, e.g. a turnout. Such an indicator might be taken to be the number or length of long-lasting total unavailability of an item. Another railway indicator found by analogy is
the fraction of passengers that suffers more than a specific number of delays per year, calculated e.g. for popular commuter routes.

In Table 4, the last row concerns extreme events. To exclude delays due to severe weather, railway stakeholders might consider adapting the statistical methodology in IEEE 1366 [30] or simply when 10% of the passengers are affected during a day.

**Table 4.** Electric power performance indicators and the loss scenarios they reflect and possible analogous indicators in the railway system. In the left-most columns the railway stakeholders to whom the performance indicators are most relevant, are indicated by ‘x’. The relevance to train traffic control personnel is indicated by ☺ or ☹. The rows with thick vertical borders are further described in the text.

<table>
<thead>
<tr>
<th>Pass T</th>
<th>TO C</th>
<th>IM C</th>
<th>Indicator</th>
<th>Loss scenario</th>
<th>Possible railway indicator and/or loss scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>x x x x</td>
<td>ASAI - Average service availability index (i.e. fraction uptime at outlet)</td>
<td>No power in outlet.</td>
<td>Fraction uptime as the time the passengers (train) move with planned speed; fraction uptime of a certain location on the line.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x x</td>
<td>SAIDI - System average interruption duration index</td>
<td>Loss of power, total customer base (time).</td>
<td>Downtime per passenger, during a given time interval.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x x</td>
<td>SAIFI - System average interruption frequency index</td>
<td>Loss of power, total customer base (#).</td>
<td># trains hindered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x x</td>
<td>MAIFI - Momentary average interruption frequency index (maximum 5 minutes).</td>
<td>Short interruptions are severe, too.</td>
<td># failures resulting in maximum 5 minutes downtime, calculated with respect to passengers, trains or IM’s items. ☺</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>MAIFI_k - Momentary average interruption frequency index</td>
<td>Interruptions in rapid sequence are together considered as bad as the last, longest, interruption.</td>
<td># gradual failures such as obstructing wagon door which finally fails, turnout needs gradually more attempts to be thrown and finally fails. ☹</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>CAIDI - Customer average interruption duration index</td>
<td>Loss of power and loss of energy, per affected customer.</td>
<td>How much each passenger on a journey has been delayed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>CAIFI - Customer average interruption frequency index</td>
<td>Loss of power and loss of energy, per affected customer. # each passenger has been delayed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>CTCAIDI - Customer total average interruption duration index</td>
<td>Loss of power and loss of energy, per affected customer.</td>
<td>How much each passenger that has been delayed any time, has been delayed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>Value of lost energy to customer.</td>
<td>Damage to equipment, property or production.</td>
<td>Passenger’s delay cost.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>CEM_m - Customers experiencing multiple interruptions</td>
<td>Frequent loss of power for some customers.</td>
<td>Fraction of passengers that are delayed more than a certain number of times</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8. Classification of Availability Indicators

It is suggested to label availability indicators with respect to their scope:

- **Stakeholder**: stakeholder and level relevant to the indicator. The level might range from availability of the entire system, of concern to e.g. end customers, to availability of a tiny component, of concern to e.g. IM’s component procurers.

- **External disturbances**: to which extent the effects of disturbances external to the system, e.g. sabotage, are included. In the military, such disturbances are included in the Design adequacy factor.

The indicators might be further classified with respect to how they are calculated as belonging to one of the categories:

1. As a function of punctuality or delay.
2. As the quotient Obtained capacity / Planned capacity. It might be helpful to define the Obtained capacity with respect to the Planned capacity or vice versa.
3. As the product Operational readiness · Mission reliability · Design adequacy. Analogue factors in the railway system might be Departure punctuality, Underway punctuality and Arrival punctuality, the last two defined as conditional probabilities.
4. As a frequency. As opposed to indicators of categories 1-3, such an indicator does not usually lie in the range [0, 1]. An example is the number of inspection remarks per year (in Table 1).

The majority of the indicators in Table 1, Table 2 and Table 4 belong to category 1 or 2.

9. Conclusions

Making analogies from the electric power industry is rewarding, since stakeholder structure and production characteristics are similar to those of the railway system. Many power industry indicators directly concern the power customer, from which passenger-related indicators might be adapted. As different indicators might give different rankings of causes for unavailability, the choice of indicators is important.
Availability indicators encountered in the railway literature, but not applied in Sweden, concern slack, travel time variation and wagons. New railway indicators found by analogy to the power industry include indicators that concern passengers to a large extent, traffic work not delivered and how to subtract the effect of adverse weather.

Most of the availability indicators are calculated based on punctuality or on capacity. In order to be able to compare availability, the definitions in both cases need to be further standardised. The degree of abstraction of the timetable is important. The fact that unpunctual trains have a value, but not as high as punctual trains, advocates availability being defined relative to punctuality. In other words, availability concepts in railway should be defined relative to or in relation with punctuality.

It is suggested to categorise availability indicators with respect to concerned stakeholder (e.g. TOC), inclusion or not of external disturbances (e.g. effects of sabotage) and how they are calculated (one of the alternatives 1-4 given above).

To summarise, there are many potential useful railway availability indicators. However, these examples are not developed by the use of real life data, and therefore there is a need to study these indicators in real life. Furthermore, potential indicators should be tested by individual stakeholders of the railway.

Acknowledgements

The author would like to thank Banverket and the EU regional fund for financial support. The author is grateful to Professor Per Anders Akersten at Växjö University, Dr Peter Söderholm and Professor Uday Kumar, at Luleå University of Technology for useful suggestions on earlier versions of the manuscript. Thanks are also due to Professor Ulf Sandberg, BU Services, Vattenfall, for sharing his knowledge in electric power and the anonymous referees, whose comments helped improve the paper.

References


**Birre Nyström** received his MS in industrial engineering and ergonomics and his PhD in operation & maintenance from Luleå University of Technology in Sweden. His research interests include railway maintenance and experience feedback.