information

Article

Delay Indices for Train Punctuality

Enrico Denti *,† and Luca Burroni †

Department of Computer Science and Engineering, Alma Mater Studiorum-Università di Bologna, I-40136 Bologna, Italy; luca.burroni@studio.unibo.it

* Correspondence: enrico.denti@unibo.it; Tel.: +39-051-209-3015

† These authors contributed equally to this work.

Abstract: Indicators of expected quality of service in public contracts are often based on some kind of "punctuality", usually defined in terms of the percentage of trains arriving at the final destination (and/or at intermediate significant stops) within a given delay. Passengers, however, tend to use the word "punctuality" with a more general meaning, mostly as a synonym for expected delay at their own destination, and especially in case of commuters are much less tolerant of even smaller delays than train operators would normally allow. In particular, measuring the delay only at the final destination is perceived as largely inadequate, leading to underestimation of the actual percentage of late trains, and in turn undermining passengers' trust in official performance statistics. In this paper, we take the passengers' perspective, introducing a family of delay indices called *D-indices* aimed at capturing the overall performance of a train "as a whole", taking into account both the delays at the sampling points and the mutual location and order of such sampling points. In this paper, all indicators have the physical dimension of time in order to be easily replaceable to other delay measures. We first present typical approaches and definitions of punctuality in the literature, then introduce D-indices while exploring their features, pros and cons, and relevant properties. We validate and discuss our approach by comparing this model with existing approaches both theoretically and by comparison with selected datasets consisting of about one hundred trains transcribed over the last three years.

Keywords: punctuality; delay indices; performance indicators; train punctuality

check for updates

Citation: Denti, E.; Burroni, L. Delay Indices for Train Punctuality. *Information* **2023**, *14*, 269. https://doi.org/10.3390/info14050269

Academic Editor: Michele Ottomanelli

Received: 6 February 2023
Revised: 21 April 2023
Accepted: 29 April 2023
Published: 1 May 2023

1. Introduction

In an ideal world, trains would always operate perfectly on time. In the real world, of course, train operations are actually subject to many factors which may cause small or larger delays; thus, the overall quality of train service is a complex issue from the technical, economical, and user experience viewpoints.

From the technical viewpoint, *train delay* can be defined as "the deviation from a scheduled event or process time of this train" [1]; there, *punctuality* is seen as an aggregate measure, defined as "the percentage of the trains arriving (or departing, or passing) a location with a delay less than as certain time in minutes" [1,2].

More generally speaking, as highlighted in [3], delays and punctuality are different in that delays refer to measures in time units, while punctuality is expressed as a percentage.

From the microeconomic/performance accounting viewpoint, the definition of expected quality of service in public contracts can take into account "punctuality indicators" [4,5], typically defined in terms of the percentage of trains arrived "on time", i.e., within a given delay, at stations, which can mean either at their final destination or also at intermediate "significant" stops. The delay threshold varies from country to country [6] (e.g., 3 min in the Netherlands [7], 5–10 min in the UK [6], 6–10 min in Belgium [8], 5 min in Germany [6], and 5–15 min in Italy for commuter and long-distance trains, respectively [5,9–12]; for a full list, readers are referred to [13,14]). Measures can also be based on the "average delay per train" [4,15], on the standard deviation of the travel time [4], and on other metrics [13]. Indeed, punctuality is "claimed to be one of the most important quality indicators" [3] in railways, and is crucial for customer satisfaction. Such performance targets, especially the commuter ones, can refer to specific lines in a given region, or possibly to the typical weekday, or to peak hours only. In certain cases, a bonus/penalty system is introduced for awarding/fining train operators whose indices meet/do not meet the expected standards, in which case forms of compensation can be established in favor both of the public commissioner and/or final customers.

Widening the discourse, analogous considerations hold for schedule-based bus operations, where a service is often considered on time if "it runs between a certain amount of minutes early and minutes late" [16], and punctuality is usually measured at the origin and at several stops along the route, in particular, at "regulation stops" [17], i.e., key stops along the route. Suitable punctuality indicators are defined for business models purposes [18], and a bonus/penalty system can be established [17] based on whether operators exceed or fall short of a pre-defined standard.

Customers, however, tend to use the word "punctuality" with quite a more general meaning, mostly as a synonym for "expected delay" [19,20] experienced at their own destination on the specific lines and trains they normally use, borrowing the semantics from their everyday life where "punctual" means no or little delay on a given clearly-identified event. Accordingly, the customers' viewpoint is both "multidimensional, subjective" [13], and quite a lot *less forgiving*: indeed, passengers experience even small delays of a minute or so quite negatively, which is much stricter than the 5-minute allowance adopted by many train operators in considering a train "on time". Delays of 30 min or more exacerbate passenger frustration, with notable policy implications for train operators [21]. Notably, the passengers' viewpoint is *situated*, i.e., specifically focused on the passengers' own experience, leading to a possible difference in perceptions of delays and punctuality with respect to the statistical definition of punctuality that is traditionally adopted by train operators and infrastructure managers, which is rather meant to provide a comprehensive view of performance for professional use (including line planning and timetabling [13,22], timetable stability analysis and optimisation [23], delay management [24], simulation [23], and performance evaluation in terms of reliability and availability of service [4,6]). In fact, the average delay has been found to be inconsistent with travelers' preferences [4].

In this paper, we focus on the performance accounting viewpoint, yet from the *passengers' perspective*, which calls for "conformity with observable behaviour" [25]. In this context, measuring train delays only at the final destination – a long-standing habit in the field – is often misleading. This habit was rooted in the professional context, especially for long-haul trips, and was later extended to all kinds of trips, including short-haul and commuter lines; however, it can lead to possible mismatches between the officially reported data and passengers' feeling, thereby undermining their trust [19]. The possible alternative of measuring delays at selected intermediate stations may appear to be unsatisfactory as well, both because of the arbitrariness of the number, location, and importance of such reporting points and most importantly because it does not provide a single comprehensive indicator that can be interpreted as "the performance of that specific train" as a whole. Of course, one could combine the set of different measures in several ways, e.g., by averaging the measured delays to provide a single number; however, apart from the arbitrariness of any combining formula, an important aspect is to preserve a clear and generally-acknowledged passengers' perception of number significance. In principle, what is needed from the passengers' viewpoint is a "natural" way to capture their daily experience.

When speaking about "punctuality" and punctuality indicators (e.g., [13,23]), the purpose is typically to obtain a comprehensive statistical measure of a set of trains in a given time period and/or line(s), often providing percentage probabilities that a train will arrive with a given delay. This is the purpose of the railway metrics defined, e.g., in [26], where the focus is on the punctuality *of the railway*, measured by taking the "lateness" of each train, in principle, at each intermediate station and at termination, providing equal weighting to each recorded stop. In this study, instead, the focus is on trying to capture the behavior of a *single* train as experienced by its passengers, with a single number with the physical dimension of time, which can then be used for subsequent elaborations.

Accordingly, in this paper we introduce a set of delay indices (*D-indices* henceforth) precisely aimed at this purpose. The basic idea is that such a number could replace the unsatisfactory measure at the final destination (or any other) in potentially any kind of subsequent elaboration related to the perceived quality of service, be it the synthesis of a statistical performance indicator over a large number of trains in a given period for a given line, a microeconomic model, or others. As discussed in the next sections, a family of indices is defined, rather than a single one, in order to enable the train journey to be observed from (slightly) different standpoints depending on the purpose of such observation, whether one is mainly interested in the train performance per se, or in weighting it by the number of passengers or other aspects, in comparing the train with other similar ones, in highlighting the presence of excessive recovery time, etc.

Conversely, it is of key importance to highlight what this paper is *not* about. Intentionally, our goal is not to face any microeconomic or technical aspects, such as defining general economic evaluation/performance indicators, estimating/modeling train delays to perform statistical analyses (e.g., distribution, parameter estimation, etc.), or infrastructure modelling or simulation, as is commonly done, e.g., in the timetable design process to avoid/minimise conflicts; we forward interested readers to [4,13,23]. Accordingly, concepts such as dwell time, recovery times, extra times, and buffer times, which pertain the train operator and the infrastructure manager in their design and development activities, are outside the scope of this paper, and the same holds more generally for every study or activity performed a priori before the train service is established.

Instead, we position ourselves ex post, trying to capture the quality of the travel experience from the passenger, especially commuter, perspective. This perspective has gained growing interest in the last 15-20 years, both in official reports from train operators or customer associations (e.g., [19,20,27]) and in notable research papers (see for instance [4,7,13,20,28,29]).

The rest of this paper is structured as follows. Section 2 introduces the basic terminology and related works, then Section 3 provides a short overview of the typical approaches used to measure train delays, highlighting their pros, cons, and weaknesses. In Section 4, D-indices are introduced and thoroughly discussed with suitable examples; a number of relevant properties are discussed in Section 5. Section 6 is devoted to the validation of the proposed approach, discussing the results obtained in characterising suitable sets of trains via D-indices with respect to other possible alternatives. Finally, a discussion and conclusions are provided in Section 7.

2. Delays and Punctuality: Background and Related Work

According to [30], quoting [31], train punctuality is "related to trains running according to schedule", while [32] refers to punctuality as a "numerical measurement (...) part of trains arriving on time to the stations." Olsson [33] reports that "in most cases, this is measured to the terminus, but in some cases, it could also be measured at intermediary stops." In fact, in [34], punctuality is measured at any stop.

Of course, this requires defining in which circumstances an arrival is counted as a delay. Indeed, one of the basic measures used to classify a train as "on time" or "late" has long been the delay at the final destination. On top of this basic value, punctuality indicators are usually defined in terms of the percentage of trains arriving within a given delay, which is typically lower for short-haul and commuter services (e.g., 3–5–6 min, depending on the country) and larger for long-haul services (e.g., 10–15 min) [13,14]. This approach is both easy to understand and somewhat natural from a train operator's viewpoint, as the goal is to guarantee the overall stability of the timetable from the operational perspective, which clearly depends on rolling stock being available for the next train at the expected time in the expected location to ensure that the delay of one train is not transferred to the next.

However, both in commuter lines and long-haul journeys, only a minority of passengers actually makes the whole trip from the train origin to the final destination, with the possible exception of short shuttle services with few or no intermediate stops. Indeed, most passengers get on and off at intermediate stops, especially if the train serves one or more major hubs. For such passengers, delays at intermediate stops do matter, as they cause delays in reaching their work site, school, business meeting, etc. [4]; yet, if the train recovers from the delay in the subsequent part of its journey, such delays can virtually disappear in the statistics and in the official quality of service reports, undermining passengers' trust in official punctuality figures and, overall, in the service itself [19]. This is why, as pointed out by U. Martin in 2008 [6], the reporting points where punctuality is recorded should be distributed over the whole network (at least at major intermediate stations) and not limited to terminal stations. Indeed, more recent approaches [26] aim to measure the delay of each train, at least in principle, at each intermediate station, other than its final destination.

In schedule-based bus operation, a service is usually considered on time if it runs between a certain amount of minutes early and minutes late (in the USA, commonly between 1m30s early and 5m late per bus stop [16], typically measured at the origin and at several stops along the route), in particular at the so-called "regulation stops" [17] (short early arrivals (30s) are often considered on time as well). Accordingly, punctuality can be defined as "the share of buses departing within a certain time window with respect to the planned timetable from several key stops along the route" [17].

Over the last fifteen years, awareness has been growing around the importance of taking the passengers' viewpoint into account. Among the major train operators, Infrabel [8] explicitly states the idea of measuring punctuality "during the entire train journey", namely, "at 95 strategic measurement points on the network", to consider the case when a train succeeds in eliminating a previously-accumulated delay and arrives on time at its final destination. Train operators in UK and Sweden used to consider the "average delay per train" as a valuable indicator [4,20]. Other authorities (e.g., [5]) establish that the delay be evaluated at "relevant" intermediate stations, where the relevance is based on a pre-defined classification that takes into account the number of served passengers, interchange stations, provincial capitals, etc.

From a different perspective, the comprehensive study by Transport Focus (the operating name of the Passengers' Council in the UK) [19] investigated the passengers' viewpoint more in detail. Essential findings include the fact that punctuality is a vital prerequisite for building trust between passengers and a train company, and that the passengers' concept of a train being 'on time' is much stricter (about one minute) than most industry standards (usually 5 min, often 10–15 min for long-distance trains; see [13] for a comprehensive table). Moreover, passenger satisfaction appears to decline quickly for each extra minute of lateness, especially for commuters. As a consequence, quoting [19], "a significant degree of passenger satisfaction is lost when trains are officially on time according to the industry measure, but late in passengers' eyes." Not surprisingly, passengers observed that "punctuality should be measured at all stations, not just where a train terminates." The passenger survey conducted in [4] confirmed that "the common practice of using average delay as performance indicator is misleading, if the aim is to reflect travelers' preferences." In fact, more recently the passengers' perspective is being taken into account in railway timetabling [29].

Another interesting finding is that the actual indicator to be measured should possibly be the number of *passengers* arriving on time, rather than number of trains. This suggests that in order to find the actual picture of the inconvenience caused to people, delays should somehow be weighted by the number of passengers who experience the delay itself. This is why one of the D-indices is defined by taking this aspect into account.

Finally, passengers in [19] objected to the standard practice of adding extra time into the timetable on approach to the destination station, seen as a way to "adjust" the train performance rather than an effective technique to absorb delays and prevent a delay from being transferred to the subsequent trains. This point further stresses the need for a kind of measure able to see "the whole picture" of a train journey from the passengers' eyes.

Technical studies take a different viewpoint, and are mostly rooted in the operations research, transportation, and microeconomic areas. There, the train delay is typically defined as "the deviation from a scheduled event or process time of the train" [1], whereas punctuality is "the percentage of the trains arriving (or departing, or passing) a location with a delay less than as certain time in minutes" [1,2]. All the relevant technical aspects of train operation are considered, from delay distribution models [28], line planning and timetabling [22], timetable stability analysis and optimisation [23], delay management [24], simulation [23], and performance evaluation in terms of reliability and availability of service [4,6,13]. In [14], a systemic classification of delay causes was proposed based on a comprehensive literature review which includes the main definitions of punctuality and reliability, together with delay thresholds in many European countries. In [35], the focus is on punctuality reporting systems, aimed at performing an in-depth analysis of the delay causes and of the train run.

For timetable planning purposes, in particular, the train movement is first carefully modelled taking into account its acceleration/deceleration and the allowed speed on different sections of the railway line based on its characteristics in order to find the pure running time, then the dwell time at stations is taken into account and recovery times are further added to allow for small delay compensation (usually, a percentage of the running time). Train paths are then modelled (which in turn require proper infrastructure modelling) and proper separation schemes (e.g., block sections, protected zones) and infrastructure (e.g., signalling) are considered to guarantee the necessary train separation [23]. Buffer times are added to prevent the transmission of small delays. Train scheduling is then performed, usually in a computer-assisted way. Finally, train running times are carefully estimated for different train configurations. Clearly, such a complex process strongly builds on optimisation techniques, the desired result being a *robust* timetable, that is, a timetable which does not suffer considerably from small perturbations [23]. Other aspects, as mentioned above, range from timetable stability analysis, which includes, among others, the study of delay propagation, to simulation tools, which allow for a very cost-effective way to check hypotheses, verify interactions, highlight possible conflicts, etc., not to mention optimisation of energy and cost issues.

As pointed out in the introduction, most of this huge corpus of knowledge is fundamentally exploited ex ante to design and implement the train service. Following the passengers' viewpoint, our approach is basically ex post view, which observes the train's behaviour when the service has already been designed and implemented, judging its quality in terms of "distance" from the published schedule; the closer, the better. Recent approaches, such as [26] in the UK, have actually aimed to monitor train behaviour for quality of service purposes, defining a notable set of metrics and potentially measuring the delay of each train at all stations in terms of whether such a delay is within a given threshold. These data are then exploited to derive the percentage of "on time" trains, i.e., the punctuality, *of the railway*; in principle, the same could obviously be done for single lines of interest.

3. Typical Approaches

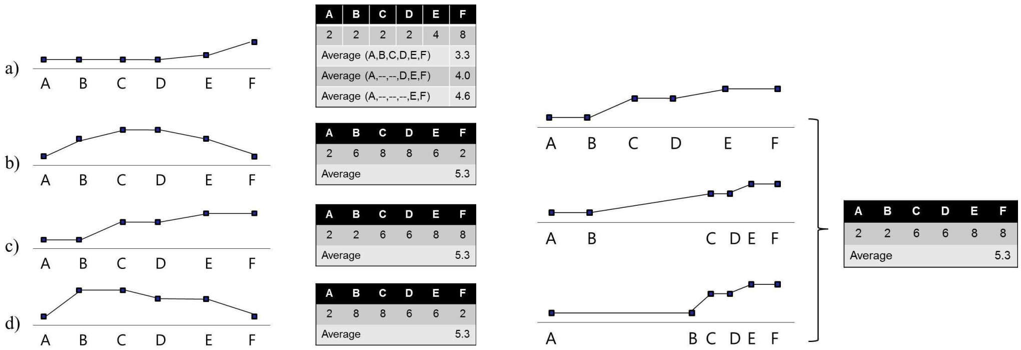
As highlighted above, a long-used approach for building performance indicators is to measure the delay of trains at their final destination, then use these data to build punctuality statistics. The threshold within which a train is considered "on time" is usually lower for commuter and regional services than for long-haul and inter-city services, the first ranging from 3 to 6 min and the other from 10 to 15 min. Passengers, however, are much less tolerant, and tend to consider their trains "on time" within much stricter margins of about one minute or thereabouts.

Of course, measuring the delay of a train at its final destination is quite an approximate way of evaluating its quality of service, because *(a)* a train might either arrive nearly 'on time' at the final destination despite performing poorly in the earlier part of its journey thanks to suitable recovery time in the schedule, extra time, or even dwell time at intermediate major hubs; conversely, *(b)* it might be counted as 'late' if something delays it in the latter part of its journey, despite good performance in all the previous part of the journey. The first situation is perhaps most typical in commuter, regional, and short-haul services if there are one or more major hubs in the middle of the journey; the second can equally affect regional and long-haul trains, which tend to connect major cities end-to-end.

A natural idea to improve on the final-destination approach is to average the delays measured at a set of intermediate reporting points (intermediate stops, stations, etc.); unsurprisingly, this is precisely what emerged following a 2011 dialog between passengers and a company inspector [27]. At first sight, an average (possibly with its standard deviation) appears "naturally" more indicative of the whole train journey than a single measure; yet, the result reflects mainly the available samples rather than the actual progress of the journey. This weakness is noted in [4,13] with respect to a set of trains (comparing a situation where "all trains are one minute late, and two out of ten are five minutes late") rather than the samples of a single train, as in our case. The point is that the average is *(i)* highly sensitive to the presence of more/less samples in the journey, *(ii)* does not consider the specific locations where the samples are taken (which reflects the intermediate stop sequence), and *(iii)* is insensitive to the corresponding inter-station distances, which indirectly influence the delay inter-dependence.

To illustrate such aspects, in the figures throughout this paper we adopt a kind of distance–time diagram—henceforth called the *delay diagram*—to map the registered delay at each sampling point (train stops or other points of interest along the train line). The diagram reports the sampling points on the X-axis (possibly reflecting their mutual distance, as described below), and the corresponding delays on arrival on the Y-axis; in this way, delays are not merely "counted" in a formula, instead being directly related to the expected train journey (the sequence of sampling points) and schedule (a perfectly on time train would have its graph fully overlapping the X-axis). Actually, many kinds of time–distance diagrams are of standard use in railway operation and planning; while an overview is outside the scope of this paper, interested readers can refer to [23] for an in-depth discussion.

Figure 1(left (a)) shows the first issue, i.e., that the number of sampling points may lead to different average values for the same journey. This suggests that such numbers do not represent the *journey*, only the *sampling points* selected during the journey. It should be noted that the average results may be different even if the delay remains the same throughout the whole section A–D, while a reasonable indicator could be expected to not be influenced by the presence/absence of intermediate samples such as B and C between A and D.



A	B	C	D	E	F
2	2	2	2	4	8
Average (A,B,C,D,E,F)					3.3
Average (A,--,--,D,E,F)					4.0
Average (A,--,--,--,E,F)					4.6

A	B	C	D	E	F
2	6	8	8	6	2
Average					5.3

A	B	C	D	E	F
2	2	6	6	8	8
Average					5.3

A	B	C	D	E	F
2	8	8	6	6	2
Average					5.3

A	B	C	D	E	F
2	2	6	6	8	8
Average					5.3

Figure 1. Critical issues defining punctuality as the average of delays. (**Left**): delays at stations A–F, with corresponding average values. (**Right**): delays plotted in proportion to the inter-station distance.

Figure 1(left (b–d)) illustrate the second complementary issue: different delay sequences representing quite different journeys may result in the same average if the sampled delays are numerically the same, regardless of *where* they were registered. In other words, such numbers lack *situatedness*, representing only the *unordered set* of the sampling points.

Figure 1(right) illustrates the third issue: the same delays, even when registered in the same locations and sequence, may lead to the same average *independently of the mutual placement* of the measurement stations, while in fact, for physical reasons, the more the sampling stations are closer to each other, the more delays are inter-dependent, because the registered delay tends to be similar due to sampling the same (or a similar) delay multiple times. Conversely, the farther the sampling stations are from each other, the less dependent the sampled delays are, representing a larger part of the train journey. This suggests that a passenger-oriented indicator could take into account the *mutual location* of the sampling points, possibly in terms of sequence and mutual distance other than the pure delay values, before combining this information in an effective and self-coherent way.

More recent approaches, as mentioned in Section 2, are based on counting the number of trains arriving within a (tight) threshold at (all) stations, and use such data to derive the percentage of "on time" trains. This approach is widely adopted and very effective, as it makes it possible to sample many trains at several sampling points in a practical way; yet, it is somehow "digital" in that a train at each station/sampling point either fits the threshold or does not (although this aspect can be partially compensated by taking tight thresholds). The above-mentioned lack of situatedness applies as well, as the location of the sampling stations does not influence the result.

In principle, all approaches based on taking frequent samples, possibly obtained by real-time information on the train journey and hence actually interpolating the delay diagram, could provide an analogous information base; the average of many samples would then accurately describe the train behaviour. However, such (big) data would likely be available mainly to train operators, not directly to passengers or the general public, which could possibly reproduce the "undermining trust" issue between official data and passengers' feelings reported in [19], as such big data elaboration would likely be out of reach for passengers, commuters' associations, etc.

Moreover, all the above approaches do not explicitly deal with possible early arrivals, which are normally counted as "on time" trains, even though early arrivals represent deviations from the expected schedule.

The D-indices, introduced below, aim to achieve a similar result from a set of sampling points (in practice, stations and intermediate stops) that are commonly available to the general public, combining them with the notion of *situatedness*, that is, relating them to the physical mutual location of the sampling points, and explicitly dealing with *early arrivals*, providing a set of views that aim to capture different aspects of train performance.

4. D-Indices

Following the above considerations, an indicator aimed at actually representing the train journey from the passengers' perspective should not only measure the delays at (several) sampling points, it should combine them in such a way that the result is:

- closely related to the measured delays (as well as to possible early arrivals);
- easily understandable to passengers;
- *situated*, that is, somehow related to the physical location of the sampling points.

The latter requirement is a distinguishing feature of the proposed model and is at the base of the contribution of this work. Accordingly, the above requirements can be refined as follows:

- delays should be sampled at a multiplicity of sampling points;
- the sequence and mutual order of such sampling points should be taken into account in terms of their inter-distance.

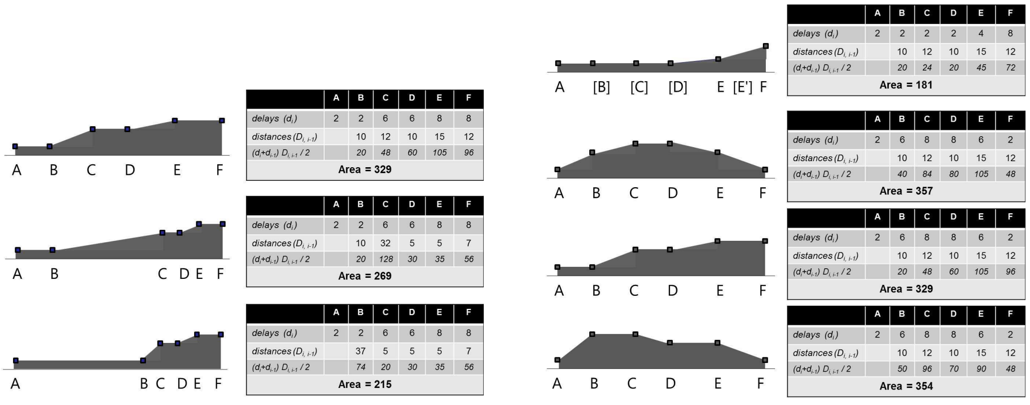
Delay diagrams make a step in this direction, *provided that the sampling points on the* X-*axis are scaled in proportion to their mutual distance*. In this way, items on this axis are not just categories, as in conventional histograms, they are measurement points plotted according to a kind of scale based on their "distance".

Moreover, such distance need not necessarily be the obvious spatial distance (e.g., in kilometres) among the sampling points, but could be a *time-based* distance, such as the scheduled travel time; the X-axis would then be labelled in time units (e.g., minutes) and the sampling points would be mapped based on the expected time of arrival/passage of the train at that point. While the first alternative is clearly absolute, as the physical location of the sampling points is immutable and is the same for all trains, the second is relative to the planned schedule of each specific train, suggesting an interesting dimension to be explored.

4.1. The Basic Idea

Moving on from the above considerations, the basic idea beyond D-indices is to assume *the area* resulting from the interpolating line in the delay diagram as representing the overall delay accumulated by the train during its whole journey, in a *schedule-proportional way* (Figure 2a). In principle, other interpolation curves could be used; however, a linear approach seems preferable both for simplicity and for the physical nature of delays, as if sampling points are reasonably close to each other the inherent inertia of delays due to the physical nature of train movement makes sudden changes rather unlikely.

In order for the resulting index to have the physical dimension of time, the area is then normalised by the total travel distance in the formal definitions provided below. By construction, this area inherently provides each travel portion with its own "weight", bringing about several benefits; in particular, it helps to address issues *(i)*–*(iii)* in the average approach. In fact, because the sampling points are in sequence and to the scale, if these are closer then the enclosed area is smaller, while if they are farther then the area becomes larger, thereby inherently taking into account and compensating for the number and mutual location of the sampling points themselves.



	A	B	C	D	E	F
delays (d_i)	2	2	6	6	8	8
distances $(D_{i,\,i\text{-}1})$		10	12	10	15	12
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*20*	*48*	*60*	*105*	*96*
Area = 329						

	A	B	C	D	E	F
delays (d_i)	2	2	6	6	8	8
distances $(D_{i,\,i\text{-}1})$		10	32	5	5	7
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*20*	*128*	*30*	*35*	*56*
Area = 269						

	A	B	C	D	E	F
delays (d_i)	2	2	6	6	8	8
distances $(D_{i,\,i\text{-}1})$		37	5	5	5	7
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*74*	*20*	*30*	*35*	*56*
Area = 215						

(a)

	A	B	C	D	E	F
delays (d_i)	2	2	2	2	4	8
distances $(D_{i,\,i\text{-}1})$		10	12	10	15	12
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*20*	*24*	*20*	*45*	*72*
Area = 181						

	A	B	C	D	E	F
delays (d_i)	2	6	8	8	6	2
distances $(D_{i,\,i\text{-}1})$		10	12	10	15	12
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*40*	*84*	*80*	*105*	*48*
Area = 357						

	A	B	C	D	E	F
delays (d_i)	2	6	8	8	6	2
distances $(D_{i,\,i\text{-}1})$		10	12	10	15	12
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*20*	*48*	*60*	*105*	*96*
Area = 329						

	A	B	C	D	E	F
delays (d_i)	2	6	8	8	6	2
distances $(D_{i,\,i\text{-}1})$		10	12	10	15	12
$(d_i+d_{i-1})\ D_{i,\,i-1}\ /\ 2$		*50*	*96*	*70*	*90*	*48*
Area = 354						

(b)

Figure 2. (**a**) The basic idea; the area obtained by plotting stations on the X-axis represents the the overall delay accumulated by the train in its whole journey in a schedule-proportional way. (**b**) The area represents the globally accumulated delay independently of both the delay at the final destination and the average of delays.

As shown in Figure 2b, if a train arrives late only at a (limited) subset of stops, a situation in which the delay affects only a small part of the journey, the corresponding area is smaller, providing only a limited contribution to the global figure *even if the delay affects the final part of the journey* (top diagram in Figure 2b). Considering only the delay at the final destination, such a train would be improperly tagged as 'late', making it virtually undistinguishable from a 'really late' train (third diagram). Conversely, if a train arrives late for a large part of its journey, the corresponding area is bigger *even if the delay decreases or becomes zero at the final destination* (second and fourth diagrams in Figure 2b), preventing a poorly performing train from being optimistically counted as 'on time'. Overall, this approach avoids the situation in which a train that performs well for most of its journey is judged poorly due to its behavior only on the final part of the journey, and conversely a train that performs poorly for most of its journey is judged well only because it was able to absorb the delay in the last part of the journey thanks to recovery and extra times incorporated in the schedule.

It is worth noting that because we assume the passengers' viewpoint, *only the delay on arrival at each station is relevant*; an analogous approach is mentioned in [13]: "a departure delay can be recovered and thus bears no impact for users." In train timetable and planning techniques, of course, things are different, and many other factors (e.g., dwell time, departure times, etc.) must be considered [23,29]. It could be argued that due to the dwell time and possible recovery time included in the schedule a train can leave a station with a different (shorter or longer) delay than it arrived; thus, in principle, each train stop should be modelled more precisely as a pair of possibly-distinct delays. However, as noted above, from the passengers' viewpoint a delay on departure is only relevant as long as it is not absorbed before the next stop; otherwise, it all comes down to a shorter trip, with no consequence at all. This is why, for the sake of simplicity, we opted to represent only the delay on arrival; if necessary, delays on departure are indirectly considered by the delay on arrival at the subsequent stops. The consequences of this choice are discussed in Section 4.4.

Based on this idea, a family of indices can be defined which share a set of features and properties yet aim to capture specific aspects and viewpoints, for instance, static-oriented and dynamic-oriented views, absolute vs. relative views, etc., as discussed below. The resulting indices effectively synthesise global information into a single number that retains the physical dimension of time, and as such remains interpretable as a "the train delay", yet in the more comprehensive form of *delay indicator*; as such, it could potentially be used in punctuality statistics or in more complex performance indicators or microeconomic models as a replacement for other measures of delay.

4.2. Basic Indices

The above-outlined framework intentionally leaves a degree of freedom in the kind of 'distance' to be used on the X-axis, accounting for a *family* of indices, namely, *spatial-* and *time-*based indices, yielding absolute vs. relative views, respectively.

Basic indices, defined below, are the simplest form of indices, built upon the plain area underlying the interpolation lines among delays on arrival at the sampling points.

Definition 1. *Delay-Distance (DD_0) basic index*

$$DD_0 = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d_{i-1}) L_i}{\sum_1^N L_i} \text{ or just } DD_0 = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d_{i-1}) L_i}{L_{tot}} \tag{1}$$

where d_i is the delay on arrival at the i-th station and L_i is the distance (in kilometres or miles) between the i-th station and the *(i-1)*-th station. The departure station is conventionally considered the *0*-th station. By convention, the delay d_0 at such a station is either the delay on departure, if the train actually originates there, or the delay on arrival from a previous section of the journey if there is one. The next index functions analogously in the time domain.

Definition 2. *Delay-Time (DT_0) basic index*

$$DT_0 = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d_{i-1}) T_i}{\sum_1^N T_i} \textit{ or just } DT_0 = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d_{i-1}) T_i}{T_{tot}} \tag{2}$$

where T_i is the scheduled travel time between the *i*-th station and the *(i-1)*-th station; the other symbols have the same meaning as above. Note that because T_i represents *travel* times, T_{tot} represents the total travel time net of all dwell times at all stops.

It is worth noting that both formulas interpolate the delay values by simply connecting delay points with each other, which is clearly unrealistic when the sampling points are stations, as trains obviously cannot leave a stop earlier than scheduled; this is why basic indices should be seen only as a proof-of-concept towards more comprehensive definitions, as developed in the next Subsections. In fact, as an extreme case, under these definitions a train arriving early at most/all stations would result in a negative index. As the implications of negative contributions (and indices) are all but trivial in terms of opportunity, representation fidelity, and the related pros and cons, we discuss this issue in depth separately (Section 4.5), introducing *bounded* indices to capture this aspect properly.

Figure 3 illustrates the basic indices compared to the average of delays for the same situations previously shown in Figures 2a,b. No distinction is made here between the DD and DT indices, as the spatial/temporal nature of the 'distance' on the *X*-axis is irrelevant at this stage.

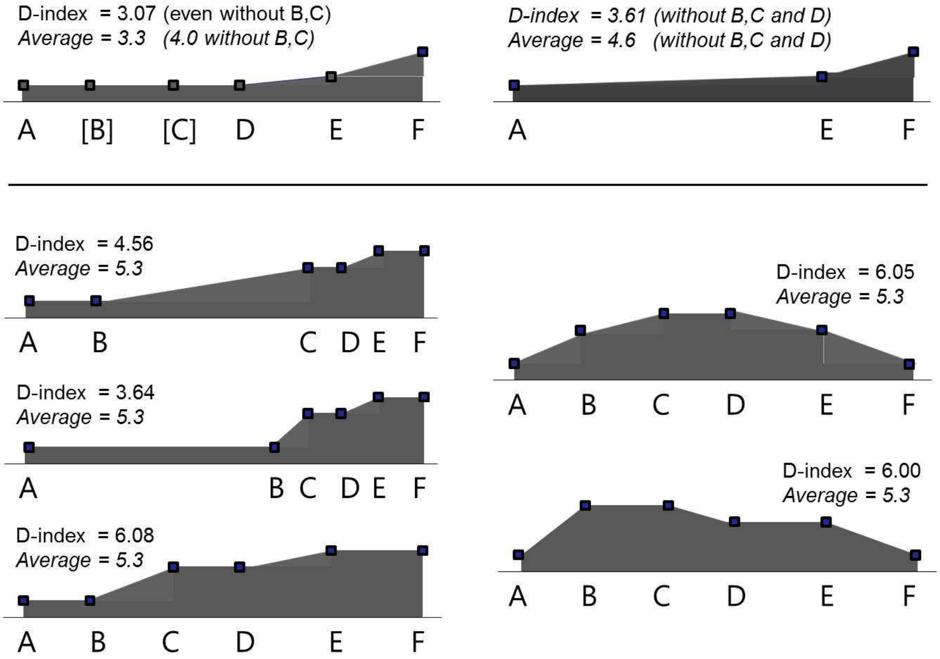


Figure 3. D-indices vs. averages in the above cases: ***(top)*** in the presence/absence of intermediate sampling points, ***(bottom)*** the representation of five different journeys with the same average.

In Section 3, we noted that the average is *(i)* highly sensitive to the presence of more/less samples in the journey, that *(ii)* does not consider the specific locations where samples are taken (which reflects the intermediate stop sequence), and that *(iii)* is insensitive to the corresponding inter-station distances. Issue *(i)* means that the number of sampling point can yield different results for the same journey; in particular, the presence/absence of the intermediate sampling points at B and C, even though sampling the same delay as A and D, influences the average. Issue *(ii)* implies that different delay sequences, representing different journeys, can result in the same average if the sampled delays are numerically the same, independently of where they were registered. Issue *(iii)* concerns the benefit of taking into account the inter-distance of the measurement stations in order to indirectly account for the interdependency of delays. As shown in Figure 3(top left), the D-indices do not suffer from these issues; as long as they sample the same delay as the adjacent points, the presence/absence of intermediate sampling points such as B and C is irrelevant. Moreover, if another sample (such as D in Figure 3(top right)) is removed, the index obviously changes (+17%); however, the impact is lower than would be the case when using the average (+40%). Another key feature is that different journeys result in correspondingly different D-indices (Figure 3(bottom)), even when the average is the same.

Thus, the D-indices seem promising in combining the relevant information into a single more comprehensive indicator; the physical dimension of time makes it an easy replacement for the delay at the final destination or the average of delays. Yet, basic indices may result in inaccuracies when modelling specific situations, as discussed below; for this reason, we derive a *family* of indices (Sections 4.4–4.7), providing a multiplicity of viewpoints.

4.3. Spatial vs. Temporal Indices

DD indices adopt a spatial metrics for the sampling points on the *X*-axis; thus, the resulting indicators are “absolute” in the sense that the physical location of the sampling points is immutable and the same for all trains of a given line. This is clearly a natural choice which allows for easy comparisons among different trains on the same line, yet it corresponds to a “static” view in which train performance is evaluated based on static coefficients determined by the intrinsic route characteristics.

DT indices instead adopt a temporal metrics, that is, the distance between the sampling points on the *X*-axis is the time spent by the train to cover that journey according to its schedule. Such a metrics is “relative”, as the coefficients are specific to that train; this approach emphasizes a more “dynamic” view, with train performance evaluated based on what that specific train was supposed to do, rather than a fixed scale. Of course, in periodic timetables adopting the same schedule for all trains in a given category, the DT indices of trains in the same category are mutually comparable, the same as if DD indices were used.

An identical DD index captures the fact that such trains actually showed the same delays in the same points of the same route, providing (at least apparently) the same quality of service; however, it can be argued that this tells only part of the truth, as a slower train is actually allowed a longer time to cover the same distance. As a consequence, it should perhaps “not have been so late” as a faster one under the same conditions. Said differently, the slower train had better opportunities to compensate for its delay than the faster train, but did not (as its DD index is identical); thus, it actually under-performed with respect to the faster one. DT indices are conceived precisely to reveal this issue; by taking the time schedule into account by construction, their synthesis of train behaviour embeds the idea that “good performance means making good use of the allowed time”.

Time-based indices, however, may involve interpretative subtleties. To identify the possible issues, let us first observe that by construction *the faster the train schedule, the lower the area in the time-based diagram*, and conversely, the slower the train schedule, the higher the area. This property, however, does not necessarily transfer onto DT indices; in particular, a train with a larger area (and a larger amount of ‘globally accumulated delay’) may result in a lower DT index than a train with a lower area if its scheduled journey time is more than proportionally shorter than the other.

To better understand the implications of this aspect, let us imagine two nearly identical trains covering the same route and calling at the same stations, showing exactly the same delays at all stations; the slower train results in a lower DT index, (apparently) providing a ‘better’ service than the other. Whether this is a correct representation of the passengers’ feeling is clearly questionable, and depends on the type of service. For long-distance high-speed trains, where passengers reasonably expect little/no delay (somewhat due to the ticket price of such services), the DT index is likely to fit the passengers’ expectation; the faster the train was supposed to be, the more intolerable delays are. For commuter services, things may be different; the longer the trip schedule, the more intolerable the delay becomes for passengers. In this case, the DT index might somehow be misleading.

This dichotomy should not come as a surprise, as any indicator inherently provides its own viewpoint focused on the specific aspects embedded in its definition. The viewpoint embedded in DT indices is that "good performance means making good use of the allowed time"; the passengers' experience being multi-facet and heterogeneous [13], if the viewpoint changes, as in the commuter's case, a different index may be needed to properly capture the new aspect. We address this issue in Section 4.7 by introducing "absolute" indices uninfluenced by the journey time.

4.4. Standard Indices

As noted above, the main limitation of the basic indices is that they interpolate the sample delays by simply connecting delay points with each other, actually adopting the delay on arrival (the right side of each area section in the delay diagram) as representative of the delay on departure (the left side of each area section) for the next area section. This leads to a clear misrepresentation in the case of early-arriving trains, which certainly cannot leave the stop earlier than scheduled. In the extreme case, a train arriving early at most stations would have a negative index.

A more correct model should then allow negative delays *at stations* to appear only as delays on arrival, i.e., as right-hand terms in the formulas (d_i), not as delays on departure, i.e., as left-hand terms (d_{i-1}), where they should be lower-bounded to zero. Note that this correction only concerns train stops, not other possible sampling points en route. The following definitions henceforth replace Definitions 1 and 2 above.

Definition 3. *standard Delay-Distance (DD) index*

$$DD = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d^*_{i-1}) L_i}{\sum_1^N L_i} \quad \textit{or just} \quad DD = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d^*_{i-1}) L_i}{L_{tot}} \tag{3}$$

Definition 4. *standard Delay-Time (DT) index*

$$DT = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d^*_{i-1}) T_i}{\sum_1^N T_i} \quad \textit{or just} \quad DT = \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d^*_{i-1}) T_i}{T_{tot}} \tag{4}$$

where d^*_i is the possibly bounded delay on arrival; if the i-th sampling point is a stop, the (negative) delay is replaced by zero. All the other symbols maintain their previous meaning.

The delay diagram can now contain discontinuities, as the left-hand side of each area section is zero-bounded while the right-hand side is not. Figure 4 illustrates the difference between the basic DD index (top) and the standard DD index (center) for the same train. Analogous diagrams could be made for DT indices.

Philosophically speaking, early arrivals in this context become a sort of 'benefit' which reduces the area (and the index) of a well-behaving train in the same way as late arrivals punish a poorly-behaving train. At first sight, this seems reasonable, yet such a benefit inherently rewards larger schedules in which trains are intentionally assigned longer travel times than they can actually meet. Although train operators could have good reasons to do this, from the passengers' viewpoint a longer-than-necessary travel time could hardly be seen as 'quality'. Moreover, the above situation could be a sign of imprecise schedule planning, which is not the kind of situation one would want to reward.

This is why it might be desirable to neutralise the effect of such early arrivals in selected situations; *bounded indices*, discussed below, are introduced for this purpose.

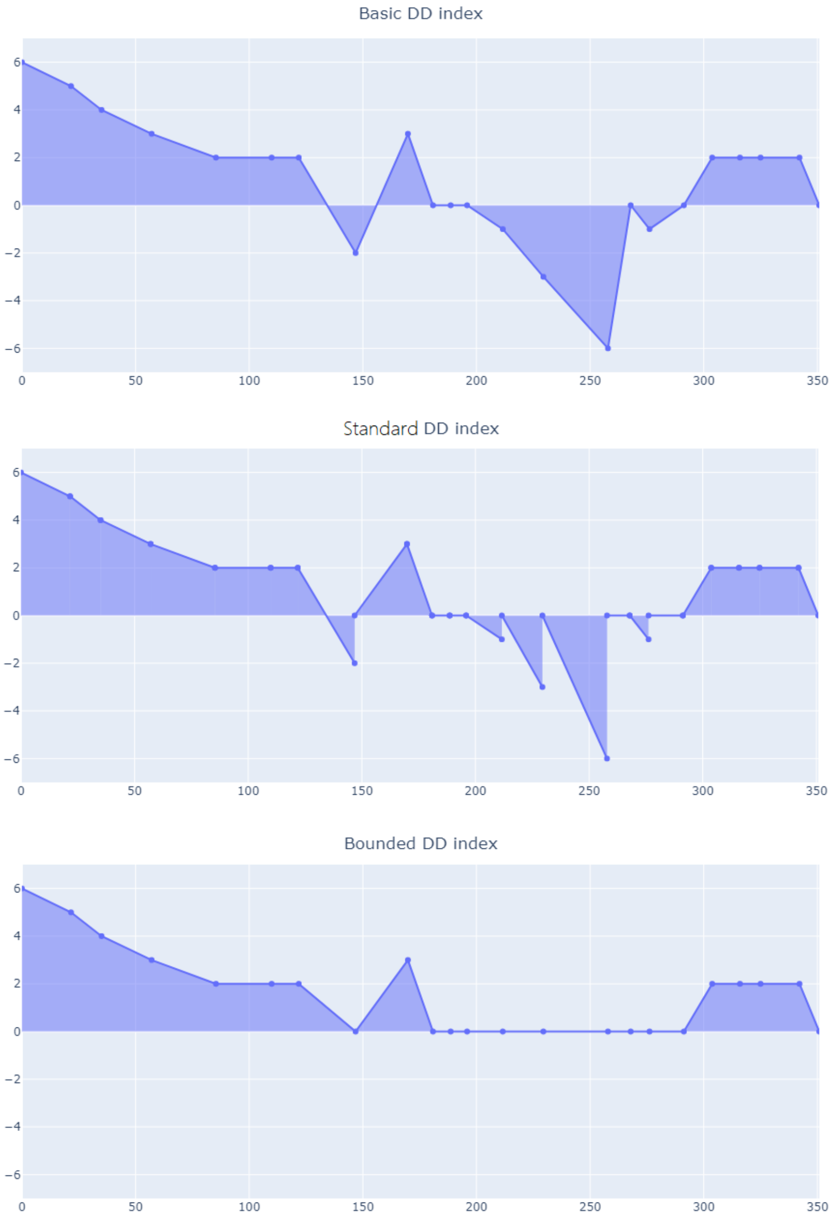


Figure 4. Delay diagram of basic DD, standard DD, and bounded DD indices of a given train.

4.5. Bounded Indices

In order to not reward early arrivals as 'good performance', two further indices can be defined.

Definition 5. *Bounded Delay-Distance (DD*) index:*

$$DD^* = \frac{1}{2} \cdot \frac{\sum_1^N (d_i^* + d_{i-1}^*)L_i}{\sum_1^N L_i} \text{ or just } DD^* = \frac{1}{2} \cdot \frac{\sum_1^N (d_i^* + d_{i-1}^*)L_i}{L_{tot}} \quad (5)$$

Definition 6. *Bounded Delay-Time (DT*) index:*

$$DT^* = \frac{1}{2} \cdot \frac{\sum_1^N (d_i^* + d_{i-1}^*)T_i}{\sum_1^N T_i} \text{ or just } DT^* = \frac{1}{2} \cdot \frac{\sum_1^N (d_i^* + d_{i-1}^*)T_i}{T_{tot}} \quad (6)$$

where all delays, not just the left-hand terms, are now zero-bounded, i.e., $d_i^* \geq 0$ *for any i.*

Figure 4(bottom) illustrates this situation in comparison with the two previous cases. Obviously, negative areas disappear. With this view, systematic early-arrival scenarios or intentionally large (or poorly-designed) schedules are represented by a 'perfect' 0-index with no intrinsic reward at all. Whether to use standard (Definitions 3 and 4) or bounded (Definitions 5 and 6) indices depends on the analysis to be performed and on the issues one aims to highlight; for most situations, bounded indices can be a natural choice due to their inherent interpretation of early arrivals. However, standard indices can be an effective choice if one is interested in capturing the behaviour of the train "as is", with the least elaboration and no intrinsic interpretation of early arrivals, as the reward implicit in admitting delay compensation highlights the actual fact that passengers *did* arrive less late at subsequent stations. Overall, bounded indices are better able to reveal out-of-standard situations and to stress the performance of a train under worst-case assumptions when no compensation for early arrivals is permitted.

4.6. Weighted Indices

As mentioned in Sections 1 and 2 and in [19], delay indicators can be weighted to the proportion of passengers experiencing the delay, providing a measure of the overall impact of such a delay on everyday life. This could be done by weighting the train stops (e.g., assigning higher weights to more important stations and major hubs) and/or specific trains, etc.

Whatever the weighting criteria, adding weights simply amounts to inserting suitable coefficients in the formulas, thereby amplifying certain delays with respect to others, e.g., stating that 5 min late to the hub station H counts the same as, say, 10 min late to stations A or B. A normalising factor is inserted, down-scaling the resulting area in proportion to the applied weights to ensure that the weighted indices remain comparable with the standard ones.

Definition 7. *Weighted D-indices:*

$$WDx = \frac{1}{2} \cdot \frac{\sum_1^N (W_i d_i + W_{i-1} d^*_{i-1}) D_i}{\sum_1^N D_i} \cdot \frac{N+1}{\sum_0^N W_i} \tag{7}$$

where WDx is the weighted version of any of the DD/DT indices, $W_i \geq 1$ are the weights, and D_i are the spatial or temporal distances (i.e., either L_i for DD or T_i for DT), as in Definitions 3 and 4, etc.; the other symbols have the usual meaning.

Weights express the relative strength of each delay contribution; these indices model situations in which train stops contribute non-homogeneously to the perceived quality of service. For instance, if a train is nearly on time to the hub station (actually serving most passengers well) but accumulates delay in other less relevant parts of the journey, suitable weights can prevent its D-indices from being unfairly poor (Figure 5(top right)). Conversely, if the delay is experienced mostly at the hub, impacting the majority of passengers, the resulting D-index becomes worse than in the plain formulas, matching the perceived quality of service (Figure 5(top left)). The difference is evident in Figure 5(bottom), where the delay diagrams for the standard and weighted DD index are plotted together; the hub near km 150 is weighted twice, making the weighted graph (red) larger in that section.

Weights can be chosen in a variety of ways. In principle, the number of alighting passengers would be the ideal; however, as such data are not typically available, less precise yet still indicative weights could be used to capture the station weights indirectly depending on which information is available from transportation and/or local authorities. For instance, a first raw choice could be to simply use the population of the city intended as a generic indirect indicator of its relevance to indicate the amount of involved passengers. A better choice might be the average amount of travelers (commuters) involved in that hub, either generally, or maybe in the time slot where the specific train runs. If available, a more precise alternative could be the average amount of travelers (commuters) on that specific line, again, either generally or in the time slot where the specific train runs. In the end, it depends on available data, as weighted indices are neutral per se with respect to the chosen coefficients.

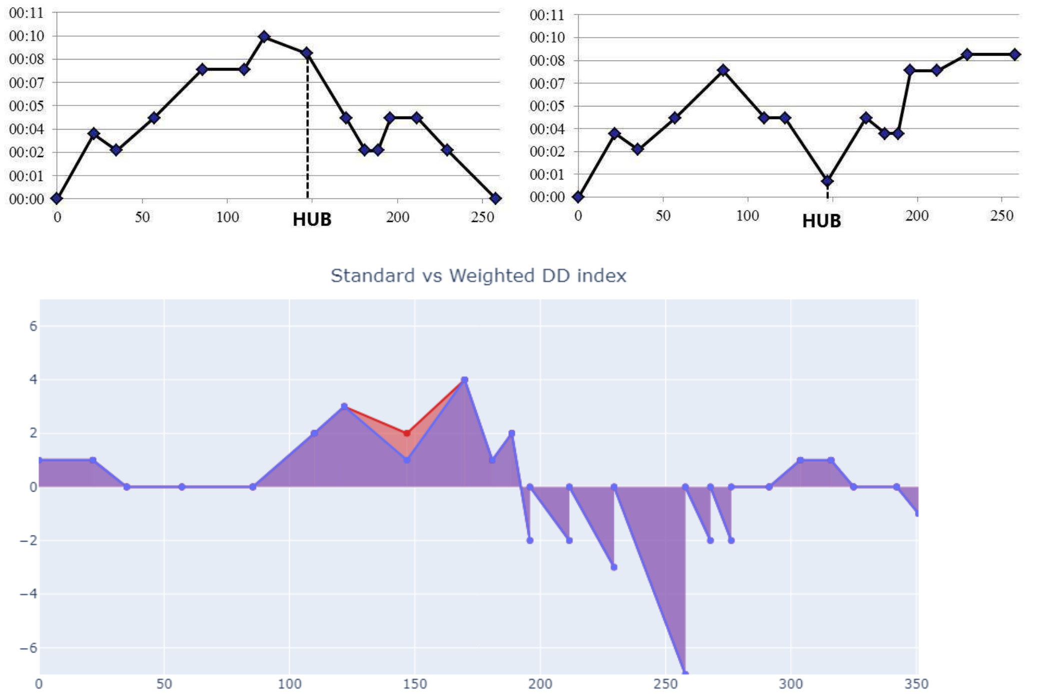


Figure 5. Weighted indices. ***(top left)*** A case where the delay is higher in the hub station, where applying a weight makes the weighted D-index worse. ***(top right)*** The opposite case. ***(bottom)*** Standard (blue) vs. bounded (red) DD indices for a given train; the hub near km 150 is weighted twice.

4.7. Delay-Absolute Time Indices

As discussed in Section 4.3, the DT index might be misleading in situations where it is necessary to evaluate the train performance in a way that is uninfluenced by the journey time (and length); this is what *delay-absolute* time indices are for.

The starting point is that in the D-indices framework the faster the train schedule, the lower the area in the time-based diagram; this is an absolute property uninfluenced by the journey. As such, in principle, it can be seen as an indicator per se, usable precisely in the cases where the goal is to compare two trains operating on the same route *without any normalisation*, as implied in standard DT indices. However, using the plain area as an index would be doubly impractical due to the lack of both a direct physical meaning and the proper physical dimension (as areas are time to the square). As here we are only interested in actual delays (i.e, non-negative delays), it is safe to take the square root of such area, i.e., of the numerator of the DT index, as follows.

Definition 8. *Delay-Absolute time index:*

$$DA = \sqrt{\frac{1}{2} \cdot \sum_{1}^{N} (d_i + d^*_{i-1}) T_i} \tag{8}$$

While the previous indices are representative of the train ‘delay’ from some viewpoint, the latter has no direct physical counterpart, yet fits the purpose of comparing late trains, most likely (though not necessarily) on the same route, to each other *in an absolute way*, as unconsciously done by most passengers.

To compare the DA indices of different trains, one possible approach is to compute their ratio; the resulting values then represent "how much worse" or "how much better" a given train performs with respect to the one assumed as the reference.

For instance, if two trains cover the same line and maintain the same delay at all stations, the standard indices would be equal, while if the two trains have two (slightly) different planned schedules, the delay of the train with a larger schedule is proportionally worse than that of the other, which has less recovery time in its schedule. DA indices make this kind of situations explicit, capturing the difference in actual behaviour and in the provided quality of service. A numerical example can be found in Section 6.3.

5. Relevant Properties

D-indices feature a number of relevant properties, discussed more in detail in Appendix A:

- *proportionality*;
- *compositionality*;
- *insensitivity to sampling point insertion*;
- *comparability*.

Certain properties hold for all kinds of indices, others only for some of them.

Proportionality means that two trains A and B with the same delays, in the same order, and with the same sampling points have the same D-indices (either DD or DT) if their inter-sampling point distances (L_i or T_i, respectively) are in the same proportion. For instance, a train B that is twice as slow as train A (i.e $T_i^B = 2 \cdot T_i^A$ for any index i) but exhibits the same delays has the same DT and DD index as train A. This can be interpreted in the sense that D-indices are capable of 'extracting' the actual delay information from the train journey independently of its spatial/temporal characteristics.

Compositionality derives from the fact that areas in the delay diagram are inherently additive; thus, the D-indices D1 of a train covering route segment A–B and D2 of the same train continuing to route segment B–C can be composed together using the respective fractions of distance as coefficients. In the case of time-based indices, this requires the extra assumption that the timeline is not altered, that is, the presence of the intermediate point (station) B does not change the global trip duration ΔT of the whole route A–C, which remains the sum of the two durations $\Delta T1$ of the route segment A–B and the $\Delta T2$ of the route segment B–C, as obviously occurs for spatial distances.

Insensitivity to sampling point insertion is a consequence of compositionality, and holds under the same assumptions; adding an extra sampling point does not alter the D-index if the delay measured at the new point is either the same of the two adjacent ones or is in their proportion (i.e., geometrically, it fits the line conjuncting them). In other words, if a train maintains the same delay in the whole section A–C, adding the extra measure of the (same) delay at the intermediate sampling point B does not alter the index, as it would for the average of delays. This also holds if the delay either grows or decreases linearly in section A–C, and then has the "proper" proportional intermediate value at intermediate point B. While this might seem a strong constraint to satisfy, chances are that it is not uncommon in practice, at least in short-to-medium route sections, as due to the physical nature of the train movement delays cannot change suddenly with spikes in the absence of major dwell times at an intermediate station. As recalled above, in timetable planning the time flow would certainly be altered by the insertion of an intermediate stop; here, however, the viewpoint is different, as the X-axis simply represents the time flow in terms of scheduled arrival/passing times at specified locations and is all-inclusive as perceived from the passengers' eyes.

Comparability is a consequence of insensitivity to sampling point insertion, and holds under the same assumptions; it means that the performance of local vs. express trains serving the same route can be compared on a uniform basis. For instance, if train T1 (the 'local' train) makes more stops than train T2 (the 'express' train), the above property, as far as it holds, makes their D-indices mutually comparable despite the different number of stops made. Of course, such a comparison is valid only as far the previous property holds, i.e., if the delay is constant or linearly increasing in the section where the local train makes the extra stops. Outside of such scenario, the more the real situation deviates from such an ideal setting, the less meaningful the comparison becomes.

6. Validation

In order to validate the D-indices framework, we transcribed from public sources the behaviour of about one hundred of trains over a period of 3 years between 2020 and 2022. Both regional (commuter) and long-distance inter-city trains (though not high-speed trains) were selected; regional trains were chosen based on the most common working hours and long-distance trains according to a more widespread approach during the whole day. For regional trains, we focused on a line taken with *(a)* several stops (though not all trains actually stopped at all of them) and *(b)* a major city hub in between, ensuring that the typical cases discussed in Section 4 were all present. Validation was organised in two steps:

- first, choosing eight trains (four regional and four long-distance) monitored for over a week;
- then, by monitoring fourteen trains (ten regional and four long-distance) over a two-year period to obtain 126 samples per train (95 actually usable due to lack of data or incomplete data on particular days).

The first dataset was intended to provide a basis for preliminary considerations and comparisons, while the latter aimed to provide a basis for wider data analysis.

Starting from these datasets, the D-indices were compared with both the delay on arrival-based and average-based approaches. Moving from the first dataset, several interesting analyses can be made, discussed below:

- the situation in a typical day;
- the situation in a typical week;
- considerations about the DA index.

Then, in Section 6.4, we consider specific ad hoc trains to place in evidence further situations and corner cases. The wider dataset, discussed in Section 6.5, aims to explore the pros and cons of the D-indices with respect to other possible indicators; to this end, a specific train among the fourteen trains monitored was chosen as representative for a discussion of its indices, delay distributions, and other relevant properties.

In order to aggregate the results of single trains for both the typical week and the considerations in the wider dataset, we opted to *average* the values of interest of the corresponding trains. Although other approaches would be possible, e.g., counting the number of trains with a D-index within a given threshold (see Section 7), using the average seems more suited to the "analog" nature of the D-indices, as their values are in minutes and fractions. Additional details are provided in the specific subsections.

6.1. Findings in a Typical Day

For this comparison, we use four regional (commuter) trains and four inter-city trains, 50% southbound and 50% northbound. For each train, Table 1 shows D-indices compared with the average of delays and the delay at the final destination. The maximum delay and advance are reported as a further reference. All the sampling points are train stops, i.e., there are no sampling points en route where trains do not stop. For regional trains, the selected line section includes one major hub in between. For weighted indices, we arbitrarily weight the hub stations as double the other stations.

The delay at the final destination, rather expectedly, turns out to be quite an imprecise measure. REG 2, for instance, arrives 3 min late at its final destination despite good performance at most of the previous stops; indeed, its DD and DT indices are close to 0 and better than all the other trains which arrived at their final destination on time. REG 3 even arrived early at its final destination, though only thanks to the recovery time embedded in its schedule; actually, passengers getting off at intermediate stations experienced non-negligible delays, as captured by DD and DT indices of 4.2 and 4.17 min, respectively. It is worth noting that all trains occasionally experienced non-negligible deviations from their schedule here and there at certain stops, as shown by the maximum delay and advance, obviously affecting the passengers getting off there; however, such peaks were not enough to characterise the overall train behaviour.

Table 1. D-indices vs. other delay measures for the selected trains in a typical day (minutes).

	REG 1	REG 2	REG 3	REG 4	IC 1	IC 2	IC 3	IC 4
Journey length (km)	350	350	350	350	1011	568	568	1011
DD	1.91	0.09	4.20	0.30	1.47	0.04	−0.22	2.30
DT	1.91	0	4.17	0.31	1.64	0.02	−0.22	2.27
Bounded DD	2.20	0.41	4.31	0.75	2.23	0.61	0.41	2.65
Bounded DT	2.23	0.38	4.29	0.77	2.32	0.58	0.39	2.61
Weighted DD	2.23	0.06	5.05	0.23	1.36	0.04	−0.25	2.33
Weighted DT	2.30	−0.04	5.06	0.24	1.50	0.02	−0.27	2.30
Delay at final dest.	1	3	−2	−1	1	0	1	1
Average of delays	1.57	−0.3	3.28	−0.13	1.26	−0.50	−0.83	2.29
Maximum delay	+9	+3	+15	+7	+15	+3	+2	+10
Maximum advance	−3	−6	−3	−5	−6	−5	−7	−9

The average of delays may appear to be a better indicator, yet it actually “mixes” delays independently of the inter-station distances, leading to over-sampling of journey sections in which the train stops are closer to each other. REG 1, for instance, travels with a non-negligible delay for over 175 km, within which it makes 10 stops; then, in the next 155 km with 13 more stops, it recovers. Because 13 stops is weighted more heavily than 10 stops, the second part of the journey (which is actually shorter) contributes more to the average than the first (which is longer), leading to an average delay of 3.28 min that makes the train performance to appear better than it was. Its D-indices (DD = 4.2, DT = 4.17), however, tell a different story, returning a closer view to what passengers actually experienced.

Looking at long-distance trains, the delay at the final destination is negligible (0–1 min) in all cases, while the actual behaviours are different, and the D-indices are as well. IC 4, for instance, is about 2 min late throughout its journey, as confirmed both by the D-indices and, in this case, the average of delays, as the delay is rather uniform. IC 3 instead fits its schedule almost perfectly, often arriving slightly earlier than scheduled.

Bounded indices, which do not award early arrivals as good performance, could be of interest in these scenarios; IC 1, for instance, thanks to early arrivals at 11 intermediate stations, has low standard DD/DT indices (1.47/1.64 min, respectively), while its bounded indices are significantly higher (2.23/2.32 min, respectively), bringing out the hidden compensation of previous delays, which, in fact, occasionally reached 15 min.

Weighted indices can be useful for differentiating early/late arrival at minor stations from major hubs. REG 3 arrives 15 min late at the intermediate major hub, while REG 4 arrives there early. In fact, the weighted indices for REG 3 (5.05/5.06 min, respectively) are higher than its standard indices (4.20/4.17 min, respectively), while the opposite holds for REG 4, whose weighted indices (0.23/0.24 min, respectively) are lower than the standard ones (0.30/0.31 min, respectively).

6.2. Findings in a Typical Week

For each of the above trains, we registered the behaviour on all days in a given sample week; in order to aggregate them, we then averaged the results for each train stop. Table 2 reports the results, which are similar to the single-day results reported above.

It is worth noting here that the "hub effect" at the major hub, highlighted by weighted indices, is more perceived on the regional trains (intentionally chosen at peak commuter hours) than on the long-distance trains, which connect major cities by definition. The opposite holds for bounded indices, which are more sensitive for long-distance trains; this is reasonable, as their schedule is more likely planned to account for more recovery and extra times considering the longer journey. Accordingly, early arrivals are more likely to occur when everything goes well.

Table 2. D-indices vs. other delay measures for the selected trains in a typical week (minutes)

	REG 1	**REG 2**	**REG 3**	**REG 4**	**IC 1**	**IC 2**	**IC 3**	**IC 4**
Journey length (km)	350	350	350	350	1011	568	568	1011
DD	0.27	0.63	1.22	0.63	0.18	0.96	0.21	0.69
DT	0.33	0.57	1.20	0.66	0.18	0.76	0.22	0.64
Bounded DD	0.72	0.95	1.54	1.04	1.11	1.64	1.15	1.39
Bounded DT	0.80	0.93	1.49	1.07	1.09	1.49	1.10	1.38
Weighted DD	0.37	0.66	1.42	0.59	−0.01	0.97	0.16	0.71
Weighted DT	0.46	0.61	1.40	0.62	−0.07	0.77	0.16	0.66
Delay at final dest.	4.43	2.57	−1.57	1.00	0.29	5.29	−1.29	0.57
Average of delays	−0.06	0.31	0.69	0.46	−0.45	0.08	−0.37	0.12

Other interesting analyses could be made by weighting the weekdays differently than the weekends based on the type of service; commuter services could be weighted more heavily on weekdays, while tourist-oriented services could be assigned higher weights on the weekends and selected major holidays. Table 3 reports the average indices for weekdays only; a quick comparison with Table 2 shows that, as expected, commuter services tend to perform worse on the weekdays, when the load and possible interference in the network are higher. To a certain extent, the same holds for long-distance trains, though in certain cases (es. IC 2) the behaviour appears to be better, possibly due to higher load on the weekends.

Table 3. D-indices vs. other delay measures for the selected trains in a typical week (minutes)–weekdays only

	REG 1	**REG 2**	**REG 3**	**REG 4**	**IC 1**	**IC 2**	**IC 3**	**IC 4**
Journey length (km)	350	350	350	350	1011	568	568	1011
DD	0.50	0.86	1.61	0.95	0.58	0.01	0.42	1.14
DT	0.61	0.80	1.58	0.98	0.56	−0.04	0.43	1.09
Bounded DD	0.94	1.17	1.87	1.33	1.40	0.64	1.31	1.78
Bounded DT	1.04	1.14	1.83	1.35	1.36	0.60	1.27	1.77
Weighted DD	0.62	0.91	1.87	0.90	0.41	0.02	0.38	1.16
Weighted DT	0.75	0.85	1.85	0.93	0.34	−0.03	0.36	1.11
Delay at final dest.	7.60	3.40	−1.60	1.40	2.00	1.40	−1.60	2.00
Average of delays	0.33	0.55	1.08	0.80	0.01	−0.70	−0.13	0.70

In order to further put in evidence the difference between weekdays and weekends, it would be possible to compute a new weighted average among the indices computed on the weekdays and weekends; as an example, Tables 4 and 5 report the case when weekdays are weighted twice the weekends and vice versa.

Table 4. D-indices vs. other delay measures for the selected trains in a typical week (minutes) with weekdays having twice the weight of weekends

	REG 1	**REG 2**	**REG 3**	**REG 4**	**IC 1**	**IC 2**	**IC 3**	**IC 4**
Journey length (km)	350	350	350	350	1011	568	568	1011
DD	0.36	0.72	1.39	0.76	0.35	0.57	0.30	0.88
DT	0.45	0.67	1.36	0.80	0.34	0.43	0.31	0.83
Bounded DD	0.81	1.04	1.67	1.16	1.23	1.22	1.22	1.55
Bounded DT	0.90	1.02	1.64	1.19	1.20	1.12	1.17	1.54
Weighted DD	0.47	0.76	1.61	0.72	0.16	0.57	0.25	0.89
Weighted DT	0.58	0.71	1.59	0.75	0.10	0.43	0.24	0.85
Delay at final dest.	5.75	2.92	−1.58	1.17	1.00	3.67	−1.42	1.17
Average of delays	0.10	0.41	0.85	0.60	−0.26	−0.24	−0.27	0.36

Table 5. D-indices vs. other delay measures for the selected trains in a typical week (minutes) with weekends have twice the weight of weekdays

	REG 1	**REG 2**	**REG 3**	**REG 4**	**IC 1**	**IC 2**	**IC 3**	**IC 4**
Journey length (km)	350	350	350	350	1011	568	568	1011
DD	0.13	0.50	1.01	0.46	−0.04	1.49	0.10	0.43
DT	0.18	0.45	0.98	0.49	−0.03	1.21	0.11	0.39
Bounded DD	0.60	0.83	1.35	0.88	0.95	2.20	1.05	1.17
Bounded DT	0.66	0.82	1.31	0.91	0.94	1.98	1.01	1.16
Weighted DD	0.22	0.52	1.17	0.42	−0.25	1.50	0.05	0.46
Weighted DT	0.29	0.48	1.16	0.45	−0.30	1.21	0.05	0.41
Delay at final dest.	2.67	2.11	−1.56	0.78	−0.67	7.44	−1.11	−0.22
Average of delays	−0.27	0.17	0.47	0.27	−0.71	0.51	−0.50	−0.20

6.3. About the DA Index

As discussed above, the DA index has no physical counterpart; its purpose is to enable a comparison (most likely, yet not necessarily, among trains operating on the same line) in a more "absolute" and passenger-oriented way. For this purpose, let us consider, e.g., trains REG 1 and REG 2, which cover the same line in opposite directions, and let us further suppose that on a particular day they maintain the same delay, say, 5 min, at all stations. Then, of course, the standard indices, the delay at the final destination, and the average of delays would all be 5 min.

However, this hides the fact that the planned schedule of such trains is (slightly) different; because REG 2 has a larger schedule, its 5-minute delay is proportionally worse than the same 5-minute delay of REG 1, which has less recovery time in its schedule. This is what DA indices are for; indeed, REG 1 has a DA value of 36.54, while REG 2 has a value of 37.1, capturing the difference in their actual behaviour and quality of service.

6.4. Corner Cases

In order to highlight specific situations that could occasionally be of interest, we consider and compare two more trains:

- a high-speed train that arrives at the final destination more than 15 min late despite good performance for most of the previous part of its journey;
- a regional train which arrives at its final destination quite late after arriving late at all/most of the previous stops.

The first case is interesting because, according to the quality parameters defined in [5,13], arriving more then 15 min late at the final stop leads to the train being classified as "not on time", yet its D-indices (DD = 2.52, DT = 2.26) tell a much less severe story, indicating that most of its delay was probably accumulated only in the last section of the journey.

The second case is interesting for the opposite reason, as the delay at the final destination and all other indices, both the average of delays and the D-indices, are over 20 min, indicating a persistent delay affecting passengers throughout the journey.

6.5. Analysis of a Larger Dataset

We monitored and transcribed from public sources fourteen trains (ten regional and four long-distance) over a period of 2 years between 2021 and 2022; each train was sampled 126 times, though only 95 records per train were actually usable due to a lack of data on certain days.

In order to better investigate the pros and cons of the D-indices with respect to other indicators, we focus on one of these trains, namely, a regional service with a 350 km journey, 4h27 total journey time (3h50 of actual travel time without dwell times at stops), and 22 stops (plus the departure station).

Figure 6 shows the distribution of delays at the 23 stops (for the first, the delay on departure is considered instead of the one on arrival) for the 95 samples of the selected train. It turns out that the distribution is in most cases approximately log-normal, and delays are less variable in the initial than in the central (and, to a minor extent, final) part of the train journey. Despite peaks, i.e., days where the train ran considerably out of schedule (which are quite of interest for the passengers' experience), such a distribution suggests that the mean of the 95 train trips is actually significant enough to depict the behaviour of the "typical" journey of the train; indeed, the median (not shown) is very close to the mean. Due to the sum-based nature of the D-indices, we assume that aggregating these 95 train runs (which refer to the same train sampled 95 times) by averaging their single D-indices can be a meaningful approach for our comparison purposes. Accordingly, in the following we take the averaged D-indices as the reference for our subsequent data analysis. In order to support intuition, the resulting average values can be thought of as if they were the actual values of a sort of "imaginary train". We adopt this terminology below in both the text and figure captions.

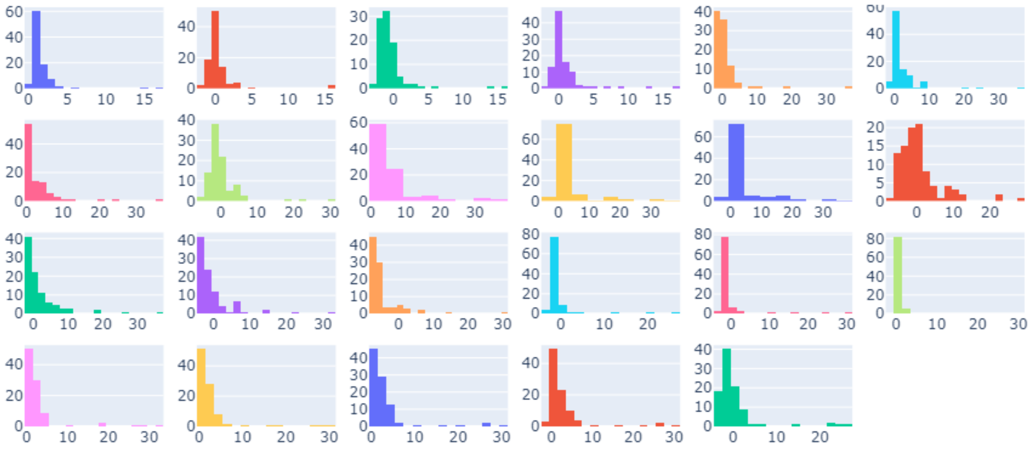
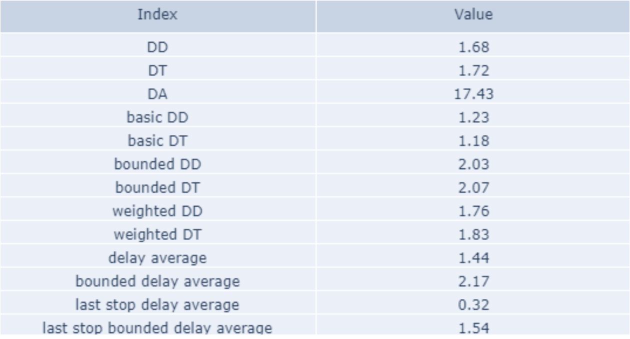


Figure 6. Distribution of delays at 23 stops (95 train samples), plotted in different colors. For the first stop, the delay on departure is considered instead of the delay on arrival.

Figure 7 shows the corresponding results; in order to ease the comparison with the delay at the final destination and the average of delays, the latter are provided zero-bounded, i.e., computing 0 for early arrivals. First, it can be seen that, as expected, the basic indices are smaller than the standard indices, and in turn these are lower than the bounded indices. This is coherent with their definitions, as DD and DT (Definitions 3 and 4) consider the possibly negative delay on arrival and zero-bound the delay on departure, while bounded DD and DT (Definitions 5 and 6) zero-bound both. In this sense, DD and DT can be seen as an intermediate viewpoint, "fair enough" from the passengers' perspective in that they do award early arrivals yet with no over-estimation. Moreover, the average of delays appears between the basic and standard indices, while its bounded version is greater than the bounded indices. As discussed above, this is because the average assigns the same weight to all samples, independently of their order and inter-distance. It is worth noting that the bounded DD index, perhaps the most natural reference, is 2.03, while the corresponding bounded average delay is 2.17; although this might appear low, a nearly 7% difference is not negligible, in particular as it is the result of averaging over 95 trains.



Index	Value
DD	1.68
DT	1.72
DA	17.43
basic DD	1.23
basic DT	1.18
bounded DD	2.03
bounded DT	2.07
weighted DD	1.76
weighted DT	1.83
delay average	1.44
bounded delay average	2.17
last stop delay average	0.32
last stop bounded delay average	1.54

Figure 7. Aggregated D-indices obtained as the average of the 95 single D-indices (82 for DA).

For the sake of completeness, Figure 7 includes the average delay at the final destination, although a more typical performance indicator is the number of trains arriving within a given delay threshold. Here, the purpose is merely to compare its order of magnitude with respect to the D-indices. It is worth highlighting that the above indices, being the result of averaging 95 trains, indicate unsatisfactory performances even when their values are apparently not very high; for instance, while DD = 1.68 min may seem low, it has to be read together with the corresponding average delay at the final destination of 0.32, which is about five times lower.

Because quality of service indicators in contracts are often expressed in terms of how many trains, as a percentage, arrive at the final destination (say) 5–10–15 min late with respect to their schedule, it may be interesting to do the same on our dataset. Figure 8 shows the results; four trains out of 95 arrived more than 15 min late, one train arrived more than 10 min late, and a further train arrived more than 5 min late, leading to an impressive 93.7% of (apparently) “on time” trains. Yet, the D-indices tell another story. Figure 9 shows the DD delay diagram for the “imaginary” train obtained by averaging the 95 single D-indices. The right side of the graph, close to the last stop, actually drops, in accordance with the above percentage; however, the overall behaviour of the train shows that the passengers’ experience has been different, with notable delay peaks (as well as an early arrival) at several intermediate points. This is why the resulting DD is 1.68, and is the reason why such an apparently low value has to be properly interpreted.

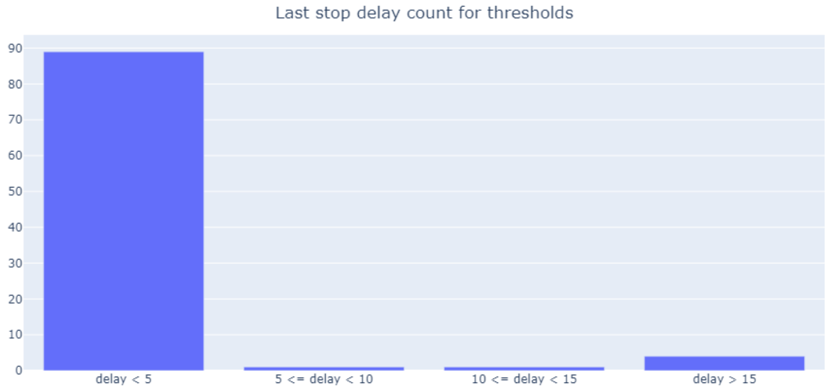


Figure 8. Number of late trains at the last stops plotted per range of delay (95 trains).

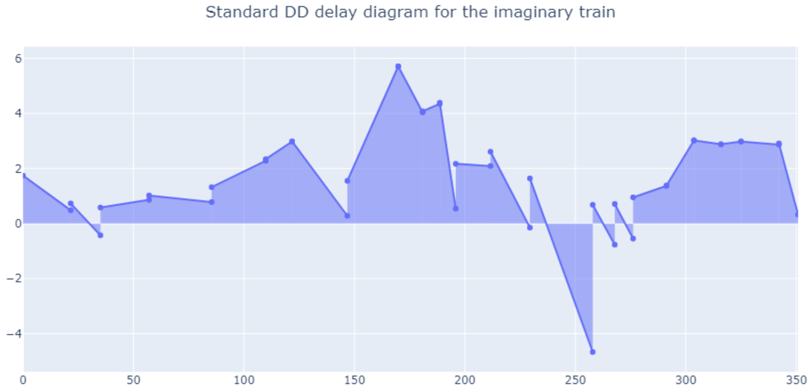


Figure 9. DD delay diagram for the imaginary train obtained as the mean of 95 single DDs.

In order to further investigate the difference between the D-indices and the average of delays, it is worth comparing space- and time-based representations of delay at each stop (i.e., DD/DT basic indices) vs. a representation in which the sampling points are equidistant (i.e., uninfluential), which recalls the standard average of delays (Figure 10) in which space/time distances on the *X*-axis are normalised to an a-dimensional 0-1 interval. Blue areas represent basic D-indices, while red areas the equidistant version, recalling the standard average. Overall, these diagrams confirm that the viewpoints provided by DD and DT indices, that is, space vs. scheduled time, are similar as concerns their ability to capture the overall train behaviour, whereas equidistant points, recalling the plain average approach, are inherently different. In particular, DD and DT are obviously similar for lines where the schedule is planned assuming about the same commercial speed for all/most of the trains; in such cases, what actually matters is clearly just the spatial distance among the sampling points.

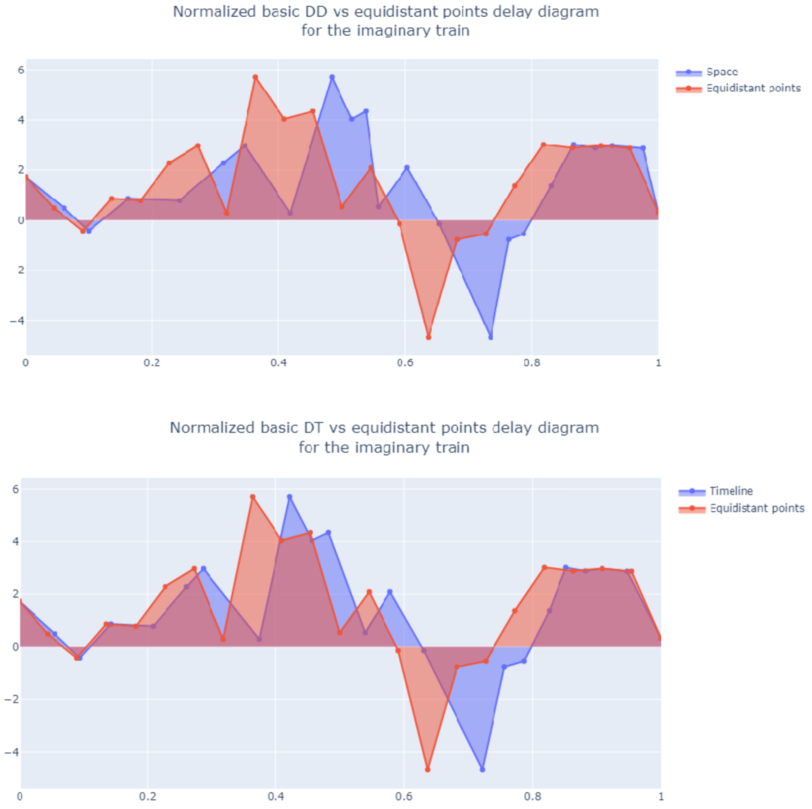


Figure 10. Normalized basic DD (**top**) and DT (**bottom**) delay diagram vs. equidistant point delay diagram. Distances in space and time are normalized to the a-dimensional [0–1] interval.

7. Discussion and Conclusions

The D-indices aim at providing a view of train performance that retains the physical dimension of time while being closer to the actual passengers' experience, and possibly easier to understand by passengers, commuters' associations, etc.

As such, D-indices are intended to capture the quality of the travel experience from the passengers' perspective for commercial and performance purposes; they are not conceived for use ex ante, in the planning phase, for train operators and infrastructure managers in their design and development activities, or for any technical aspects, nor to estimate/model train delays for statistical analyses, infrastructure modelling, simulation, etc., the passengers' perspective in railway timetabling being a wholly different research area [29].

Rather, they could in principle be used as replacements for commercial quality of service indicators [13], usually defined as the percentage of trains "on time" or travelling within a given delay, once a suitable *aggregation policy* is defined for aggregating the single D-indices of a set of train runs into a single indicator. Potentially, D-indices allow for a plurality of aggregation policies; the aggregated numbers, however obtained, could then be used to verify the selected performance targets. In Section 6.5, we explored the idea of averaging the single D-indices, but one could also evaluate, e.g., the number/percentage of trains whose D-index is below a given threshold, or others as well. Whatever the aggregation policy, aggregating the D-indices of single train runs makes it possible to identify and keep separate the performance of each train run and the aggregated measure, which in turn makes the aggregated indicator more transparent and easier to understand by passengers, an important requirement for ensuring trust in the official statistics. Of course, as discussed above, train operators possess huge amounts of much more detailed data, possibly from real-time measurements in the infrastructure and trains themselves, which can be used to build more sophisticated representations for any possible use; yet, such processing is likely out of reach for passengers, consumers' associations, and possibly even authorities, risking to reproduce the "undermining trust" issue between official data and passengers' feelings reported in [19].

The reason for defining a family of indices rather than a single one is that a single indicator can hardly represent the multi-facet reality of passengers' experience [13]; in particular, whether to either award early arrivals as a sign of good performance or conversely to consider them as poor performance is a key issue leading us to distinguish among basic, standard, and bounded indices. In addition, whether to prefer a "static" view, such as the one provided by space-based indices, or to opt for a more relative and "dynamic" metrics, such as the one provided by time-based indices, is another issue, and reflects the user's preference for evaluating train performance, either on a fixed scale or based on what the specific train was supposed to do. The delay-absolute index makes it possible to evaluate the train performance in a way that is uninfluenced by the journey time, while weighted indices make it possible to assign more weight to those delays that impact more passengers.

The drawback of having this many indices is clearly that there is a degree of uncertainty as to which should be chosen; thus, Table 6 summarises their features, limitations, and intended use cases.

In short, standard indices are a natural choice whenever one is interested in obtaining the picture of a train "as is", with no intrinsic interpretation; while they do award early arrivals, according to the passengers' feeling that arriving earlier is generally considered good, early arrivals are not transferred on departure as the raw basic indices would do, avoiding misinterpretation and over-awarding. Bounded indices, in turn, discourage any deviation from the expected schedule, not rewarding early arrivals. In this sense, they probably fit the most common expectation and are adequate for a variety of situations and passengers' perceptions. They can be helpful for detecting larger schedules, when passengers are forced to spend more time on the journey than actually needed. At the same time, their approach may appear counter-intuitive in that passengers often do not disdain arriving earlier, at least, as long as their journey does not become artificially longer as a result. Weighted indices, straightforwardly, make it possible to assign more weight, regardless of the weighting criterion, to hubs and more generally to stations with many alighting passengers, even if the train continues its journey thereafter.

Table 6. The D-indices family: feature summary.

Index	**Features**	**Limits**	**Use Cases**
Basic DD	Static view; early arrivals awarded	early arrivals counted also on departure, over-compensate delays	N/A
Basic DT	Dynamic view; early arrivals awarded	early arrivals counted also on departure, over-compensate delays	N/A
Standard DD	Static view; fair treatment of early arrivals	early arrivals can compensate actual delays	scenarios where passengers are sensitive to early arrivals
Standard DT	Dynamic view; fair treatment of early arrivals	early arrivals can compensate actual delays	scenarios where passengers are sensitive to early arrivals in a dynamic view
DA	view uninfluenced by the journey time	no direct physical meaning	absolute comparison of trains w/each other (how worse/better)
Bounded DD	Static view; early arrivals do not compensate delays	worst-case approach, early arrivals seen as negative	identify any deviation from schedule in the most common scenario
Bounded DT	Dynamic view; early arrivals do not compensate delays	worst-case approach, early arrivals seen as negative	identify any deviation from schedule in common scenarios, in a dynamic view
Weighted DD	Static view; hubs can weight more	early arrivals as in DD	most common scenarios as in DD
Weighted DT	Dynamic view; hubs can weight more	early arrivals as in DT	most common scenarios as in DT

As concerns the spatial vs. temporal issue, the DT indices are perhaps more useful when evaluating the performance of a single train, while the DD indices (and DA, from a different standpoint) can be more helpful when comparing multiple trains on the same line. In fact, by mapping delays onto its own scheduled timeline, DT evaluates each train in terms of what it was expected to do (the "dynamic" view), while DD evaluates all the trains of a given line on the same (fixed, spatial) "static" basis, which is inherently insensitive to the scheduled planning, working in a certain sense as if all trains were supposed to share the same schedule. For this reason, DD can be a better choice if the goal is to compare different trains (typically, with different schedules) on the same line. If a time-based yet more absolute view is needed, however, the DA index could work better despite its lack of direct physical meaning.

Under certain conditions, the D-indices feature a number of interesting properties that can be of help in composing and comparing train performance.

On the other hand, the D-indices suffer from some limitations. First, time-based indices are potentially less easy to interpret than space-based indices due to subtleties that it is important to be aware of. In this sense, spatial indices, with their static yet easier-to-interpret view, are probably the better choice to start from in the analysis process. Time-based indices can play a role later in the process to investigate whether a train is "making good use" of its allowed time.

Another point worthy of attention is that if the sampling points are close enough to each other, then the difference with the average of delays tends to be reduced. In principle, very frequent delay samples, possibly obtained via a real-time delay check, would accurately describe the train behavior, leading the D-indices approach becoming identical to the average one, apart from the consideration of early arrivals at intermediate stops. However, such data, provided that they are made available to passengers and the general public, would require ad hoc elaborations, likely out-of-reach for passengers, re-introducing the risk of possibly "undermining trust" in the official statistics [19]. In this sense, the D-indices can be seen as way to achieve a similar result from a set of sampling points (in practice, stations and intermediate stops) that are easily available to the general public, in combination with the notion of situatedness which relates them to the physical mutual location of the sampling points.

Lastly, the definitions of the D-indices intentionally exclude the delay on departure, according to the assumption (as in [13]) that passengers are only interested in their arrival time and that delays on departure are only relevant to the extent that they cause a delay on arrival at subsequent stops. Although a more precise model could in principle be devised in which delays on departure are explicitly modelled, it is unclear how such delays should be regarded; if, on the one hand, a train leaving late is a clear deviation from the planned schedule, on the other hand it could allow late passengers to make an unhoped-for train. Moreover, the actual relevance of a delay on departure depends on the impact on subsequent delays on arrival, if any.

As mentioned above, an open research direction concerns the exploration of other aggregation policies other than the plain D-indices average adopted in Section 6.5. In particular, following more recent approaches in the literature (Section 2), it would be worthwhile to evaluate a policy based on counting the number/percentage of trains with a D-index below a given threshold, then comparing it with current best practices.

Other future work will be devoted to further analysis on wider and possibly more specific datasets, especially as concerns the DD/DT dichotomy, in order to better investigate their representativeness in specific contexts.

Author Contributions: Basic idea, E.D.; Conceptualization and formal analysys, E.D. and L.B.; methodology, E.D. and L.B; validation, E.D. and L.B.; data set and data curation, L.B.; writing—original draft preparation, E.D. and L.B; writing—review and editing, E.D. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The data presented in this study are available on request from the corresponding author, due to anonymisation issues.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this Appendix, we discuss in more depth the mathematics behind the properties presented in Section 5.

Appendix A.1. Proportionality

Proportionality means that two trains A and B with the same delays, in the same order, and with the same sampling points have the same D-indices if their inter-sampling point distances (L_i or T_i, respectively) are in the same proportion:

$$\exists k \in \mathit{Int} : \forall i \in [1..N], L_i^A / L_i^B = k \text{ for DD}$$
$$\exists k \in \mathit{Int} : \forall i \in [1..N], T_i^A / T_i^B = k \text{ for DT}$$

For instance, a train B half as fast as train A (i.e., $T_i^B = 2 \cdot T_i^A$ for any index i) but exhibiting the same delays has the same DT index as train A. This is precisely what makes D-indices meaningful for performance; the lower the index, the better the performance, and vice versa, independent of the spatial/temporal characteristics of the train journey.

In the very special case when all the delays are identical, i.e., $d_i = d \; \forall i \in [1..N]$, the two DD and DT indices are identical as well, and both are equal to d, the same value that both the delay-on-arrival and the average approaches would compute. The DA index, instead, results in $\sqrt{d \cdot T_{tot}}$, confirming its lack of direct physical meaning.

Appendix A.2. Compositionality

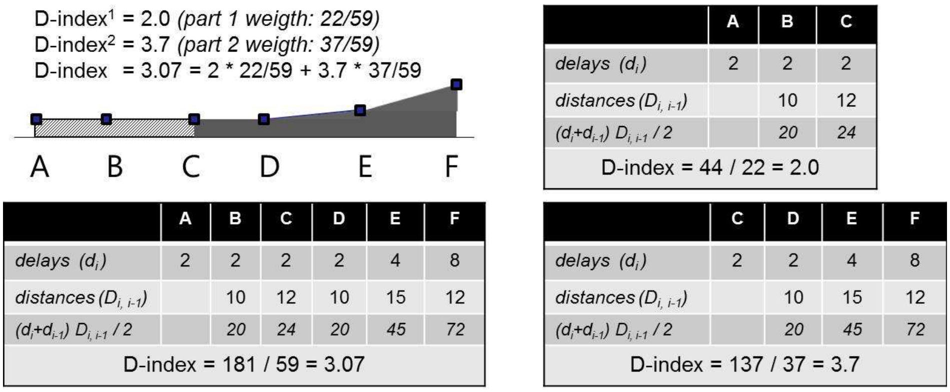
Compositionality derives from the fact that areas in the delay diagram are inherently additive; thus, the D-indexes D1 of a train covering the route segment A–B and D2 of the same train continuing in the route segment B–C can be composed together using the respective fractions of (spatial or temporal) distance as coefficients. In the case of time-based indices, this requires the extra assumption that the timeline is not altered, that is, that the addition of the intermediate point (stop) B does not change the global trip duration ΔT of the whole route A–C, which remains the sum of the two durations $\Delta T1$ of the route segment A–B and $\Delta T2$ of the route segment B–C, as obviously occurs for spatial distances. For the sake of simplicity, in the following we refer to DT index only, though analogous considerations hold for the other DD indices.

Let us consider two sub-parts 1 and 2 of a given train journey, and let DX^1 and DX^2 be their corresponding sub-indices. More precisely, let us say that part 1 covers station 0 to station P, meaning that its sum in the DX^1 index ranges from stations 1 through P, while part 2 covers from station P to station N and its sum in the DX^2 index ranges from stations P+1 through N. The global D-index DX^G of the whole journey can then be expressed as follows (where D_i indifferently stands for spatial or temporal distances):

$$\begin{aligned} DX^G &= \frac{1}{2} \cdot \frac{\sum_1^N (d_i + d^*_{i-1}) D_i}{D_{tot}^{0-N}} = \\ &= \frac{1}{2} \cdot \frac{\sum_1^P (d_i + d^*_{i-1}) D_i}{D_{tot}^{0-N}} + \frac{1}{2} \cdot \frac{\sum_{P+1}^N (d_i + d^*_{i-1}) D_i}{D_{tot}^{0-N}} = \\ &= \frac{1}{2} \cdot \frac{\sum_1^P (d_i + d^*_{i-1}) D_i}{D_{tot}^{0-P}} \cdot \frac{D_{tot}^{0-P}}{D_{tot}^{0-N}} + \frac{1}{2} \cdot \frac{\sum_{P+1}^N (d_i + d^*_{i-1}) D_i}{D_{tot}^{P-N}} \cdot \frac{D_{tot}^{P-N}}{D_{tot}^{0-N}} = \\ &= DX^1 \cdot \frac{D_{tot}^{0-P}}{D_{tot}^{0-N}} + DX^2 \cdot \frac{D_{tot}^{P-N}}{D_{tot}^{0-N}} = DX^1 \cdot \frac{D^1_{tot}}{D_{tot}} + DX^2 \cdot \frac{D^2_{tot}}{D_{tot}} \end{aligned} \tag{A1}$$

which can be rephrased to say that the global index is the weighted average of the two sub-indices, where the weights are the respective fractions of distance.

Figure A1 illustrates this situation, elaborating on the example already presented in Figure 3(top). As shown in the tables, the D-index of the first A–C section of the journey is 2.0, while the D-index of the second C–F section of the journey is 3.7; as the A–C section is 22 units long and the C–F is 37 units long, the global D-index can be obtained by composing the two sub-indices by factors of 22/59 and 37/59, respectively.


D-index¹ = 2.0 (part 1 weigth: 22/59)
D-index² = 3.7 (part 2 weigth: 37/59)
D-index = 3.07 = 2 * 22/59 + 3.7 * 37/59
A
B
C
D
E
F

	A	B	C
delays (d_i)	2	2	2
distances ($D_{i,\,i-1}$)		10	12
(d_i+d_{i-1}) $D_{i,\,i-1}$ / 2		*20*	*24*
D-index = 44 / 22 = 2.0			

	A	B	C	D	E	F
delays (d_i)	2	2	2	2	4	8
distances ($D_{i,\,i-1}$)		10	12	10	15	12
(d_i+d_{i-1}) $D_{i,\,i-1}$ / 2		*20*	*24*	*20*	*45*	*72*
D-index = 181 / 59 = 3.07						

	C	D	E	F
delays (d_i)	2	2	4	8
distances ($D_{i,\,i-1}$)		10	15	12
(d_i+d_{i-1}) $D_{i,\,i-1}$ / 2		*20*	*45*	*72*
D-index = 137 / 37 = 3.7				

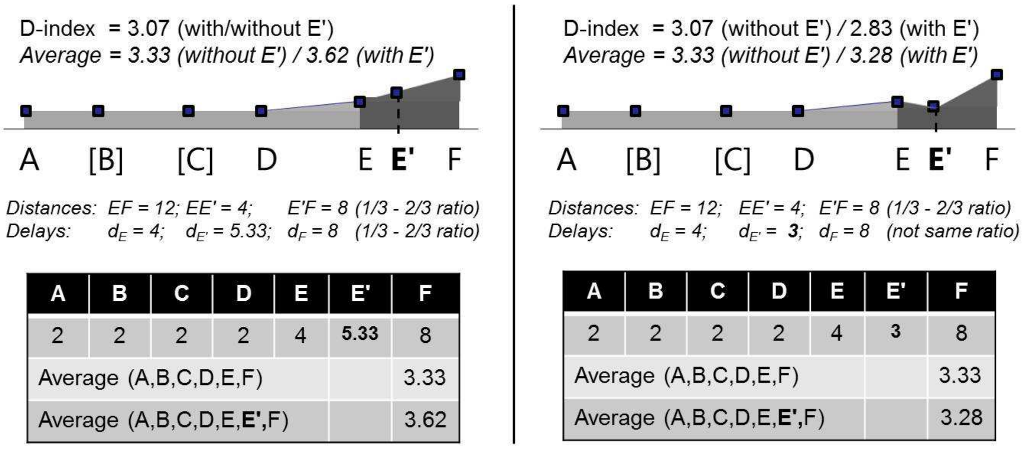
Figure A1. Compositionality: the global D-index is the weighted average of the two sub-indices.

Appendix A.3. Insensitivity To Sampling Point Insertion

Insensitivity to sampling point insertion is a consequence of compositionality, and holds under the same assumptions; it means that, unlike what happens with the average, adding an extra sampling point does not alter the D-index if the delay measured at the new point is either the same of the two adjacent ones or is in their proportion (i.e., fits the line conjuncting them).

To better illustrate this aspect, we can consider two trains T1 and T2 covering the same route A–F with the same schedule, with the only difference being that train T2 stops at intermediate stations B and C while train T1 does not (Figure 1(left (a))). Clearly, the average of delays is sensitive to the presence/absence of extra values sampled at intermediate points, even in the (frequent) case when the delays sampled at the extra stations simply confirm the adjacent values. The area-based approach of the D-indices helps to address this issue; in fact, as long as the area below the corresponding section of the diagram remains the same (as in the case shown in Figure 3, *top*, for stations B,C), the resulting indices do not change either *provided that the distance between the two adjacent stations remains the same* (Figure 3(*top left*)).

This continues to hold if the delay either grows or decreases linearly in section A–F, meaning that it has the "proper" proportional intermediate value at the new intermediate point. Graphically speaking, it means that the new delay $d_{E'}$ measured at the extra point E′ is aligned with the two adjacent delays (Figure A2(*left*)). If these conditions do not hold (Figure A2(*right*)), the property does not hold either; of course, the lower the deviation from the scenario in which it holds, the more it can be taken as "approximately" valid for practical purposes.



A	B	C	D	E	E'	F
2	2	2	2	4	5.33	8
Average (A,B,C,D,E,F)						3.33
Average (A,B,C,D,E,**E'**,F)						3.62

A	B	C	D	E	E'	F
2	2	2	2	4	3	8
Average (A,B,C,D,E,F)						3.33
Average (A,B,C,D,E,**E'**,F)						3.28

Figure A2. Insensitivity to sampling point insertion. (***left***) A case where property still holds (aligned delay) and (***right***) a case where it does not.

While the above might seem a strong constraint, the depicted situation might not be unlikely to occur (at least approximately) in practice, as delays by nature feature a kind of 'inertia' in that cannot suddenly disappear or have sharp edges for physical reasons, meaning that smooth trends are the most probable. As recalled above, in timetable planning the time flow would be certainly altered by the insertion of an intermediate stop; here, however, the viewpoint is different, as the *X*-axis simply represents the time flow in terms scheduled arrival/passing times at specified locations all-inclusively as perceived from the passengers' eyes.

Appendix A.4. Comparability

Comparability is a consequence of insensitivity to sampling point insertion and holds under the same assumptions, making it possible to compare the performance of local vs. express trains serving the same route on a uniform basis. In fact, if train T1 (the 'local' train) makes more stops than train T2 (the 'express' train), the above property makes their D-indices mutually comparable so far as it holds, despite the different number of stops made, as the number of sampling points is not directly relevant.

In the same situation, the average of delays would instead provide two hardly comparable values, as one average (A1) would be made on more samples than the other (A2); removing the effect of the extra stops not present in A2 from A1 would be impractical, the only way being to effectively recalculate the average itself, provided that the basic data are available.

Of course, comparability is valid only if the delay is constant or linearly increasing in the section where the local train makes the extra stops (Figure A3(*top left*)); the more the real situation deviates from such an ideal setting, the less meaningful the comparison becomes (Figure A3(*top right*)).

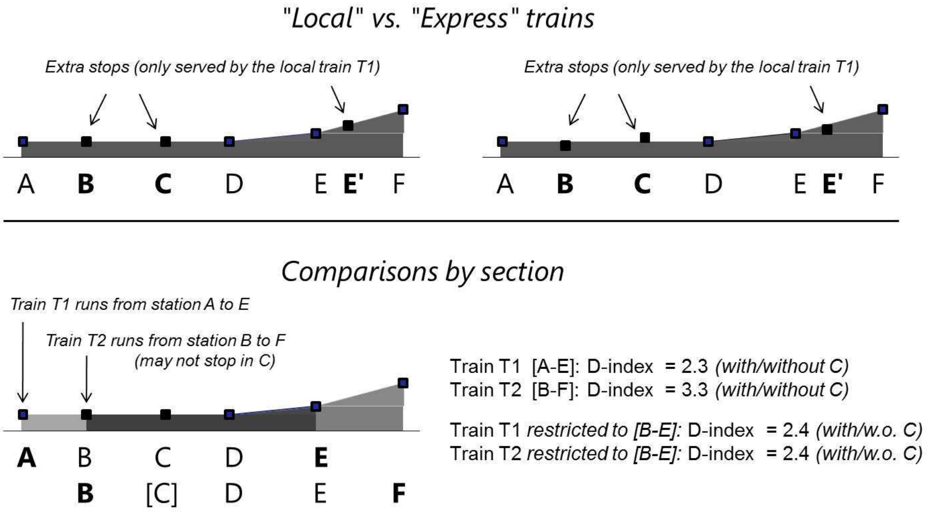

"Local" vs. "Express" trains
Extra stops (only served by the local train T1)
Extra stops (only served by the local train T1)
A B C D E E' F
A B C D E E' F
Comparisons by section
Train T1 runs from station A to E
Train T2 runs from station B to F (may not stop in C)
A B C D E
B [C] D E F
Train T1 [A-E]: D-index = 2.3 (with/without C)
Train T2 [B-F]: D-index = 3.3 (with/without C)
Train T1 restricted to [B-E]: D-index = 2.4 (with/w.o. C)
Train T2 restricted to [B-E]: D-index = 2.4 (with/w.o. C)

Figure A3. Comparability: ***(top)*** local vs. express trains (perfect/approximate comparison) and ***(bottom)*** comparison restricted to a common route section.

In addition, it is possible to compare trains by section, that is, restricting the comparison to a relevant section of their routes; in Figure A3(*bottom*), the D-index of train T1 running from A to E is 2.3, while the D-index of train T2 running from B to F is 3.3. While a direct comparison of these values does provide relevant information about the overall performances of the two trains during their whole journeys, a comparison restricted to the common B–E section would focus on the trains' performance on the busier and more critical central section. Thanks to the property of invariance to sampling point insertion, so far as it holds, the resulting indices are valid independently of the presence of the intermediate C and D stops.

References

1. Yuan, J. Statistical Analysis of Train Delays. In *Railway Timetable & Traffic*; Hansen, I.A., Pachl, J., Eds.; Eurail Press: Hamburg, Germany, 2008; Chapter 10.
2. Hansen, I.A. Improving Railway Punctuality by Automatic Piloting. In Proceedings of the 2001 IEEE Intelligent Transportation Systems Conference, Oakland, CA, USA, 7 August 2002; IEEE: Oakland, CA, USA, 2001; pp. 792–797. https://doi.org/10.1109/ITSC.2001.948761.
3. Økland, A.; Olsson, N.O.E. Punctuality development and delay explanation factors on Norwegian railways in the period 2005-2014. *Public Transp.* **2021**, *13*, 127–161. https://doi.org/10.1007/s12469-020-00236-y.
4. Börjesson, M.; Eliasson, J. On the use of average delay as a measure of train reliability. *Transp. Res. Part Policy Pract.* **2011**, *45*, 171–184. https://doi.org/10.1016/j.tra.2010.12.002.
5. ART-Autorita di Regolazione dei Trasporti. Allegato A alla Delibera n. 16 dell'8 febbraio 2018: Atto di regolazione recante "Condizioni minime di qualità dei servizi di trasporto passeggeri per ferrovia, nazionali e locali, connotati da oneri di servizio pubblico, ai sensi dell'articolo 37, comma 2, lettera d), del decreto-legge 6 dicembre 2011, n. 201, convertito, con modificazioni, dalla legge 22 dicembre 2011, n. 214". ART Web Site. 2018. Available online: https://www.autorita-trasporti.it/wp-content/uploads/2018/02/All.-A_delibera-n.-16_2018.pdf (accessed on 12 January 2023)

6. Martin, U. Performance Evaluation. In *Railway Timetable & Traffic*; Hansen, I.A., Pachl, J., Eds.; Eurail Press: Hamburg, Germany, 2008; Chapter 10.
7. Kroon, L.G.; Dekker, R.; Vromans, M.J. Cyclic Railway Timetabling: A Stochastic Optimization Approach. In *Proceedings of the Railway Optimization 2004*; Geraets, F., Ed., Springer: Berlin/Heidelberg, Germany, 2007; pp. 41–66.
8. Infrabel. How do We Measure Punctuality. Available online: https://infrabel.be/en/punctuality (accessed on 12 January 2023).
9. Stagni, G. I Contratti di Servizio (Service Contracts). Available online: http://www.stagniweb.it/Rifor04b.htm (accessed on 16 January 2023).
10. Altroconsumo. Pendolari, il 39% dei treni è in ritardo (Commuters, 39% of Trains is Late). Available online: https://www.altroconsumo.it/vita-privata-famiglia/viaggi-tempo-libero/news/inchiesta-pendolari (accessed on 16 January 2023).
11. Trenitalia. Carta dei Servizi Regione Emilia Romagna 2019. Available online: https://www.trenitalia.com/content/dam/tcom/allegati/trenitalia_2014/in_regione/carta-dei-servizi/FSI_Carta_servizi_EMILIA_ROMAGNA.pdf(inItalian)(accessed on 12 January 2023).
12. Trenitalia. Carta dei servizi Regione Piemonte 2022. Available online: https://www.trenitalia.com/it/treni_regionali/piemonte/carta-dei-servizi-piemonte.html (accessed on 12 January 2023).
13. Blayac, T.; Stéphan, M. Are retrospective rail punctuality indicators useful? Evidence from users perceptions. *Transp. Res. Part A Policy Pract.* **2021**, *146*, 193–213. https://doi.org/10.1016/j.tra.2021.01.013.
14. Grechi, D.; Maggi, E. The importance of punctuality in rail transport investigation on the delay determinants. *Eur. Transp. Eur.* **2018***12*, 1–23.
15. Nathanail, E. Measuring the quality of service for passengers on the hellenic railways. *Transp. Res. Part Policy Pract.* **2008**, *42*, 48–66. https://doi.org/10.1016/j.tra.2007.06.006.
16. Tirachini, A.; Godachevich, J.; Cats, O.; Muñoz, J.; Soza-Parra, J. Headway variability in public transport: a review of metrics, determinants, effects for quality of service and control strategies. *Transp. Rev.* **2021**, *42*, 1–25. https://doi.org/10.1080/01441647.2021.1977415.
17. Cats, O. Regularity-driven bus operation: Principles, implementation and business models. *Transp. Policy* **2014**, *36*, 223–230. https://doi.org/10.1016/j.tranpol.2014.09.002.
18. Comi, A.; Nuzzolo, A.; Brinchi, S.; Verghini, R. Bus dispatching irregularity and travel time dispersion. In Proceedings of the 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Naples, Italy, 26–28 June 2017; pp. 856–860. https://doi.org/10.1109/MTITS.2017.8005632.
19. Transport Focus. Train punctuality: The passenger perspective. Available online: https://www.transportfocus.org.uk/publication/train-punctuality-the-passenger-perspective/ (accessed on 12 January 2023).
20. Börjesson, M.; Eliasson, J. Train Passengers' Valuation Of Travel Time Unreliability. In Proceedings of the 2008 European Transport Conference, Leiden, The Netherlands, 6–8 October 2008.
21. Monsuur, F.; Enoch, M.; Quddus, M.; Meek, S. Modelling the impact of rail delays on passenger satisfaction. *Transp. Res. Part A Policy Pract.* **2021**, *152*, 19–35. https://doi.org/10.1016/j.tra.2021.08.002.
22. Michaelis, M.; Schöbel, A. Integrating line planning, timetabling, and vehicle scheduling: A customer-oriented heuristic. *Public Transp.* **2009**, *1*, 211–232. https://doi.org/10.1007/s12469-009-0014-9.
23. Hansen, I.A.; Pachl, J. (Eds.) *Railway Timetable & Traffic*; Eurail Press: Hamburg, Germany, 2008.
24. Schöbel, A. Capacity constraints in delay management. *Public Transp.* **2009**, *1*, 135–154. 10.1007/s12469-009-0010-0.
25. Fosgerau, M.; Engelson, L. The value of travel time variance. *Transp. Res. Part B Methodol.* **2011**, *45*, 1–8. https://doi.org/10.1016/j.trb.2010.06.001.
26. Rail Delivery Group. Definitions of Railway Performance Metrics. Available online: https://www.raildeliverygroup.com/media-centre-docman/acop/287-rdgntfdorpm-final/file.html (accessed on 2 April 2023).
27. From the First Great Western blog. Train Punctuality Needs Review, Says Passenger Focus. Available online: www.firstgreatwestern.info/coffeeshop/index.php?topic=8368.0;wap2, (accessed on 12 January 2023).
28. Goverde, R.; Hansen, I.A.; Hooghiemstra, G.; Lopuhaa, H.P. Delay distribution in railway stations. In Proceedings of the 9th World Conference on Transport Research, Seoul, Korea, 22–27 July 2001.
29. Parbo, J.; Nielsen, O.A.; Prato, C.G. Passenger Perspectives in Railway Timetabling: A Literature Review. *Transp. Rev.* **2016**, *36*, 500–526. https://doi.org/10.1080/01441647.2015.1113574.
30. Landmark, A.D.; Seim, A.A.; Olsson, N. Visualisation of Train Punctuality - Illustrations and Cases. *Transp. Res. Procedia* **2017**, *27*, 1227–1234. https://doi.org/10.1016/j.trpro.2017.12.092.
31. Palmqvist, C.; Olsson, N.; Hiselius, L. Delays for passenger trains on a regional railway line in southern Sweden. *Int. J. Transp. Dev. Integr.* **2017**, *1*, 421–431. https://doi.org/10.2495/TDI-V1-N3-421-431.
32. Harris, N.G.; Haugland, H.; Olsson, N.; Veiseth, M. *An Introduction to Railway Operations Planning*; A & N Harris: London, UK, 2016; p. 162.
33. Olsson, N.O. Train punctuality analysis in a rolling stock perspective. *Transp. Res. Procedia* **2020**, *47*, 641–647. In Proceedings of the 22nd EURO Working Group on Transportation Meeting, EWGT 2019, Barcelona, Spain, 18–20 September 2019. https://doi.org/10.1016/j.trpro.2020.03.142.

34. Palmqvist, C.W.; Olsson, N.; Winslott-Hiselius, L. Some influencing factors for passenger train punctuality in Sweden. *Int. J. Progn. Health Manag.* **2017**, *8*, 1–13. https://doi.org/10.36001/ijphm.2017.v8i3.2649.
35. Richter, T. Systematic analyses of train run deviations from the timetable. In *Computers in Railways XII—Computer System Design and Operation in Railways and Other Transit Systems*; The WIT Transactions on The Built Environment; Ning, N., Brebbia, C.A., Tomii, N., Eds.; WIT Press: Southampton, UK, 2010; Volume 114. https://doi.org/10.2495/CR100601.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.