A good or a bad timetable: Do different evaluation functions agree?

Johann Hartleb*[1,2], Marie Schmidt[2], Markus Friedrich[1], and Dennis Huisman[3,4]

[1]Institute for Road and Transport Science, University of Stuttgart, Germany
[2]Rotterdam School of Management and Erasmus Center for Optimization in Public Transport, Erasmus University Rotterdam, The Netherlands
[3]Econometric Institute and Erasmus Center for Optimization in Public Transport, Erasmus University Rotterdam, The Netherlands
[4]Process quality and Innovation, Netherlands Railways, Utrecht, The Netherlands

April 4, 2019

Abstract

We compare different evaluation functions that are all designed to measure the quality of a timetable from passengers' perspective. Already in small examples fundamentally different timetables can be preferred by evaluation functions that seem to be similar. To investigate this effect in practice, we design a set of evaluation functions as representatives for a wide range of commonly used evaluation functions in optimization models, evaluation applications, or choice models. These functions are compared by analyzing their evaluation values of multiple timetables in three case studies. To investigate to what extent these evaluation functions agree on a good or a bad timetable, we apply cluster analysis as well as a novel methodology to quantify the similarity of pairs of evaluation functions based on the values they yield on different timetables.

We empirically show that the choice of the evaluation function can have a significant impact on the assessed quality of timetables, and thus also on which timetable is considered optimal, even though all evaluation functions are meant to evaluate the same - the quality of a timetable from passengers' perspective. Due to the structure of the designed evaluation functions, it is further possible to identify which components of the functions influence the results of an evaluation and under which conditions they this is most pronounced. This can be very beneficial when designing timetable evaluation functions for passengers.

Keywords: Public transport evaluation, timetable evaluation, consistent evaluation, comparison of evaluation functions

*hartleb@rsm.nl, johann.hartleb@isv.uni-stuttgart.de, +31 10 4082403

1 Introduction

When providing public transport, operators should aim for the highest possible quality from passengers' viewpoint, respecting physical and monetary constraints. However, there are many different definitions for 'quality from passengers' viewpoint'. The literature on public transport planning, both from Transport Engineering and Operations Research perspectives, as well as practitioners in railway companies, have come up with very different evaluations of quality . These range from very basic functions designed to be used in linear programming frameworks to sophisticated multi-variable models optimized to fit observed passenger behavior as well as possible.
In this paper, we investigate the following question: Considering a situation characterized by demand for public transport, and different public transport services provided to satisfy this demand, to what extent do different evaluation functions agree on the quality of the provided transport services? That is, will the evaluation functions considered - all designed to measure 'quality from passengers' perspective' - lead to the same evaluation of what is a good or a bad timetable?

We give an overview of different evaluation functions for timetables proposed in literature and identify three components in which the functions differ from each other. Based on this, we classify the considered evaluation functions and design a set of representative evaluation functions that differ from each other in the three components. These functions cover a wide range of the most commonly used evaluation functions in mathematical models, in evaluation applications or in choice models. Moreover, their modular structure as combination of the three components allows a purposeful analysis of their similarity.
To empirically compare these representative functions with each other and analyze how similar they are, we conduct three case studies. In each case study, we evaluate a set of possible timetables for a given demand situation with each of the representative evaluation functions. Two sets of timetables are defined for an artificial grid network and one set is defined for the real-world network of Netherlands Railways (NS). Since the sets of timetables are designed by different parties with varying methods and various objectives, the comparison of the functions should not be biased by the way the timetables were created.
Based on the resulting evaluation values of all timetables with respect to each evaluation function, we develop a methodology to quantify the degree to which the different evaluation functions coincide. The result of the methodology allows a pairwise comparison of the evaluation functions and can be interpreted as a measure of inconsistency, which we investigate in two ways. First, the pairwise inconsistency is interpreted directly, visualized with the help of heat maps and multidimensional scaling. This gives an overview of the extent of inconsistency between the evaluation functions and allows an immediate recognition of patterns of which evaluation functions are more or less consistent with each other. Second, we use cluster analysis to determine the strongest inconsistencies between the functions. The cluster analysis identifies groups of evaluation

functions that are consistent while the evaluation functions in different groups are less consistent.

The contribution of this paper is twofold. First, we develop an empirical methodology to compare multiple evaluation functions on a set of timetables. Since this methodology is independent of the structure of the evaluation functions, it can be applied to empirically compare evaluation functions in other applications as well. Second, we provide a thorough comparison of timetable evaluation functions for passengers. Our analysis shows whether and under which circumstances a component of a sophisticated evaluation function is crucial for the result of an evaluation. This can be used to either justify the simplifications made in current state-of-the-art optimization approaches to public transport planning, or to point out which aspect is still lacking and needs to be incorporated to obtain objective functions providing a valid evaluation. An earlier version of the findings was presented at the Symposium in Rail Transport Demand Management 2018 in Darmstadt, Germany.

The remainder of this paper is organized as follows. In Section 2 we give an overview of evaluation functions that are commonly used to measure the quality of public transport from passengers' viewpoint. Afterwards, in Section 3 we structure the evaluation functions used in the literature and define a set of representative evaluation functions which we use for the analysis in this research. In Section 4 we describe the data which we use in the case studies. Section 5 introduces a novel measure of inconsistency of evaluation functions and gives insight on the used methodology of comparison. We report on the main findings of our experiments in the same section. In Section 6 it is briefly demonstrated how the results can be used at the design of an evaluation function. The paper concludes in Section 7.

2 Literature on evaluation functions

Naturally, research concerned with the design of public transport also deals with the corresponding evaluation. There are various evaluation functions proposed in different research areas. Since we focus on the evaluation of public transport from passengers' point of view, we restrict ourselves to these evaluation functions. An overview of the most important factors of influence for timetable evaluation for passengers is given by Parbo et al. (2016). We consider only the planned case and neither disruptions nor robustness measures are considered, following the motto that "time savings are the single most important benefit of transport improvement projects all over the world" (Dios Ortuzar and Willumsen, 2011). In this section, we give an overview of different evaluation functions for timetables structured by the different components of timetable evaluation.

2.1 Types of evaluation functions

First, there exist many different ways to evaluate public transport. These differ from each other in the incorporated characteristics and the structure of the functions. We distinguish between two principally different types of evaluation functions, where each of them can appear in different variations.

On the one hand, most commonly used are travel time based evaluation functions. This is the default way of evaluation in both the research areas of Operations Research and Traffic Engineering. The key idea is to quantify the quality of public transport for passengers by a travel time equivalent. Travel time based evaluation functions are typically linear functions of passengers' travel time, but they vastly differ in the number and kind of incorporated characteristics (Hensher and Button, 2007). In Operations Research, the timetabling models are mostly based on the periodic event scheduling problem introduced by Serafini and Ukovich (1989) and often use the absolute time passengers spend in public transport for evaluation. In advanced evaluation functions the travel time is usually "subdivided into walking time, waiting time, time on vehicle, transfer time, and concealed waiting time" (Flyvbjerg et al., 1986). Furthermore, travel time based evaluation functions often take more influential factors into account, among them fare, frequency or temporal spread of the connections offered to passengers. In this case, they are mostly referred to as perceived travel time, generalized cost or disutility. Sometimes, also preferred departure or arrival times of passengers are modeled by penalizing early or late departures or arrivals. Kanai et al. (2011) considered late departures to be equivalent to waiting times for transfers and Robenek et al. (2016) introduced additional variables and penalty terms for the modeling of departure time preferences.

A comprehensive overview of generalized cost as evaluation functions can be found in Dios Ortuzar and Willumsen (2011). Both in research and practice the generalized costs are commonly used for evaluation purposes, although since long time there have been many publications recommending to stop using them to evaluate the quality of timetables from passengers' point of view. For instance, Grey (1978) discussed five aspects why generalized cost are unsuitable for evaluation, all following the same argument that depicting peoples' variety of perceptions in a single variable leads to an inaccurate representation.

On the other hand, we consider utility based evaluation models that are mainly known from research in choice modeling. The difference to travel time based evaluation models is that the evaluation value is not a travel time equivalent but follows the concept of passenger supplement. That means, each reasonably good connection for passengers adds to the utility and thus improves quality of the service. A comprehensive overview of utilities of alternatives is given in Ben-Akiva and Lerman (1985). Utility based evaluation functions are still almost exclusively found in choice modeling, although several publications proposed to employ them for evaluation purposes as well. For example, Jong et al. (2007) concluded that the 'logsum', a utility based evaluation function, is well suited for evaluation and a probable reason for their little success is the seemingly complex theory behind it - in contrast to travel time based evaluation functions.

2.2 Passenger distribution

Second, the assumed passenger distribution model is crucial for the evaluation. To evaluate the quality of a public transport service in a meaningful way, it is important to estimate how passengers will use it, that means, it is important to estimate how passengers distribute over the offered routes. The applied passenger distribution models in timetable related research range from very simple assumptions to highly developed choice models. In Operations Research, it is often assumed that passenger routes are known before the timetable is fixed and most publications use a priori fixed passenger loads on the connections (Nachtigall, 1998; Liebchen, 2018). Recently, there is a change in the timetabling literature visible with more publications focusing on an integrated timetable dependent passenger distribution. Since the connections passengers choose are not always reliably determinable beforehand, a shortest path search was included in timetabling models. Schmidt and Schöbel (2015) did that for the aperiodic case, Borndörfer et al. (2017) for the periodic case and Gattermann et al. (2016) also for the periodic case using a satisfiability formulation instead of a periodic event scheduling formulation. In these cases, the total travel time of all passengers on their shortest connections are evaluated, instead of the travel time on a previously defined connection. While it is often assumed that passengers only use a single route for each origin-destination pair, some timetabling papers specifically focus on a realistic distribution of the passengers. For instance, Parbo et al. (2014) developed an iterative approach where the passenger distribution is adjusted in each iteration of timetable optimization and Sels et al. (2013) used passenger loads derived from ticket sales data. We are not aware of an integrated search for multiple routes, most probably due to the high complexity of such a model.

In contrast to that, research in Traffic Engineering primarily applies passenger distribution models including multiple routes for passengers. Commonly applied for passenger route choice are choice models like the rooftop (Guis and Nijënstein, 2015), probit (Yang and Lam, 2006) or logit model. The theory of choice models is explained in Ben-Akiva and Lerman (1985) and an overview of choice models suited for passenger route choice in transit networks can be found in Dios Ortuzar and Willumsen (2011). It seems that the logit model is capable of depicting the passenger behavior best and it is therefore found most regularly. For example, Friedrich et al. (2001) designed an efficient algorithm based on the logit model to compute passenger distributions in public transport networks. Recently, Espinosa-Aranda et al. (2018) proposed a new formulation with estimation of a constrained nested logit model for connection choice in public transport.

2.3 Passenger preferences

Third, the evaluation of public transport services should be suited to the target group, which is in this case the passengers. Therefore, it is important that their preferences are reflected in the evaluation functions. These are expressed by parameters to tune the evaluation functions. Wardman and Toner (2018) showed in their analysis for the case of generalized cost that choosing the correct parameters is essential for a correct evaluation. While research in Operations Research focuses on developing algorithmic methods to compute timetables and mainly uses given or estimated parameters, there is much research in Traffic Engineering and choice modeling on parameter identification.

Usually, the parameters are found by either stated preference or revealed preference approaches. In the first case, people are asked to take decisions in a survey and their theoretical choice is used to derive rules for passenger behavior. For example, Bradley and Gunn (1990) determined the value of travel time of the Dutch population by a stated preference survey. In the second case, the actual decisions of passengers are generalized. Recently, with more data being available, more publications analyze passenger behavior with revealed preferences approaches. For example, Kusakabe et al. (2010) estimated passenger usage patterns from smart card data.

The most important parameters for public transport evaluation are of two different kinds, modeling passengers' preferences and passengers' behavior. The preference parameters specify how the different components of the passenger's journey are weighted. Different components include, but are not limited to waiting time, in-vehicle time or transfer time. Dell'Olio et al. (2010) provides passenger preference parameters measured from a bus transport service and Schittenhelm (2013) lists preferences of passengers of the Copenhagen S-train. A collection of multiple parameter settings found in various publications is published in Wardman (2004). Part of the passenger preferences but usually individually researched is the value of time, with many publications determining values under certain conditions, see for example Wardman et al. (2012). Parameters for passengers' behavior refer to the parameters used in the passenger distribution model, for example the logit parameter. The importance of correct parameter modeling for logit models is stressed in Swait and Louviere (1993).

2.4 Comparison of evaluation functions

Although there are various approaches to evaluate public transport from passengers' point of view, there is only limited research comparing different evaluation functions. Publications undertaking a comparison of evaluation functions mostly compare two evaluation functions only, a newly introduced function and the state of the art. Usually, the purpose is either to illustrate the merits of the newly introduced evaluation function, as it was done in the previously discussed integrated shortest path search (Schmidt and Schöbel, 2015; Borndörfer et al., 2017; Gattermann et al., 2016), or to better fit the evaluation to reality. As example for the latter, Jong et al. (2007) showed that in their case study a logsum based evaluation should be preferred to the currently applied evaluation since it is more precise in computing passengers' surplus when changing the public transport service. Some publications undertake a comparison of multiple evaluation functions, however, these are limited to a theoretical comparison. For example, Parbo et al. (2016) provides a literature review on public transport evaluation and focuses on the conflict of passenger's versus operator's focus. We are not aware of an empirical comparison of public transport evaluation functions or of an investigation of their inconsistency, which are the topics of this paper.

3 Timetable evaluation

From the literature review it is apparent that there are many different evaluation functions in use to measure quality of timetables from passengers' viewpoint. In this section, we classify this multitude of different functions and design a set of 16 evaluation functions to represent the evaluation functions in use.
For this purpose, we now define the terms important for the design of evaluation functions. All variables introduced are summarized in Appendix A. Passengers' demand is specified by a set of *origin-destination (OD) pairs* OD, where each of them is a directed pair of stations in the public transport network with time-dependent demand. We consider disjoint *time slices* $t \in T$ of one hour and define the hourly demand of passengers that want to depart in time slice $t \in T$ for each OD pair to be o^t_{od}. The sum of all hourly demand equals the daily demand o_{od} of each OD pair, i.e., $\sum_{t \in T} o^t_{od} = o_{od}$. To meet the demand of passengers, each timetable offers connections to the passenger. We use the term *connection* to denote a time-bound route for passengers using public transport services and denote a set of reasonable connections for each OD pair od with preferred departure time slice t by C^t_{od}. To evaluate timetables, we follow the usual approach to measure and aggregate the quality of available connections for the passengers.

3.1 Quality measurements

We give a brief introduction into important characteristics of connections for the passengers, explain how we aggregate these characteristics to the network level and define four quality measurements for timetables. The four quality measurements are based on the characteristics and reflect the evaluation functions found in the literature.

Basic characteristics of a connection

To evaluate the public transport services, we quantify five characteristics of a connection c as listed in Table 1. These characteristics are important factors of influence for a passenger's decision whether to travel on connection c or not. Note, that we do not take the fare of connections into account. We assume a fare system where the fares depend on origin and destination only, as used, e.g., at Netherlands Railways (NS), the largest Dutch railway operator. Consequently, in such a system the cost for each OD pair od is constant and does not affect the attractiveness of connections.

$\mathrm{IVT}(c)$	In-vehicle time	The time spent in public transport vehicles
$\mathrm{WKT}(c)$	Walk time	The time spent walking between platforms for a transfer
$\mathrm{TWT}(c)$	Transfer wait time	The time spent at a station waiting for the next connecting public transport vehicle
$\mathrm{NTR}(c)$	Number of transfers	The number of transfers in the connection
$\mathrm{DEP}(c)$	Departure time	The departure time at the origin

Table 1: Characteristics of connections

Derived characteristics of a connection

Using these characteristics of connections, we derive more characteristics of connections which provide a basis for commonly used evaluation functions. One derived characteristic that is commonly found in publications in the field of Operations Research to quantify the quality of a connection is the *absolute travel time* (ATT). The absolute travel time is defined as the sum of in-vehicle time and the transfer time, consisting of walk time and transfer wait time:

$$\mathrm{ATT}(c) \coloneqq \mathrm{IVT}(c) + \mathrm{WKT}(c) + \mathrm{TWT}(c).$$

More general, many evaluation functions used in literature apply a weighting of travel times of the different trip segments and include a penalty for transfers. To model this weighting, we define the *perceived journey time* (PJT) as

$$\mathrm{PJT}(c) \coloneqq \mathrm{IVT}(c) + \alpha_{\mathrm{WKT}} \cdot \mathrm{WKT}(c) + \alpha_{\mathrm{TWT}} \cdot \mathrm{TWT}(c) + \alpha_{\mathrm{NTR}} \cdot \mathrm{NTR}(c) \quad (1)$$

With the coefficients

$$\alpha_{\mathrm{WKT}}, \alpha_{\mathrm{TWT}}, \alpha_{\mathrm{NTR}} \in \mathbb{R}_{\geq 0}$$

it is possible to model different passenger preferences and the perceived journey time can be interpreted as a time equivalent expressing how long the connection c feels to a passenger.

Some publications include departure time preferences of passengers in their evaluation. To model these preferences, we introduce the *adaption time* (ADT), which is the time a passenger has to deviate from their preferred departure time slice t to take connection c. Each time slice t corresponds to a one hour interval $[\,\underline{t}, \bar{t}\,)$ of preferred departure time. Let $\hat{t} \in t$ be a time point in the time slice $t = [\,\underline{t}, \bar{t}\,)$, then the adaption time is defined as

$$\mathrm{ADT}^{t}(c) \coloneqq \mathrm{ADT}^{[\,\underline{t}, \bar{t}\,)}(c) \coloneqq \min_{\hat{t} \in [\,\underline{t}, \bar{t}\,)} \|\hat{t} - \mathrm{DEP}(c)\|.$$

The adaption time could similarly be defined for arrival times, however, for the sake of simplicity we restrict ourselves to an adaption time at departures only.

The adaption time is further explained in Example 1 for the case that the time window of preferred departure time equals the time slice t. To model stronger departure time preferences we split each time slice t in $\gamma \in \mathbb{N}$ time windows t_j of equal length, with

$$t = \dot{\bigcup_{j=1}^{\gamma}} t_j.$$

Then, we assume that o^t_{od}/γ passengers want to depart in each of the γ time windows and the adaption time generalizes to the average adaption time to the γ time windows, i.e.,

$$\mathrm{ADT}^t(c) = \frac{1}{\gamma} \sum_{j=1}^{\gamma} \mathrm{ADT}^{t_j}(c).$$

Using this, it is possible to include the impact of access time and spread of available connections in the evaluation. We define the *adapted journey time* $\mathrm{AJT}^t(c)$ of a connection c for all passengers with preferred departure time slice t by

$$\begin{aligned} \mathrm{AJT}^t(c) \coloneqq \mathrm{IVT}(c) + \alpha_{\mathrm{WKT}} \cdot \mathrm{WKT}(c) + \alpha_{\mathrm{TWT}} \cdot \mathrm{TWT}(c) \\ + \alpha_{\mathrm{NTR}} \cdot \mathrm{NTR}(c) + \alpha_{\mathrm{ADT}} \cdot \mathrm{ADT}^t(c). \end{aligned} \tag{2}$$

This number quantifies how unattractive a certain connection is perceived by a passenger who wants to start traveling in time slice t. We denote the passenger preferences by

$$\alpha \coloneqq (\alpha_{\mathrm{WKT}}, \alpha_{\mathrm{TWT}}, \alpha_{\mathrm{NTR}}, \alpha_{\mathrm{ADT}}) \in \mathbb{R}^4_{\geq 0}.$$

Note, that for $\alpha = (1, 1, 0, 0)$ the adapted journey time equals the absolute travel time ATT, and for $\alpha = (\alpha_{\mathrm{WKT}}, \alpha_{\mathrm{TWT}}, \alpha_{\mathrm{NTR}}, 0)$ the adapted journey time equals the perceived journey time PJT.

Furthermore, there also exist utility based evaluation functions in literature that are derived from choice models. To represent these functions, we consider the *evaluated total utility* of a connection (ETU) as a number expressing how useful a connection is to a passenger with preferred departure time slice t. We define the evaluated total utility of a connection c to be

$$\mathrm{ETU}^t(c) = e^{-\beta \cdot \mathrm{AJT}^t(c)}, \tag{3}$$

based on the definition of the logit model as a passenger distribution model. The logit model and its associated parameter $\beta \in \mathbb{R}_{\geq 0}$ will be explained in detail in Section 3.3.

Table 2 gives a theoretical comparison of the four quality measurements. If a characteristic is included linearly in a quality measurements, the table shows the coefficient, if the dependency is non-linear, it is only indicated by an asterisk whether the characteristic is taken into account.

	IVT	WKT	TWT	NTR	ADT
absolute travel time ATT	1	1	1	0	0
perceived journey time PJT	1	α_{WKT}	α_{TWT}	α_{NTR}	0
adapted journey time AJT	1	α_{WKT}	α_{TWT}	α_{NTR}	α_{ADT}
evaluated total utility ETU	*	*	*	*	*

Table 2: Each entry indicates which of the five characteristics (in-vehicle time IVT, walk time WKT, transfer wait time TWT, number of transfers NTR and adaption time ADT) are taken into account in the four quality measurements ATT, PJT, AJT and ETU. Linear dependencies are indicated by coefficients, non-linear by asterisks.

Example 1. *We consider an example network with four stations, O, A, B and D and one OD pair from O to D. The public transport service provides a (slow) bus line driving from O to D via A and B. Let there be two buses departing at* 9:05 *and* 10:05*, arriving in A at* 9:20 *and* 10:20 *and in D at* 9:47 *and* 10:47*, respectively. In addition to the bus, there is a direct train connecting A and D, without a detour via B. Let the train depart at* 9:23 *from A and arrive at* 9:38 *in D. In this example we are only interested in the time slice t_9^{10} between* 9 *am and* 10 *am and assume that all passengers want to depart during that time, i.e., $o_{OD} = o_{OD}^{t_9^{10}}$. A layout of the network can be found in Fig. 1.*

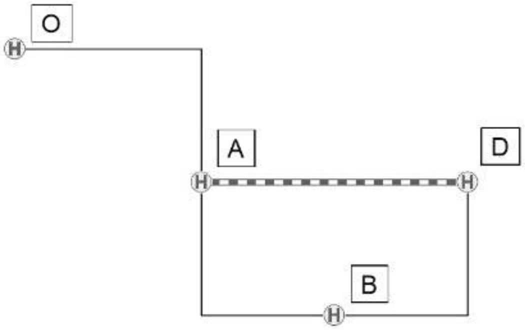


Fig. 1. Exemplary public transport network

This public transport service offers three connections for passengers traveling from O to D:

c_1	Direct bus at 9:05
c_2	Direct bus at 10:05
c_3	Bus at 9:05 and transfer to train in A

Let the walk time at station A from the bus terminal to the train platform be 2 *min and assume passenger preferences of $\alpha = (1, 1, 10, 2)$ and $\beta = 0.22$. Then,* *the connections c_1, c_2, c_3 have the characteristic values as listed in Table 3.*

	measured					derived				
	IVT	WKT	TWT	NTR	DEP	ATT	PJT	ADT	AJT	ETU
c_1	42	0	0	0	9:05	42	42	0	42	$9.7 \cdot 10^{-5}$
c_2	42	0	0	0	10:05	42	42	5	52	$5.0 \cdot 10^{-5}$
c_3	30	2	1	1	9:05	33	43	0	43	$7.8 \cdot 10^{-5}$

Table 3: Measured and derived characteristics of three connections

Connections c_1 and c_3 have an adaption time of 0 *min since both depart between* 9:00 *and* 10:00. *The second connection has an adaption time of* 5 *min as it departs* 5 *min after the preferred interval. The same adaption time would result for a connection departing at* 8:55.

Characteristics for OD pairs

The goal is to design evaluation functions for a whole timetable but, so far, just characteristics of connections were defined. For the purpose of a timetable evaluation we aggregate the characteristics of connections.

As a first step, we aggregate the characteristics over all time slices t and connections c per OD pair. In this way we obtain the characteristic values for each OD pair. Let $p^t(c)$ be the probability that connection c is chosen by passengers with preferred departure time slice $t \in T$, i.e.,

$$\sum_{c \in C^t_{od}} p^t(c) = 1 \quad \forall t \in T$$

and

$$p^t(c) \geq 0 \quad \forall t \in T, c \in C^t_{od}.$$

How we derive realistic and meaningful values for this probability is outlined in Section 3.2. Let $X^t(c)$ be a characteristic of connection $c \in C^t_{od}$, with a value that possibly depends on the preferred departure time slice t. We consider $X^t(c) \in \{\mathrm{ATT}(c), \mathrm{PJT}(c), \mathrm{AJT}^t(c)\}$ and denote these as *travel time based* characteristics. Then the average value of that characteristic over all time slices $t \in T$ and connections $c \in C^t_{od}$ for the OD pair od is defined as

$$X_{od} := \frac{\sum_{t \in T} \left(o^t_{od} \sum_{c \in C^t_{od}} p^t(c) \cdot X^t(c) \right)}{\sum_{t \in T} o^t_{od}}. \tag{4}$$

To compute the characteristic value for OD pairs, this value is weighted with the probability $p^t(c)$ that a connection c is chosen, given the preferred departure time slice t. This can heavily influence the characteristic values as the next example shows.

Example 2. *We continue with the example network from above and assume two different scenarios for passenger distribution. In the first, all passengers travel on the shortest path w.r.t the absolute travel time. This yields* $p^{t_9^{10}}(c_1) = 0 = p^{t_9^{10}}(c_2)$ *and* $p^{t_9^{10}}(c_3) = 1$. *Thus, we obtain for the absolute travel time*

$$\begin{aligned} ATT_{OD} &= \frac{\sum_{t \in T} \left(o_{od}^t \sum_{c \in C_{od}^t} p^t(c) \cdot ATT(c) \right)}{\sum_{t \in T} o_{od}^t} \\ &= \frac{o_{od}^{t_9^{10}} \cdot 1 \cdot 33}{o_{od}^{t_9^{10}}} = 33, \end{aligned}$$

the average absolute travel time over all passengers from O *to* D *is* $33\,min$.
In the second scenario we assume that all passengers distribute uniformly over the available connections, independent of their characteristics. This corresponds to a distribution $p^{t_9^{10}}(c_1) = 1/3 = p^{t_9^{10}}(c_2)$ *and* $p^{t_9^{10}}(c_3) = 1/3$ *with an absolute travel time of*

$$\begin{aligned} ATT_{OD} &= \frac{\sum_{t \in T} \left(o_{od}^t \sum_{c \in C_{od}^t} p^t(c) \cdot ATT(c) \right)}{\sum_{t \in T} o_{od}^t} \\ &= \frac{o_{od}^{t_9^{10}} \cdot (\frac{1}{3} \cdot 42 + \frac{1}{3} \cdot 42 + \frac{1}{3} \cdot 33)}{o_{od}^{t_9^{10}}} = 39. \end{aligned}$$

Since the passengers distribute over the connections, the absolute travel time increases from $33\,min$ *to* $39\,min$ *compared to the case where all passengers travel on the shortest connections. This example shows that the characteristics of the public transport service for passengers heavily depends on the passenger behavior* $p^t(c)$.

Furthermore, we define the evaluated total utility for passengers as

$$\mathrm{ETU}_{od} := \frac{\sum_{t \in T} \left(o_{od}^t \sum_{c \in C_{od}^t} \mathrm{ETU}_{od}^t(c) \right)}{\sum_{t \in T} o_{od}^t}. \tag{5}$$

This characteristic is independent of the passenger distribution $p^t(c)$ since the evaluated total utility of each connection $\mathrm{ETU}_{od}^t(c)$ is derived from the logit model which we use for choice connection. However, the assumed passenger distribution model might affect which connections are considered to be reasonable alternatives in the set C_{od}^t, which will be addressed in Section 3.2.

Characteristics of the public transport service

In accordance with many evaluation functions used in research, we define the characteristics of the public transport service X to be the weighted average of the characteristics for OD pairs, computed by

$$X := \frac{\sum_{od \in OD} o_{od} \cdot X_{od}}{\sum_{od \in OD} o_{od}},$$

for $X_{od} \in \{\mathrm{ATT}_{od}, \mathrm{PJT}_{od}, \mathrm{AJT}_{od}, \mathrm{ETU}_{od}\}$. We refer to the corresponding four characteristics of a timetable $X \in \{\mathrm{ATT}, \mathrm{PJT}, \mathrm{AJT}, \mathrm{ETU}\}$ as *quality measurements*. While the first three quality measurements are travel time equivalents, the evaluated total utility is a utility based evaluation function, where each reasonably good connection for passengers adds to the utility and thus improves quality of the service. Hence, we refer to ATT, PJT and AJT as *travel time based*, and to ETU as *utility based* quality measurement.

3.2 Passenger distribution

The decision which connections passengers choose is dependent on the characteristics of the connections. There are two fundamentally different approaches for passenger distribution used in the literature. While research in Operations Research often assumes that all passengers travel on the shortest connection available, most publications from other research areas apply more realistic passenger distribution models when evaluating timetables. To investigate this difference, we consider two passenger distribution models.

On the one hand, we rely on the logit model to obtain a realistic distribution of the passengers on multiple connections (*mc*). We assume the set C_{od} of reasonably good connections for OD pair *od* to be given. Then, the logit model can be interpreted as a function assigning a probability $p^t(c)$ to each connection $c \in C_{od}$ that it is used by passengers with preferred departure time slice t. The logit model is defined by

$$p^t(c) = \frac{e^{-\beta \cdot \mathrm{AJT}^t(c)}}{\sum_{c' \in C_{od}} e^{-\beta \cdot \mathrm{AJT}^t(c')}} \qquad \forall c \in C_{od}, \tag{6}$$

where the parameter $\beta \in \mathbb{R}_{\geq 0}$ is used to adjust the model to a specific case study. Note, that the choice set of connections C_{od} is independent of the preferred departure time slice t of the passengers. Since the logit model is based on the adapted journey time of all considered alternative connections, only connections departing in or close to the time slice t will be assigned a probability that is significantly larger than 0. This is a common way to model connection choices of passengers realistically (Ben-Akiva and Lerman, 1985).

On the other hand, we consider a shortest connection (*sc*) strategy for the passengers. That means, passengers only take connections with lowest journey time departing within or close to their preferred departure time slice. Let C^t_{od} be the set of all connections with lowest adapted journey time for passengers of an OD pair *od* that want to depart in time slice t. Then, the share of passengers using connection $c \in C^t_{od}$ is

$$p^t(c) = \frac{1}{|C^t_{od}|} \qquad \forall c \in C^t_{od}.$$

That means, in case there are multiple shortest connections available, we assume that the passenger distribute on them uniformly.

We want to point out that the passenger distribution also affects the quality measurements defined in Section 3.1. As mentioned, with the two possible passenger distribution models *sc* and *mc* we also assume two different choice sets to be given. For the logit distribution on multiple connections, we assume a set of reasonably good connections to be given and for the shortest connection assumption, we require the choice set to only contain connections with lowest adapted journey time. The respective choice set is also used for aggregating the characteristics in Equations 4 and 5, depending on which passenger distribution model will be applied.

3.3 Passenger preferences and behavior

We take different assumptions on passenger preferences and passenger behavior into account. In the definition of the perceived and adapted journey time of a connection in Equations 1 and 2 many passenger preferences are contained. The values of the coefficients $\alpha \in \mathbb{R}^4_{\geq 0}$ indicate how important in-vehicle time, walk time, transfer wait time, number of transfers and adaption time are (relative to each other) to the passenger.
In addition, we have the logit coefficient $\beta \in \mathbb{R}_{\geq 0}$ in Equation 6 to tune the sensitivity of passengers to absolute differences in the adapted journey time of connections. For example, for $\beta = 0$ all connections in the choice set will be used by passengers equivalently and the logit model reduces to a uniform distribution. The higher the coefficient β, the more passengers will use the connections with lowest adapted journey time. This coefficient also influences the evaluated total utility of a public transport service, as defined in Equation 3.
Furthermore, it is possible to set the passenger tolerance $\gamma \in \mathbb{N}$ for early or late departure. High values for γ indicate a low tolerance and already slight deviations from the preferred departure time are penalized.
To analyze the impact of modeling passenger preferences on the evaluation, we consider two user groups. These are represented by the two parameter settings

$$ps_1 = (\alpha, \beta, \gamma) \text{ with } \alpha = (1, 1, 5, 1),\ \beta = 0.13,\ \gamma = 1 \tag{7}$$

and

$$ps_2 = (\alpha, \beta, \gamma) \text{ with } \alpha = (2, 2, 10, 2),\ \beta = 0.22,\ \gamma = 60. \tag{8}$$

The first parameter setting models passengers that are mainly focused on journey time ($\alpha_{\text{WKT}} = 1, \alpha_{\text{TWT}} = 1$) and are relatively undeterred by transferring ($\alpha_{\text{NTR}} = 5$). They would also make use of connections with higher adapted journey time ($\beta = 0.13$) and are rather flexible regarding departure time ($\alpha_{\text{ADT}} = 1,\ \gamma = 1$), as long as the connections are fast.
The second parameter setting models passengers that are more convenience oriented. They prefer a public transport service that is suited to their needs with less and short transfers ($\alpha_{\text{WKT}} = 2, \alpha_{\text{TWT}} = 2, \alpha_{\text{NTR}} = 10$), preferably use connections with low adapted journey time ($\beta = 0.22$) and are inflexible regarding their desired departure time ($\alpha_{\text{ADT}} = 2,\ \gamma = 60$).

The parameters are chosen following recommendations from research and practice. For example, as of 2012, the NS used a penalty of 10 min for each transfer (De Keizer et al., 2012). Wardman (2004) provides a thorough study of values of time, among them several values for wait and walk time compared to in-vehicle time are listed. Usually, the coefficients for wait and walk time are around 2. The logit parameter β should be adjusted for each case study, but experience has shown that values of $\beta \in [0.13, 0.22]$ are a reasonable choice if minutes are used as time units. The adaption time is modeled to fit the characteristics of the two parameter sets from Equations 7 and 8.

3.4 Evaluation functions

We define an *evaluation function* as a combination of a quality measurement, a passenger distribution model and an assumption on passenger preferences. That means, applying an evaluation function consists of two steps: Given a timetable with a connection choice set for passengers, the passengers are first distributed on the connections according to the distribution model and their preferences. Second, the quality of the timetable is evaluated with respect to the quality measurement, again using the passenger preferences. Many publications focus only on the second step when describing their evaluations. However, we believe that the distribution is an integral component of the evaluation which influences the evaluation results. Hence, we also investigate the extent of this influence. When combining the four quality measurements defined in Section 3.1 with the two distribution models described in Section 3.2 and the two different assumptions on passenger preferences fixed in Section 3.3, we obtain 16 evaluation functions in total. This design of evaluation functions entails two advantages.

	sc	*mc*
ATT	Borndörfer et al. (2017)	Parbo et al. (2014)†
PJT	Wardman and Toner (2018)	Parbo et al. (2014)†
AJT	Kanai et al. (2011)	Robenek et al. (2016)
ETU	‡	Jong et al. (2007)§

Table 4: Examples for the use of different evaluation functions in recent literature. We provide one publication for each cell, exemplifying the use of a quality measurement (absolute travel time ATT, perceived journey time PJT, adapted journey time AJT and evaluated total utility ETU) in combination with a shortest connection (*sc*) or multiple connection (*mc*) passenger distribution model

† Used ATT in evaluation and PJT in distribution

‡ ETU in combination with *sc* is not used since ETU does not require a passenger distribution

§ Used slightly different utility based evaluation function

These functions cover a wide range of commonly used evaluation functions in mathematical models, in evaluation applications or in choice models as it is indicated in Table 4. In addition to that, their modular structure as combination of quality measurement, distribution model and assumptions on passenger preferences allows a purposeful analysis. Differences or similarities of evaluation functions can easily be traced down to a single component of the functions. We denote the set of the 16 evaluation functions by F.

4 Case studies

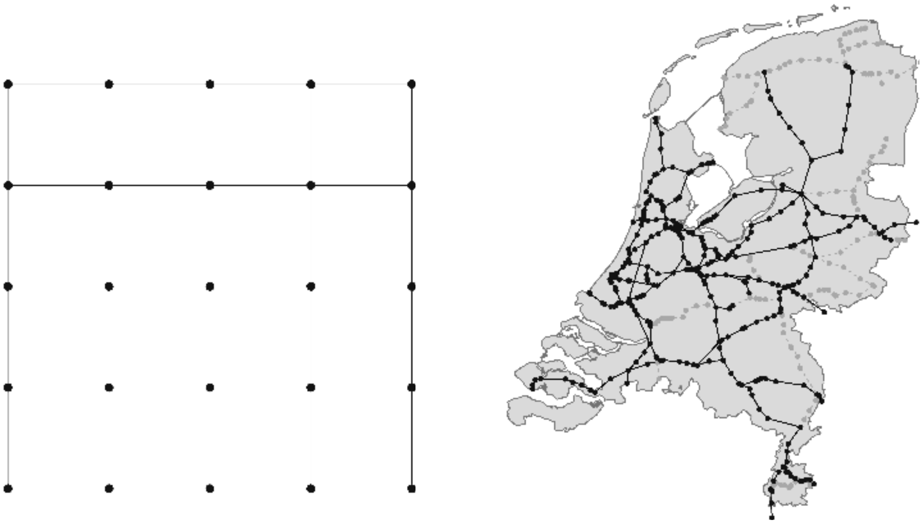
Our goal is to analyze how inconsistent the 16 different evaluation functions are by comparing their evaluation behavior on multiple public transport services. In this section, we describe three case studies in which we perform these evaluations. Each *case study* is characterized by a fixed public transport infrastructure, a demand situation on that infrastructure and a set of services supplying the demand. A *public transport infrastructure* consists of stations and direct links between them and a *demand situation* is specified by a set of origin-destination (OD) pairs OD, where each of them is a directed pair of stations with time-dependent demand. For this demand situation, we consider several public transport services supplying this demand, for comparison. Each of the *public transport services* is formalized by a line plan and a timetable which together determine the potential connections and their quality.

4.1 Case studies on a grid infrastructure

As a first infrastructure we use an artificial 5×5 grid network[1] introduced by the research group FOR2083. The infrastructure consists of 25 stations and 40 direct links and the network is depicted in Fig. 2a.

On this infrastructure we consider two demand situations with multiple corresponding benchmark services available, each of them consisting of a line plan and a timetable. Both demand situations have an almost complete demand matrix with nearly 600 non-zero entries. Although they share the same infrastructure, we treat them as two different case studies due to the different data structure of demand and supply. The first demand situation has a typical daily demand pattern and 27 suitable services that are operated throughout the whole day. All of these services were designed by traffic engineers with established methods used in transport planning. We refer to the case study as GL. The second demand situation depicts a morning peak and 28 services operating only in the morning hours are available. These service were found with different optimization models by Operations Researchers and we denote the corresponding case study by GS. The names of all 55 services as used in this paper and their corresponding names in the repository can be found in Appendix B.

[1] `https://github.com/FOR2083/PublicTransportNetworks/tree/master/Grid_5x5`, accessed November 12, 2018.



(a) Grid infrastructure

(b) Dutch railway infrastructure. Tracks and stations in black are operated by NS

Fig. 2. The evaluation functions are compared on these two infrastructures

4.2 Case study on the Dutch railway infrastructure

The second infrastructure is the Dutch railway network with roughly 270 stations as it is operated by Netherlands Railways (NS). In Fig. 2b a route map of the Dutch railway network is shown. The demand is given by a scientific demand set of more than 62000 non-zero OD pairs defined between the stations reflecting a realistic demand situation. For evaluation we consider the yearly transport services that were operated by NS in the years 2012 till 2018. Note, that due to changes in the infrastructure in the Dutch railway network, not all public transport services are defined on the same network. That means, over the years some stations and tracks might have been introduced or abolished. However, we evaluate all different services with the same demand set between the same stations, therefore the evaluation is not directly affected by the slight changes of the infrastructure. We refer to this case study by NS.

4.3 Derivation of a connection choice set

In all case studies multiple services are considered, each of them consisting of a line plan and a timetable. The evaluation functions assume a set C^t_{od} of reasonable connections for each OD pair *od* with preferred departure time slice t to be given. In this section we describe how we derive such sets from a given public transport service. To ensure better comparability of the evaluation, we derive the same choice sets for all evaluation functions within each case study.

In Section 3.2 we remark that two different connection choice sets are assumed, depending on the applied passenger distribution model. In case of a distribution on multiple connections with the logit model, we assume that a set C_{od} of reasonably good connections for OD pair *od* is given. When all passengers are assigned to the shortest connections, we assume that the set C^t_{od} of all connections with lowest adapted journey time for passengers of OD pair *od* that want to depart in time slice t is given.

Choice set for logit model

To obtain a set with all reasonably good connections for an OD pair, we consider all connections with low absolute travel time, low perceived journey time and low number of transfers. The perceived journey time of the connections is compared using the fixed parameters

$$(\alpha_{\mathrm{WKT}}, \alpha_{\mathrm{TWT}}, \alpha_{\mathrm{NTR}}) = (1.5,\, 1.5,\, 7.5).$$

These values are the arithmetic mean of the values used for α in the two parameter settings ps_1 and ps_2. In addition, we use parameters δ_{PJT}, δ_{ATT}, $\varepsilon_{\mathrm{PJT}}$, $\varepsilon_{\mathrm{ATT}}$ and $\varepsilon_{\mathrm{NTR}}$ to decide whether a connection is good enough to be considered. Then, the choice set C_{od} contains

- all connections c which have at most an absolute travel time $\mathrm{ATT}(c)$ with

  $$\mathrm{ATT}(c) < \delta_{\mathrm{ATT}} \cdot \mathrm{ATT}(c') + \varepsilon_{\mathrm{ATT}}$$

  where c' is the connection with the lowest possible absolute travel time for OD pair *od*,

- all connections c which have at most a perceived journey time $\mathrm{PJT}(c)$ with

  $$\mathrm{PJT}(c) < \delta_{\mathrm{PJT}} \cdot \mathrm{PJT}(c') + \varepsilon_{\mathrm{PJT}}$$

  where c' is the connection with the lowest possible perceived journey time for OD pair *od* and

- all connections c which have at most $\mathrm{NTR}(c)$ transfers with

  $$\mathrm{NTR}(c) < \mathrm{NTR}(c') + \varepsilon_{\mathrm{NTR}}$$

  where c' is the connection with the lowest possible number of transfers for OD pair *od*.

For the derivation of choice sets for the analysis we use the values

$$\delta_{\mathrm{PJT}} \coloneqq 1.5,\ \delta_{\mathrm{ATT}} \coloneqq 1.5,\ \varepsilon_{\mathrm{PJT}} \coloneqq 10,\ \varepsilon_{\mathrm{ATT}} \coloneqq 10 \text{ and } \varepsilon_{\mathrm{NTR}} \coloneqq 1.$$

All dominated connections are removed from the choice sets. A connection $c \in C_{od}$ is dominated by another connection $c' \in C_{od}$ if

- connection c' starts simultaneously or later and arrives simultaneously or earlier than connection c, and
- connection c' has at most as many transfers as c, and
- the perceived journey time of connection c' is at most as high as the perceived journey time of connection c and
- at least one of the three conditions is a strict inequality

Since the search is independent of the time slice t, the choice set C_{od} contains all reasonably good connections for the OD pair during the whole analysis period T. As mentioned before, the logit model assigns a share of passengers significantly different from 0 only to those connections with low adaption time.

Choice set for shortest connections

For the assumption that all passengers use the shortest connections only, one choice set for each departure time slice t is required. We define these to be the subset of the choice set C_{od} with reasonably good connections, containing only the connections with minimal adapted journey time, i.e.,

$$C_{od}^{t} \coloneqq \{c \in C_{od} \colon \ \mathrm{AJT}^{t}(c) \leq \mathrm{AJT}^{t}(c') \quad \forall c' \in C_{od}\}.$$

5 Comparison of evaluation functions

We defined 16 evaluation functions for public transport services in Section 3 and introduced the infrastructures with corresponding demand situation and multiple services for the three case studies in Section 4. In this section we describe a methodology to compare different evaluation functions and to set them into relation. Using this methodology, the 16 evaluation functions are investigated for their inconsistency in the three case studies.

The key idea is to compare the evaluation functions when applied to a number of services. We evaluate all public transport services $s \in S$ with each of the evaluation functions $f \in F$ and use the resulting evaluation values v_s^f to compare the functions in F. We conduct the evaluation of the services with PTV Visum (PTV Planung Transport Verkehr AG, 2017), a software package for macroscopic traffic analysis and forecasting. For the NS case study, Table 5 shows the evaluation values of the services operated between 2012 and 2018 for all 16 evaluation functions.

At a first glance, all public transport services in Table 5 have very similar evaluation values, suggesting that the quality of the services is effectively the same. For example, the absolute travel time on the shortest connection evaluated with the first parameter setting (evaluation function 1) ranges for all seven services between 35.94 and 36.78 minutes, implying a difference of only 0.84 minutes. While this difference sounds negligible, it actually comprises considerable differences for individual OD pairs. A total gain of 0.84 minutes in absolute travel

	ATT				PJT				AJT				$100 \cdot$ ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
NS12	36.78	37.43	36.79	37.60	37.99	38.90	41.04	42.22	38.64	40.49	58.37	53.30	38.31	55.98	1.29	1.77
NS13	36.30	36.96	36.32	37.13	37.44	38.38	40.27	41.58	38.07	39.97	57.60	52.63	34.19	53.60	1.11	1.70
NS14	36.30	36.94	36.31	37.09	37.44	38.37	40.31	41.60	38.03	39.96	56.88	52.22	36.83	56.50	1.21	1.80
NS15	36.22	36.90	36.24	37.07	37.36	38.33	40.22	41.51	37.98	39.93	56.88	52.01	36.68	56.25	1.20	1.77
NS16	36.23	36.91	36.26	37.06	37.38	38.32	40.24	41.52	37.99	39.92	56.90	51.97	38.28	55.93	1.28	1.75
NS17	36.03	36.77	36.04	36.96	37.25	38.29	40.28	41.67	37.87	39.90	56.85	52.03	40.44	57.67	1.33	1.80
NS18	35.94	36.71	35.95	36.89	37.14	38.22	40.14	41.56	37.78	39.83	56.96	51.75	39.72	59.16	1.31	1.85

Table 5: Evaluation values v_s^f in NS case study. Each column corresponds to one evaluation function $f \in F$ and each row to one public transport service s. The name of the services indicate the year in which this service was operated. The four topmost rows show the quality measurement, the used parameter setting and distribution model as introduced in Section 3 and lastly an index to identify the evaluation functions. The values for the travel time based evaluation functions (ATT, PJT, AJT) show average travel time in minutes, the values of the utility based evaluation functions (ETU) is dimensionless. For ease of exposition, all evaluation values of utility based evaluation functions are multiplied with 100.

time corresponds to an improvement of 2.3% and could for example be achieved by decreasing the travel time on all connections of the 20 biggest OD pairs by 10 minutes. This improvement would affect more than 90,000 travelers every day.

Furthermore, Table 5 also shows that the best service regarding one evaluation function is not necessarily the best service regarding another evaluation function. For example, the best services regarding evaluation functions 7 and 8 do not coincide. While NS18 provides on average the shortest perceived journey time weighted with the second parameter set on the shortest connection, NS15 yields the shortest perceived journey time on multiple connections, indicating that the passenger distribution model has an influence on the evaluation in this case. Table 6 summarizes differences in ranking of all public transport services and all evaluation functions in a 'medal count', indicating how often the respective service is classified on a certain rank.

	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}
NS12	0	0	2	0	2	0	12
NS13	0	0	0	2	1	9	4
NS14	0	1	2	0	11	2	0
NS15	1	1	7	5	0	2	0
NS16	0	2	4	8	0	2	0
NS17	3	10	0	1	1	1	0
NS18	12	2	1	0	1	0	0

Table 6: 'Medal count' from NS case study showing the number of times a public transport service is ranked on the n^{th} rank. Both row and column sum add up to 16, the number of considered evaluation functions.

The highest numbers in Table 6 appear on, or close to the antidiagonal. This shows that the evaluation functions essentially agree that the services improved from NS12 to NS18, or equivalently, improved over the years. Taking an average over all evaluations, it seems to be conclusive which service is best. However, not all of the services could be unambiguously classified. Most of the services are ranked over a range of five, some even over six ranks. Using just one evaluation function, as it is often done in research, might yield a very different ranking than the average suggests. The medal counts from the other two case studies GS and GL can be found in Appendix C. To draw inferences from this about the inconsistency of the evaluation functions, it is interesting to see whether the deviations in ranking are due to some random dispersion or whether there is a structural connection between the rankings of evaluation functions.

5.1 Inconsistency of two evaluation functions

Even when the differences in the ranking are large, actual evaluation values may be very close to each other. To avoid fallacy when comparing the evaluation functions by rank, we focus on the relative differences in objective values. Since the evaluation values v_s^f are dependent on the evaluation function and thus not directly comparable, we normalize the evaluation values. These normalized values are in the same number range and can be compared easily.
We define

$$V(f) \coloneqq \max_{s \in S} v_s^f - \min_{s \in S} v_s^f$$

to be the range of objective values of all public transport services with respect to evaluation function $f \in F$. For evaluation functions, for which smaller values are better, we define the normalized value of service $s \in S$ with respect to evaluation function $f \in F$ to be

$$\varphi_s^f \coloneqq \frac{v_s^f - \min_{s' \in S} v_{s'}^f}{V(f)}. \tag{9}$$

Equivalently, the normalized value for evaluation functions, for which larger values are better, is defined as

$$\varphi_s^f \coloneqq \frac{\max_{s' \in S} v_{s'}^f - v_s^f}{V(f)}. \tag{10}$$

The normalized values lie in the unit interval and indicate to what extent service s performs worse than the best service with respect to the same evaluation function considering the range of all other values. Therefore, the normalized values φ depend on the set of all considered services S of a case study. Tables with all normalized evaluation values for the three case studies are provided in Appendix D.
The normalized values allow a comparison of the quality of public transport services regarding different evaluation functions. To compare the evaluation functions pairwise with each other, we define the *inconsistency* of two evaluation

functions f_1 and f_2 as the mean difference in the normalized value, i.e.,

$$i_\varphi(f_1, f_2) := \frac{1}{|S|} \sum_{s \in S} |\varphi_s^{f_1} - \varphi_s^{f_2}|.$$

As the normalized values φ_s^f depend on the set of all considered services S of a case study, also the inconsistency i depends on the set S.

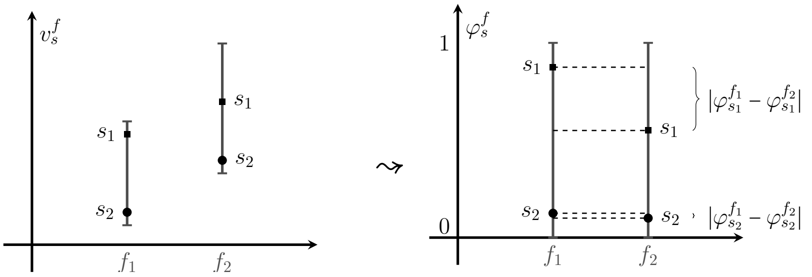


Fig. 3. Normalization of evaluation values v_s^f for two evaluation functions and indication of computation of inconsistency $i_\varphi(f_1, f_2)$ for two public transport services

For a better understanding, the normalization of evaluation values and the definition of the inconsistency as mean difference in normalized values is depicted in Fig. 3. The graph on the left shows the ranges of the evaluation values v_s^f of two evaluation functions f_1 and f_2 as vertical lines. On the lines the evaluation values of two services s_1 and s_2 are marked. As it can be seen in this graph, the two evaluation functions yield different ranges of evaluation values and therefore it is difficult to compare them. This is dealt with by the normalization of the evaluation values, which is depicted in the graph on the right. Both ranges of the two evaluation functions f_1 and f_2 cover exactly the unit interval and it is possible to compare the normalized evaluation values φ_s^f. This is shown with the same two services s_1 and s_2 from the left graph. It reveals that service s_1 is rated differently by f_1 and f_2 while the two evaluation functions nearly agree on the quality of service s_2. The vertical distance of the normalized evaluation values, averaged over all services, is defined to be the inconsistency of two evaluation functions in a certain case study.

One shortcoming of this approach is that the normalized evaluation values depend on the set of all considered services of a case study. As defined in Equations 9 and 10, all deviations in objective values between two services are compared relatively to the largest differences between any services of the respective case study. That means, in case all services are almost identical in quality, different evaluation functions might be indicated as being inconsistent although they hardly show considerable differences in the evaluation. However, when considering services that do show differences in quality, such an incorrect indication of inconsistency cannot occur.

In the three case studies NS, GS and GL we derive the pairwise inconsistencies between all 16 evaluation functions, which are tabulated in Appendix E.

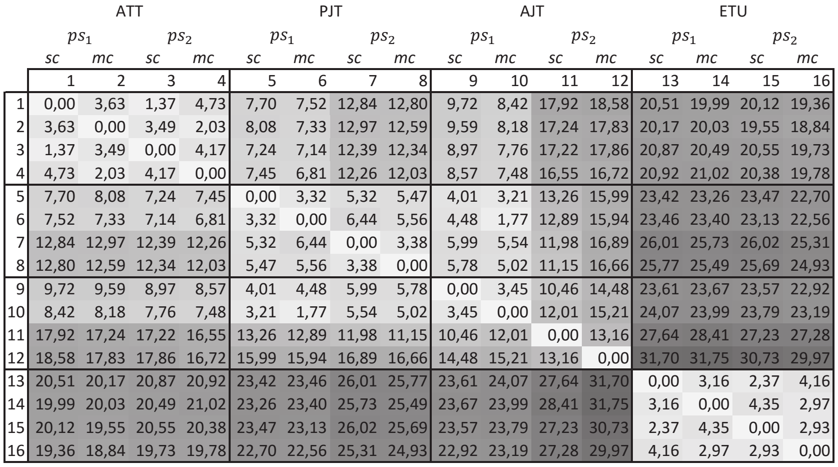
For better comprehensibility, the tables with inconsistencies are colored as heat maps, quadratic 16×16 matrices where each entry displays the inconsistency of two evaluation functions. To make differences in inconsistencies easily identifiable, high values are indicated by a dark shading and low values have a light shading. Naturally, all diagonal values of the matrix are zero as each evaluation function is fully consistent with itself and the matrices are symmetric since $i_\varphi(f_1, f_2) = i_\varphi(f_2, f_1)$ holds.

Altogether, we find qualitatively similar results, that means, the inconsistencies of the studied evaluation functions are qualitatively alike in the different case studies. Only for very few cells of the heat maps we observe a qualitative difference in the pattern of inconsistencies of evaluation functions between the case studies. This indicates that the results are not dependent on the structure of the case study but indeed on the structure of the evaluation functions. Therefore, we discuss the findings independently of the case studies where this is applicable and just highlight differences in the results.

For a collective discussion we compute the weighted average of the inconsistencies between the evaluation functions over all case studies by

$$i(f_1, f_2) = \frac{\sum\limits_{I \in \{GS, GL, NS\}} |S_I| i_\varphi^I(f_1, f_2)}{\sum\limits_{I \in \{GS, GL, NS\}} |S_I|} \quad \forall f_1, f_2 \in F,$$

where $|S_I|$ is the number of services considered in case study I and $i_\varphi^I(f_1, f_2)$ is the inconsistency of evaluation functions f_1 and f_2 derived in case study I. The weighted average inconsistencies are tabulated in a heat map in Fig. 4.



	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0,00	3,63	1,37	4,73	7,70	7,52	12,84	12,80	9,72	8,42	17,92	18,58	20,51	19,99	20,12	19,36
2	3,63	0,00	3,49	2,03	8,08	7,33	12,97	12,59	9,59	8,18	17,24	17,83	20,17	20,03	19,55	18,84
3	1,37	3,49	0,00	4,17	7,24	7,14	12,39	12,34	8,97	7,76	17,22	17,86	20,87	20,49	20,55	19,73
4	4,73	2,03	4,17	0,00	7,45	6,81	12,26	12,03	8,57	7,48	16,55	16,72	20,92	21,02	20,38	19,78
5	7,70	8,08	7,24	7,45	0,00	3,32	5,32	5,47	4,01	3,21	13,26	15,99	23,42	23,26	23,47	22,70
6	7,52	7,33	7,14	6,81	3,32	0,00	6,44	5,56	4,48	1,77	12,89	15,94	23,46	23,40	23,13	22,56
7	12,84	12,97	12,39	12,26	5,32	6,44	0,00	3,38	5,99	5,54	11,98	16,89	26,01	25,73	26,02	25,31
8	12,80	12,59	12,34	12,03	5,47	5,56	3,38	0,00	5,78	5,02	11,15	16,66	25,77	25,49	25,69	24,93
9	9,72	9,59	8,97	8,57	4,01	4,48	5,99	5,78	0,00	3,45	10,46	14,48	23,61	23,67	23,57	22,92
10	8,42	8,18	7,76	7,48	3,21	1,77	5,54	5,02	3,45	0,00	12,01	15,21	24,07	23,99	23,79	23,19
11	17,92	17,24	17,22	16,55	13,26	12,89	11,98	11,15	10,46	12,01	0,00	13,16	27,64	28,41	27,23	27,28
12	18,58	17,83	17,86	16,72	15,99	15,94	16,89	16,66	14,48	15,21	13,16	0,00	31,70	31,75	30,73	29,97
13	20,51	20,17	20,87	20,92	23,42	23,46	26,01	25,77	23,61	24,07	27,64	31,70	0,00	3,16	2,37	4,16
14	19,99	20,03	20,49	21,02	23,26	23,40	25,73	25,49	23,67	23,99	28,41	31,75	3,16	0,00	4,35	2,97
15	20,12	19,55	20,55	20,38	23,47	23,13	26,02	25,69	23,57	23,79	27,23	30,73	2,37	4,35	0,00	2,93
16	19,36	18,84	19,73	19,78	22,70	22,56	25,31	24,93	22,92	23,19	27,28	29,97	4,16	2,97	2,93	0,00

Fig. 4. Heat map showing the weighted average inconsistencies from all three case studies. For better depiction, all values are multiplied with 100.

The absolute values of the inconsistencies i allow an interpretation of the extent to which the evaluation functions agree on the valuation of the services. For example, an inconsistency of 19.36% between evaluation functions 1 and 16 can be found in the top right corner of Fig. 4. This inconsistency implies that the normalized values of all services regarding these two evaluation functions deviate by 19.36% on average. Visualized in Fig. 3, this would mean that the differences $|\varphi_s^1 - \varphi_s^{16}|$ are on average over all services s approximately one fifth of the total range of normalized evaluation values.

The heat map in Fig. 4 shows obvious patterns with dark and bright areas, indicating large and small differences in the inconsistencies between the evaluation functions. To provide a better intuition, we use multidimensional scaling to visualize the inconsistencies in Fig. 5 as distances between the evaluation functions. That means, we depict each evaluation function f as a point $x_f \in \mathbb{R}^2$ on the plane such that the Euclidean distance $d(x_{f_1}, x_{f_2})$ between each two points is representative for the inconsistency $i(f_1, f_2)$ of the corresponding evaluation functions. This is ensured by minimizing the relative deviation of Euclidean distance from the inconsistency, i.e., we solve

$$\min_{x \in \mathbb{R}^{2|F|}} \frac{\sum\limits_{f_1, f_2 \in F} (d(x_{f_1}, x_{f_2}) - i(f_1, f_2))^2}{\sum\limits_{f_1, f_2 \in F} i(f_1, f_2)^2}.$$

More on multidimensional scaling can be found in Borg and Groenen (2005). In general, the representation of inconsistencies as distances in Fig. 5 allows a faster and easier interpretation but all observations can be confirmed with the derived inconsistencies in Fig. 4.

Observations

It is obvious from Fig. 5 that the four utility based evaluation functions are separated from the travel time based evaluation functions. This is independent of the chosen parameter setting or passenger distribution model. Also the heat map indicates by a dark shading in the upper right (or equivalently lower left) part that the evaluation functions based on travel time are generally inconsistent with those based on utility. Furthermore, both figures suggest that the utility based evaluation functions are rather consistent with each other, visible from low distances between pairs of utility based evaluation functions in Fig. 5 and also from light shading in the lower right corner of Fig. 4. The utility based evaluation functions are especially far from the functions of adapted journey time although ETU and AJT are the only two quality measurements which consider the adaption time besides other characteristics, see Table 2. This shows that the shape of an evaluation function is in this case more relevant for the inconsistency than the characteristics it takes into account in the evaluation.

A second group of evaluation functions that are consistent with each other but a bit separate from other groups is formed by the evaluation functions of absolute travel time. By design, this group of evaluation functions is least

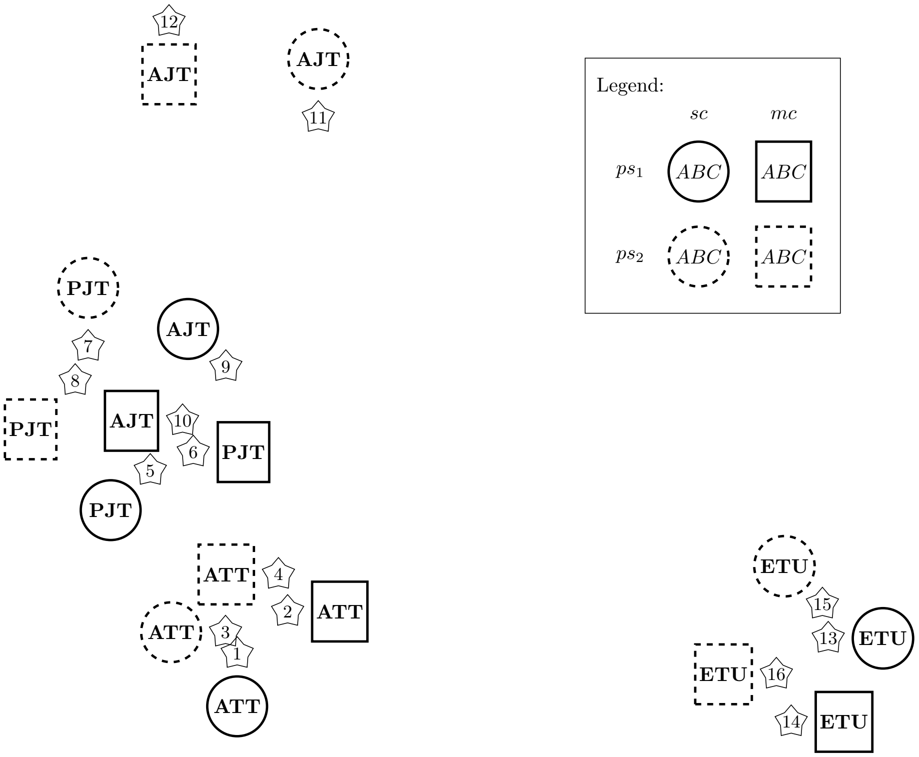


Fig. 5. The inconsistencies of pairs of evaluation functions visualized as distances on the plane. Each star corresponds to one evaluation function displaying its id. The labels next to the stars explain how the evaluation function is constructed. The quality measurement ATT, PJT, AJT or ETU is written in the labels. A round label shape indicates that passengers are distributed on the shortest connections (sc), while squared labels indicate the use of a passenger distribution model on multiple connections (mc). The used parameter setting is distinguishable by solid (ps_1) or dashed label edging (ps_2).

affected by different parameter settings and therefore we did indeed expect that evaluation functions from this group would be relatively consistent with each other. In line with this, a close inspection also shows that in our case studies the passenger distribution model has a higher impact on the inconsistency of evaluation functions of absolute travel time than the parameter setting. The group of evaluation functions of absolute travel time is far from the utility based evaluation functions and closer to other travel time based evaluation functions. The closest group to the evaluation functions of absolute travel time are the four evaluation functions of perceived journey time and the two evaluation functions of adapted journey time with the first parameter setting. Especially with the first parameter setting this closeness is plausible since the first parameter setting is very similar to the fixed parameters of absolute travel time, see Equation 7.

That the two evaluation functions of perceived journey time with the second parameter setting are a little further away indicates that the penalties for transfers and the weighting of transfer wait time have a measurable effect on the inconsistency of the evaluation functions.
In the top left corner of Fig. 5 we find the stars of both evaluation functions based on adapted journey time with the second parameter setting, separate from the other evaluation functions and also relatively far from each other. This is also reflected in the inconsistencies in the heat map in Fig. 4 where both evaluation functions 11 and 12 show fairly high inconsistencies with all other evaluation functions. A plausible explanation for this is the adaption time. The adapted journey time is the only travel time based quality measurement comprising the adaption time, and with the second parameter setting the adaption is penalized much higher than when using the first parameter setting.
A possible reason for the high inconsistency between the two evaluation functions of adapted journey time with the second parameter setting can be found in the set of services in our case studies; One kind of service provides no reasonably good alternative to the best connection(s) whereas the second kind of service additionally offers such alternatives. The evaluation of these two kinds of services is similar when considering the shortest connection since both offer comparable shortest connections. However, the adaption time in the second kind of service, which provides many comparably good connections for each OD pair, is drastically lower when considering multiple connections which leads to a different rating of the two kinds of services. The presence of both kinds of services in the case studies might account for the visible inconsistency between the two outliers for different passenger distribution models.
To summarize, Fig. 5 suggests that there are three groups of evaluation functions that are close to each other, but far from functions of other groups. One group is formed by the four utility based evaluation functions, one by the four evaluation functions of absolute travel time and one by the evaluation functions of perceived journey time and adapted journey time with the first parameter setting. Additionally, the remaining two evaluation functions of adapted journey time with the second parameter setting seem to be two outliers apart from the three groups.

5.2 Cluster analysis

In addition to an investigation of the inconsistencies, we perform cluster analyses of the evaluation functions in each of the three case studies. These help to determine which of the evaluation functions are similar to each other and which are fundamentally different. With the cluster analyses we can, on the one hand, verify the group formation that is apparent in Fig. 5 and, on the other hand, identify individual variations of the inconsistencies in the different case studies. The evaluation functions $f \in F$ are clustered based on the normalized evaluation values φ_s^f of all considered services $s \in S$. For a given $k \in \mathbb{N}$, each evaluation function is assigned to exactly one of k clusters such that the sum of all distances between the evaluation functions and their cluster center is minimal. As distance measure between an evaluation function f and a cluster center m we use the rectilinear distance of the normalized evaluation values φ to the cluster center,

$$d(m,f) = \frac{1}{|S|}\sum_{s\in S}|\varphi_s^f - m_s|. \tag{11}$$

Note, that this distance $d(m,f)$ is consistent with the definition of inconsistency $i_\varphi(f,m)$, in the sense that

$$d(f_1,f_2) = i_\varphi(f_1,f_2).$$

The complete mixed integer program we use to solve the clustering problem is specified in Appendix F. In each case study we cluster the set of 16 evaluation functions F into k clusters, for $k \in \{2,\dots,5\}$. Varying the number of clusters k helps to get a better understanding of the inconsistency of evaluation functions. In Appendix G we provide all four clusterings for each of the three case studies. These 12 clusterings are summarized in Fig. 6, each clustering represented by lines grouping several points. As before, each point corresponds to one evaluation function and for each cluster of evaluation functions there is a line surrounding the corresponding points. The thickness of a line depends on the cumulative frequency of appearance of the cluster. Hence, the number and thickness of the lines separating two evaluation functions visualize how often these two functions were separated into different clusters. Note, that in Fig. 6 the distances between evaluation functions are not representative for the inconsistencies.

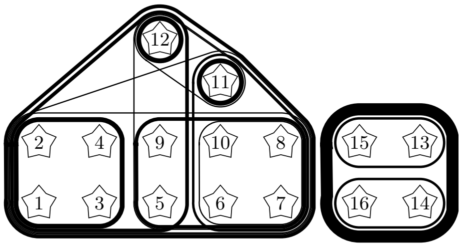


Fig. 6. The accumulated clusterings of all evaluation functions

Observations

In general, the cluster analysis confirms the observations made from a direct interpretation of the inconsistencies in Fig. 5. Additionally, it contributes some kind of ranking of which inconsistencies are more substantial.

It can be seen that the strongest separation is between the utility based evaluation functions and the travel time based evaluation functions. In no case study two evaluation functions from the two different bases were found in the same cluster. This gives evidence that the decision whether to use a travel time based or a utility based evaluation is most crucial in this setting. Also within the group of travel time based evaluation functions, we observe that the visible inconsistencies in Fig. 5 get confirmed by the cluster analysis. For the grouping of evaluation functions it seems to be important whether the absolute travel time or a weighted travel time equivalent is used. In combination with the different passenger distribution models and assumptions on the passenger preferences, this can significantly influence how the evaluation functions are separated into different clusters. This is especially visible when comparing evaluation functions of the adapted journey time in combination with the second parameter setting to other travel time based evaluation functions.

In addition to that, the cluster analysis adds a refinement of the previous observations and reveals coherences that are not or less visible in Fig. 5. For example, the cluster analysis shows that there is a difference between utility based evaluation functions for the different passenger distribution models. Functions of evaluated total utility are always clustered together when they use the same distribution model, but are occasionally separated from each other when using different distribution models. This effect is mainly found in the NS case study and only visible in the cluster analysis since the three case studies are examined individually in contrast to an investigation of averaged values as in Fig. 4. A probable explanation is that the services in this case study offer good alternative connections to the shortest connection for the main demand pairs. This affects the evaluation when considering all reasonable connections or the shortest connection only.

Fig. 6 also shows that neither the parameter setting nor the choice of the distribution model is solely decisive for a clustering of the evaluation functions across the case studies. For some combinations of parameter settings and distribution models, evaluation functions of the different quality measurements are clustered together.

6 Implications

It is interesting to see that there are structural differences in the consistency of timetable evaluation functions. In addition to a mere statement that different evaluation function might not agree on what is a good or a bad timetable, this structure can identify and explain reasons for these differences. The analysis in Section 5 helps to determine which components of the functions are responsible for the found inconsistencies. In this section, we give a brief indication how this can be used for further research dealing with the evaluation of timetables.

Often, the design of evaluation functions is restricted by different reasons, such as unavailable data, imperfect knowledge about passenger behavior or computational complexity. The observations from the inconsistencies and the cluster analysis allow implications on how to deal with these restrictions and which design element to focus on during the design or choice of an evaluation function.

On the one hand, the analysis can help to identify which simplifications of an evaluation function are justifiable. That means, it is possible to determine which simplifications have only a minor effect on the result of the evaluation. A simplification is justified if the desired evaluation function and its simplified version are rather consistent with each other, visible by not being separated into different clusters or by low values of inconsistency. For example, when designing an evaluation function based on absolute travel time without being aware of the precise parameters of the passenger preferences, approximate parameters will not drastically change the evaluation according to our case studies. This holds for both distribution models we tested, obvious from the low inconsistencies between evaluation functions 1 and 3, as well as between evaluation functions 2 and 4. Since the parameter settings for the passenger preferences affect in the case of absolute travel time only the connection choice, the validity of this simplification is expected and the analysis gives confirmation to that. This implies for the case of absolute travel time as quality measurement that the negative impact of a non-reflected modeling of passenger preferences can be disregarded as the resulting error is rather negligible.

On the other hand, this research helps to identify possibilities for improving a currently used evaluation function most effectively. Knowing that the evaluation function in use does not fully depict reality, it can be improved in various ways. The main categories of improvement are the quality measurement including which characteristics are considered, the modeling of passenger preferences and behavior, as well as the connection choice. Since modifying an evaluation function often involves elaborate data acquisition or expensive remodeling, it is desirable to make an estimate of the effects of possible modifications beforehand. For example, assume that a public transport operator applies the adapted journey time on a logit distribution for the evaluation of their services. For modeling passenger preferences they use estimated parameters. In this case, it is highly recommended to identify the correct parameters for modeling preferences and behavior of their customers properly. Using wrong parameters can lead to very different evaluation results as this research identified a high inconsistency between evaluation functions 10 and 12.

As mentioned, simplifying evaluation functions can be useful or necessary for several reasons. However, it is only reasonable if the evaluation results are consistent. It is therefore of utter importance to estimate the impact of a simplification on the evaluation. While this is important for any evaluation application, it is especially relevant at the design of timetables. Using a wrong evaluation function might not only give a wrong indication of what is a good or a bad timetable but can even misdirect the search for good solutions.

7 Conclusion

In this paper, we structured evaluation functions for public transport timetables that are commonly used in the literature and identified three components in which the functions differ from each other. Based on this, we designed a set of evaluation functions representing a wide range of commonly used evaluation functions used in mathematical models, evaluation applications, and choice models.

Furthermore, we introduced a novel methodology to quantify the inconsistency between evaluation functions. This is, unlike existing comparisons, an empirical approach based on the evaluation values of multiple timetables. Therefore, this definition is generally applicable for comparing evaluation functions and is not limited to the set of evaluation functions presented in this paper.

With this methodology, we provided an analysis of the inconsistency of the designed evaluation functions. This analysis was conducted on three sets of timetables for an artificial grid network and the real-world network of Netherlands Railways. The findings are qualitatively similar for both infrastructures even though the networks and the timetables considered are structurally different. This suggests that a generalization of the results is possible.

In our experiments, we found that there are high inconsistencies between different evaluation functions although they are all designed to measure the same - the quality of timetables from passengers' perspective. In all case studies it appears most crucial whether a travel time based or a utility based evaluation is used, which raises the question why utility based evaluation functions are commonly accepted for choice models but hardly used for evaluation. Furthermore, we observed that also within the group of travel time based evaluation functions high inconsistencies can appear. It seems most important which quality measurement is used but also different parameter settings and passenger distributions can significantly impact the inconsistency between evaluation functions. These inconsistencies can be used to validate simplifications of evaluation functions or to identify aspects of an evaluation function that need to be incorporated for a valid evaluation.

This research confirms the impression that even within a set of evaluation functions which are all meant to evaluate the quality of timetables for passengers, the choice of the evaluation function can have a significant impact on the assessed quality of timetables, and thus also on which timetable is considered optimal. This observation is particularly crucial for Operations Research models in public transport as optimizing on the wrong objective function could make the world worse rather than better.

Acknowledgements

This research was supported by Stichting Erasmus Trustfonds. We further thank Niek Guis, Gábor Maróti and Joël van't Wout from Netherlands Railways for providing us with data for the NS case study and for their valuable comments and suggestions throughout this work.

References

Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Ralf Borndörfer, Heide Hoppmann, and Marika Karbstein. Passenger routing for periodic timetable optimization. *Public Transport*, 9(1-2):115–135, 2017.

Mark A Bradley and Hugh F Gunn. Stated preference analysis of values of travel time in the netherlands. *Transportation Research Record*, (1285), 1990.

B De Keizer, K Geurs, and G Haarsman. Interchanges in timetable design of railways: A closer look at customer resistance to interchange between trains. In *Proceedings of the European Transport Conference, Glasgow*, pages 8–10, 2012.

Luigi Dell'Olio, Angel Ibeas, and Patricia Cecín. Modelling user perception of bus transit quality. *Transport Policy*, 17(6):388–397, 2010.

Juan de Dios Ortuzar and Luis G Willumsen. *Modelling transport*. John Wiley & Sons, 2011.

José Luis Espinosa-Aranda, Ricardo García-Ródenas, María Luz López-García, and Eusebio Angulo. Constrained nested logit model: formulation and estimation. *Transportation*, 45(5):1523–1557, 2018.

Bent Flyvbjerg, Kjeld Kahr, Peter Bo Petersen, and Johs Vibe-Petersen. Evaluation of public transport: method for application in open planning. *Transportation*, 13(1):23–52, 1986.

Markus Friedrich, Ingmar Hofsaess, and Steffen Wekeck. Timetable-based transit assignment using branch and bound techniques. *Transportation Research Record: Journal of the Transportation Research Board*, (1752):100–107, 2001.

Michael R Garey and David S Johnson. *Computers and intractability: a guide to NP-completeness*. WH Freeman and Company, San Francisco, 1979.

Philine Gattermann, Peter Großmann, Karl Nachtigall, and Anita Schöbel. Integrating passengers' routes in periodic timetabling: A SAT approach. In *OASIcs-OpenAccess Series in Informatics*, volume 54. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2016.

Alexander Grey. The generalised cost dilemma. *Transportation*, 7(3):261–280, 1978.

Niek Guis and Sandra Nijënstein. Modelleren van klantvoorkeuren in dienstregelingsstudies (in dutch). *Colloquium Vervoersplanologisch Speurwerk. Antwerpen: NS*, 2015.

David A Hensher and Kenneth J Button. *Handbook of transport modelling.* Emerald Group Publishing Limited, 2007.

Gerard de Jong, Andrew Daly, Marits Pieters, and Toon van der Hoorn. The logsum as an evaluation measure: review of the literature and new results. *Transportation Research Part A: Policy and Practice*, 41(9):874–889, 2007.

Satoshi Kanai, Koichi Shiina, Shingo Harada, and Norio Tomii. An optimal delay management algorithm from passengers' viewpoints considering the whole railway network. *Journal of Rail Transport Planning & Management*, 1(1): 25–37, 2011.

Takahiko Kusakabe, Takamasa Iryo, and Yasuo Asakura. Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation*, 37(5):731–749, 2010.

Christian Liebchen. On the benefit of preprocessing and heuristics for periodic timetabling. In *Operations Research Proceedings 2017*, pages 709–714. Springer, 2018.

Karl Nachtigall. Periodic network optimization and fixed interval timetables. *Deutsches Zentrum für Luft–und Raumfahrt, Institut für Flugführung, Braunschweig*, 1998.

Jens Parbo, Otto Anker Nielsen, and Carlo Giacomo Prato. User perspectives in public transport timetable optimisation. *Transportation Research Part C: Emerging Technologies*, 48:269–284, 2014.

Jens Parbo, Otto Anker Nielsen, and Carlo Giacomo Prato. Passenger perspectives in railway timetabling: a literature review. *Transport Reviews*, 36(4): 500–526, 2016.

PTV Planung Transport Verkehr AG. Visum user manual, 2017.

Tomáš Robenek, Yousef Maknoon, Shadi Sharif Azadeh, Jianghang Chen, and Michel Bierlaire. Passenger centric train timetabling problem. *Transportation Research Part B: Methodological*, 89:107–126, 2016.

Bernd Hermann Schittenhelm. *Quantitative Methods for Assessment of Railway Timetables*. PhD thesis, Technical University of Denmark, Transport, 2013.

Marie Schmidt and Anita Schöbel. Timetabling with passenger routing. *OR spectrum*, 37(1):75–97, 2015.

Peter Sels, Thijs Dewilde, Dirk Cattrysse, and Pieter Vansteenwegen. Expected passenger travel time for train schedule evaluation and optimization. In *Proceedings of the 5th international seminar on railway operations modelling and analysis, Copenhagen, Denmark Google Scholar*, 2013.

Paolo Serafini and Walter Ukovich. A mathematical model for periodic scheduling problems. *SIAM Journal on Discrete Mathematics*, 2(4):550–581, 1989.

Joffre Swait and Jordan Louviere. The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of marketing research*, pages 305–314, 1993.

Mark Wardman. Public transport values of time. *Transport policy*, 11(4):363–377, 2004.

Mark Wardman and Jeremy Toner. Is generalised cost justified in travel demand analysis? *Transportation*, pages 1–34, 2018.

Mark Wardman, Phani Chintakayala, Gerard de Jong, and Diego Ferrer. European wide meta-analysis of values of travel time. *ITS, University of Leeds, Paper prepared for EIB*, 2012.

Liu Yang and William Lam. Probit-type reliability-based transit network assignment. *Transportation Research Record: Journal of the Transportation Research Board*, (1977):154–163, 2006.

Appendices

A Notation

Greek letters

α	Scaling parameter for passenger preferences
β	Scaling parameter for logit model
γ	Scaling parameter for departure time tolerance
δ	Filter coefficient for ATT and PJT
ε	Filter parameter for ATT, PJT and NTR
φ	Normalized value of a service w.r.t. an evaluation function

Latin capitals

ADT	Adaption time
AJT	Adapted journey time
ATT	Absolute travel time
C	Set of connections
DEP	Departure time
ETU	Evaluated total utility
F	Set of evaluation functions
GL	Case study on grid infrastructure
GS	Case study on grid infrastructure
I	Index for case studies
IVT	In-vehicle time
NS	Case study on infrastructure of Netherlands Railways
NTR	Number of transfers
OD	Set of OD pairs
PJT	Perceived journey time
S	Set of public transport services
T	Analysis period
TWT	Transfer wait time
WKT	Walk time

Latin lower case letters

c	Index for connection
f	Index for evaluation function
i	Inconsistency
k	Number of clusters
mc	Distribution model on multiple connections
o	Passenger load
od	Index for OD pair
p	Probability for connection choice
ps	Parameter setting
s	Index for public transport service
sc	Distribution model on shortest connection
t	Index for time slice
v	Value of a service w.r.t. an evaluation function

B Public transport service names for grid instance

The public transport services for the grid infrastructure are available on a repository, see Section 4.1. We provide a table with all names of the services as used in this paper and their corresponding name as used in the repository.

name paper	name repository
GL1	Solution_NDP_S001
GL2	Solution_NDP_S002
GL3	Solution_NDP_S003
GL4	Solution_NDP_S018
GL5	Solution_NDP_S019
GL6	Solution_NDP_S020
GL7	Solution_NDP_S021
GL8	Solution_NDP_S022
GL9	Solution_NDP_S024
GL10	Solution_NDP_S025
GL11	Solution_NDP_S026
GL12	Solution_NDP_S027
GL13	Solution_NDP_S028
GL14	Solution_NDP_S029
GL15	Solution_NDP_S030
GL16	Solution_NDP_S031
GL17	Solution_NDP_S032
GL18	Solution_NDP_S035
GL19	Solution_NDP_S036
GL20	Solution_NDP_S037
GL21	Solution_NDP_S038
GL22	Solution_NDP_S039
GL23	Solution_NDP_S040
GL24	Solution_NDP_S041
GL25	Solution_NDP_S042
GL26	Solution_NDP_S043
GL27	Solution_NDP_S044

name paper	name repository
GS1	Solution_NDP_S004
GS2	Solution_NDP_S005
GS3	Solution_NDP_S006
GS4	Solution_NDP_S007
GS5	Solution_NDP_S008
GS6	Solution_NDP_S009
GS7	Solution_NDP_S010
GS8	Solution_NDP_S011
GS9	Solution_NDP_S012
GS10	Solution_NDP_S013
GS11	Solution_NDP_S014
GS12	Solution_NDP_S015
GS13	Solution_NDP_S016
GS14	Solution_NDP_S017
GS15	Solution_NDP_S023
GS16	Solution_NDP_S033
GS17	Solution_NDP_S034
GS18	Solution_NDP_S045
GS19	Solution_NDP_S046
GS20	Solution_NDP_S048
GS21	Solution_NDP_S049
GS22	Solution_NDP_S050
GS23	Solution_NDP_S051
GS24	Solution_NDP_S052
GS25	Solution_NDP_S053
GS26	Solution_NDP_S054
GS27	Solution_NDP_S055
GS28	Solution_NDP_S056

Table 7: Service names as used in this paper and corresponding names in the repository

C Medal counts

The tables 8 and 9 show how often each public transport service in the case studies GS and GL was ranked on the n^{th} rank. Similar to the medal count for the case study NS in Table 6, at a first glance it appears to be possible to decide on a ranking of all services in the case study GL in Table 8. However, since many services span over a large range of ranks, the evaluation with a single evaluation function is likely to show a different result. In the GS case study, it is less easy to derive a ranking between most services by visual inspection of Table 9.

It is apparent in both case studies, especially in GS, that the evaluation functions are more or totally consistent when ranking the best and worst public transport services. This can be explained by the choice of services in the set S. All services are designed to meet the demand, however, with respect to different

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
GL24	16																										
GL23		14	2																								
GL1				4	1	4	6	1																			
GL22			11				1									4											
GL5			1		1			2		2										1							
GL7						4	4	4	2	2																	
GL21				11												1		4									
GL2						5		3	3	4	1																
GL4					1	3		4	5	2								1									
GL13								2	5	1	3			4	1												
GL15		2	2	1								2		4	1	3			1								
GL12										4	7	4		1													
GL10					4		1							4	4	2	1										
GL11											1	9	4		2												
GL14											4		9	2		1											
GL9							4		1							4			4	1	2						
GL6													2		3	1	6	2	1			1					
GL17													1	1	4		4	2	4								
GL3												1			1		4		2	8							
GL16																	1	7	4	4							
GL8										1										2	9	4					
GL19																					5		1	4	6		
GL27																						1			2	4	
GL26																							1	2	2	2	
GL18																						1	5	6	4		
GL20																								4	2	1	
GL25																											16

Table 8: 'Medal count' from GL case study showing the number of times a public transport service is ranked on the n^{th} rank. Zeros are omitted for better visibility.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
GS16	10	5				1																						
GS15	5	5	2	4																								
GS25	1	2	13																									
GS28		4	1				2	4	1	2		1								1								
GS9				2	4	1			3	1				1	1	1			2									
GS8				2	2	2	2	1	1		2									3	1							
GS1				2		4				1	3	1	1		1	1	2											
GS27							1	1	5	2	4	1	1				1											
GS26					2	2	4	2								1	1		3		1							
GS2					2			1			3	4	2	3	1													
GS17				4		3	1	2						1							1	1	1		1	1		
GS12				2			2	2	1	2		1				1		2						1	1	1		
GS3										3		2	4	1	3	1	1	1										
GS22					2	2			1	1	1		1	1		1		2		1	1		1			1		
GS18							2		1	1			2	2	1	3		3	1									
GS20										1	1	5	1	1	2	3						1	1					
GS14						1			1	1	1	1	1	4			2								1		2	1
GS7													2		4	1	4	5										
GS10							2						1	1	1	1	1	1	2	1	5							
GS11							2						1	1	1	1	1	1	2	1	5							
GS5					4				1												1		2	3		3	2	
GS6																1	2		5		1	3	1	2				1
GS4								1	1	1	1							1					2	1	1	4	3	
GS13															1		1			6	2	1	1			4		
GS24																			1	2	2	4	4	1	2			
GS21																					1		1	5	4	2	1	2
GS23																								2	6		5	3
GS19																						1	2	1			3	9

Table 9: 'Medal count' from GS case study showing the number of times a public transport service is ranked on the n^{th} rank. Zeros are omitted for better visibility.

objectives. While most services are designed with the aim to maximize quality for the passengers in one way or another while respecting a reasonable budgetary restrictions, some services aim to be as cheap as possible, usually at the expense of quality for passengers, and others had no cap on costs which would not be feasible in practice. This would eliminate at least the best services from the list of alternatives, making it hard to decide on a ranking of the remaining services.

D Values

The tables in this section show the normalized evaluation values of all public transport services with respect to all 16 evaluation functions from each of the three case studies.

	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
GL1	0,38	0,35	0,38	0,34	0,44	0,40	0,49	0,45	0,41	0,41	0,34	0,21	0,44	0,46	0,39	0,39
GL2	0,40	0,37	0,40	0,37	0,46	0,42	0,50	0,46	0,43	0,43	0,34	0,22	0,46	0,47	0,40	0,40
GL3	0,63	0,61	0,63	0,61	0,67	0,64	0,63	0,60	0,64	0,65	0,50	0,29	0,69	0,69	0,72	0,72
GL4	0,39	0,37	0,39	0,37	0,43	0,41	0,45	0,42	0,44	0,42	0,62	0,21	0,52	0,53	0,51	0,51
GL5	0,34	0,33	0,34	0,33	0,39	0,37	0,40	0,38	0,41	0,38	0,68	0,18	0,50	0,51	0,51	0,51
GL6	0,51	0,47	0,51	0,47	0,54	0,50	0,63	0,60	0,54	0,51	0,73	0,34	0,64	0,65	0,60	0,60
GL7	0,38	0,35	0,37	0,34	0,44	0,40	0,50	0,47	0,42	0,41	0,45	0,23	0,55	0,56	0,50	0,50
GL8	0,64	0,63	0,65	0,62	0,69	0,66	0,69	0,66	0,65	0,67	0,49	0,40	0,84	0,85	0,88	0,88
GL9	0,50	0,47	0,50	0,47	0,61	0,58	0,71	0,68	0,58	0,58	0,46	0,36	0,48	0,49	0,42	0,42
GL10	0,47	0,44	0,47	0,44	0,53	0,49	0,61	0,57	0,51	0,50	0,45	0,30	0,45	0,46	0,39	0,39
GL11	0,41	0,38	0,41	0,38	0,47	0,44	0,55	0,51	0,46	0,45	0,49	0,26	0,64	0,65	0,60	0,61
GL12	0,41	0,38	0,40	0,37	0,46	0,43	0,52	0,49	0,45	0,44	0,51	0,25	0,62	0,63	0,60	0,60
GL13	0,40	0,37	0,40	0,37	0,45	0,42	0,53	0,50	0,45	0,43	0,54	0,26	0,64	0,65	0,60	0,60
GL14	0,44	0,41	0,44	0,41	0,49	0,45	0,55	0,52	0,47	0,46	0,55	0,27	0,60	0,61	0,59	0,59
GL15	0,45	0,44	0,46	0,45	0,53	0,50	0,54	0,51	0,53	0,51	0,63	0,19	0,37	0,38	0,35	0,35
GL16	0,54	0,52	0,54	0,52	0,58	0,55	0,63	0,60	0,56	0,54	0,56	0,32	0,79	0,79	0,79	0,80
GL17	0,50	0,48	0,50	0,48	0,54	0,51	0,59	0,55	0,52	0,51	0,51	0,29	0,77	0,78	0,79	0,79
GL18	0,82	0,80	0,81	0,80	0,86	0,84	0,88	0,86	0,83	0,85	0,78	0,67	0,93	0,93	0,94	0,94
GL19	0,81	0,80	0,81	0,80	0,87	0,86	0,94	0,92	0,83	0,86	0,71	0,71	0,83	0,83	0,82	0,82
GL20	0,98	0,98	0,98	0,97	0,98	0,98	0,98	0,96	0,94	0,97	0,79	0,78	0,94	0,94	0,95	0,95
GL21	0,26	0,24	0,26	0,24	0,29	0,27	0,32	0,29	0,28	0,28	0,29	0,30	0,76	0,75	0,76	0,76
GL22	0,22	0,21	0,22	0,21	0,24	0,23	0,26	0,25	0,23	0,24	0,21	0,21	0,66	0,65	0,67	0,67
GL23	0,11	0,10	0,11	0,10	0,14	0,12	0,16	0,14	0,12	0,13	0,08	0,09	0,37	0,37	0,36	0,36
GL24	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
GL25	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
GL26	0,74	0,77	0,74	0,77	0,78	0,80	0,81	0,85	0,76	0,78	0,80	0,84	0,98	0,98	0,94	0,94
GL27	0,66	0,68	0,66	0,69	0,70	0,71	0,72	0,75	0,74	0,70	0,90	0,87	0,97	0,97	0,97	0,97

Table 10: Normalized evaluation values φ in GL case study

	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
GS1	0,67	0,64	0,64	0,58	0,53	0,52	0,50	0,48	0,45	0,48	0,24	0,29	0,70	0,73	0,68	0,68
GS2	0,61	0,64	0,58	0,60	0,50	0,52	0,46	0,49	0,43	0,49	0,29	0,33	0,84	0,84	0,83	0,80
GS3	0,62	0,64	0,59	0,61	0,51	0,52	0,47	0,50	0,45	0,49	0,36	0,36	0,88	0,87	0,87	0,83
GS4	0,93	0,95	0,89	0,88	0,96	1,00	0,95	0,98	0,81	0,91	0,44	0,61	0,79	0,82	0,78	0,79
GS5	0,93	0,88	0,89	0,81	0,96	0,95	0,97	0,96	0,79	0,87	0,36	0,57	0,61	0,65	0,60	0,61
GS6	0,71	0,84	0,69	0,88	0,62	0,72	0,54	0,61	0,57	0,66	0,43	0,45	1,00	0,96	0,97	0,91
GS7	0,66	0,73	0,63	0,76	0,55	0,62	0,48	0,54	0,50	0,58	0,42	0,43	0,88	0,91	0,86	0,87
GS8	0,48	0,52	0,46	0,50	0,38	0,43	0,33	0,37	0,35	0,40	0,38	0,38	0,92	0,95	0,89	0,89
GS9	0,50	0,52	0,48	0,51	0,39	0,41	0,32	0,35	0,36	0,39	0,30	0,44	0,88	0,92	0,85	0,86
GS10	0,91	0,87	0,87	0,80	0,68	0,67	0,56	0,55	0,57	0,62	0,31	0,34	0,84	0,87	0,86	0,86
GS11	0,91	0,87	0,87	0,80	0,68	0,67	0,56	0,55	0,57	0,62	0,31	0,34	0,84	0,87	0,86	0,86
GS12	0,32	0,53	0,30	0,47	0,48	0,62	0,52	0,66	0,81	0,57	0,71	0,34	0,83	0,79	0,77	0,78
GS13	0,80	0,85	0,81	0,87	0,64	0,70	0,51	0,57	0,57	0,66	0,43	0,54	0,98	0,98	0,98	0,97
GS14	0,65	0,71	0,63	0,71	0,49	0,54	0,39	0,45	0,44	0,50	0,39	0,42	1,00	1,00	0,99	0,96
GS15	0,00	0,02	0,00	0,02	0,00	0,03	0,00	0,04	0,00	0,04	0,10	0,29	0,54	0,55	0,53	0,54
GS16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,34	0,00	0,00	0,00	0,00
GS17	0,48	0,47	0,46	0,45	0,39	0,39	0,39	0,40	0,36	0,37	0,42	0,67	0,93	0,98	0,92	0,93
GS18	0,70	0,67	0,67	0,62	0,55	0,55	0,51	0,49	0,48	0,51	0,45	0,40	0,79	0,79	0,80	0,77
GS19	1,00	0,96	0,95	0,89	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,92	0,97	0,94	0,96
GS20	0,66	0,63	0,63	0,59	0,53	0,52	0,50	0,48	0,56	0,54	0,65	0,56	0,82	0,83	0,83	0,81
GS21	0,97	0,94	0,93	0,88	0,71	0,72	0,58	0,58	0,77	0,74	0,69	0,64	0,97	1,00	1,00	1,00
GS22	0,38	0,51	0,36	0,47	0,51	0,61	0,55	0,65	0,61	0,59	0,71	0,51	0,82	0,81	0,81	0,82
GS23	0,98	1,00	1,00	1,00	0,77	0,82	0,64	0,67	0,82	0,83	0,78	0,76	0,94	0,97	0,97	0,98
GS24	0,94	0,91	0,90	0,84	0,70	0,69	0,58	0,57	0,67	0,67	0,56	0,52	0,91	0,95	0,93	0,94
GS25	0,01	0,06	0,01	0,03	0,02	0,09	0,02	0,05	0,04	0,10	0,10	0,00	0,19	0,13	0,13	0,13
GS26	0,48	0,48	0,46	0,45	0,43	0,45	0,41	0,43	0,43	0,43	0,50	0,52	0,87	0,90	0,87	0,88
GS27	0,59	0,56	0,57	0,51	0,48	0,47	0,45	0,44	0,46	0,46	0,44	0,38	0,78	0,82	0,78	0,79
GS28	0,56	0,52	0,53	0,48	0,47	0,45	0,44	0,42	0,45	0,45	0,47	0,28	0,15	0,17	0,12	0,12

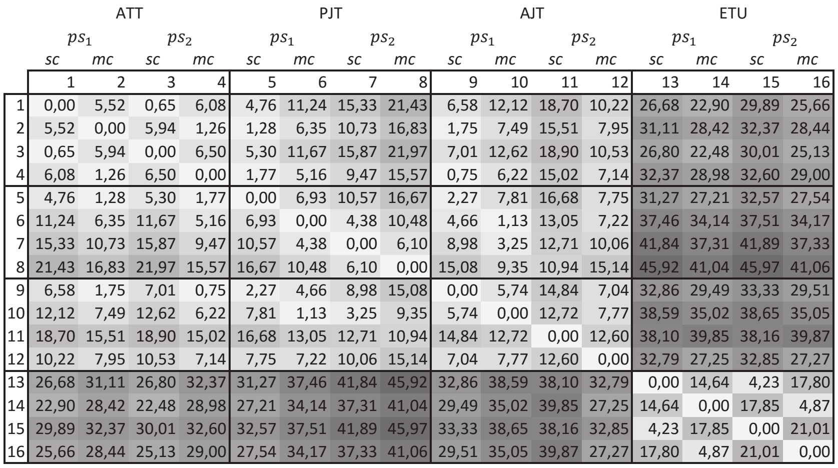
Table 11: Normalized evaluation values φ in GS case study

	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
NS12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,34	0,57	0,20	0,51
NS13	0,43	0,35	0,44	0,35	0,35	0,23	0,14	0,10	0,34	0,22	0,49	0,57	1,00	1,00	1,00	1,00
NS14	0,43	0,31	0,43	0,29	0,35	0,22	0,19	0,13	0,29	0,20	0,02	0,31	0,58	0,48	0,56	0,34
NS15	0,34	0,27	0,35	0,26	0,26	0,17	0,09	0,00	0,23	0,15	0,02	0,17	0,60	0,52	0,60	0,49
NS16	0,35	0,28	0,38	0,24	0,28	0,16	0,10	0,01	0,25	0,14	0,03	0,15	0,35	0,58	0,24	0,66
NS17	0,12	0,08	0,11	0,11	0,13	0,11	0,15	0,23	0,11	0,11	0,00	0,18	0,00	0,27	0,00	0,29
NS18	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,00	0,00	0,07	0,00	0,12	0,00	0,10	0,00

Table 12: Normalized evaluation values φ in NS case study

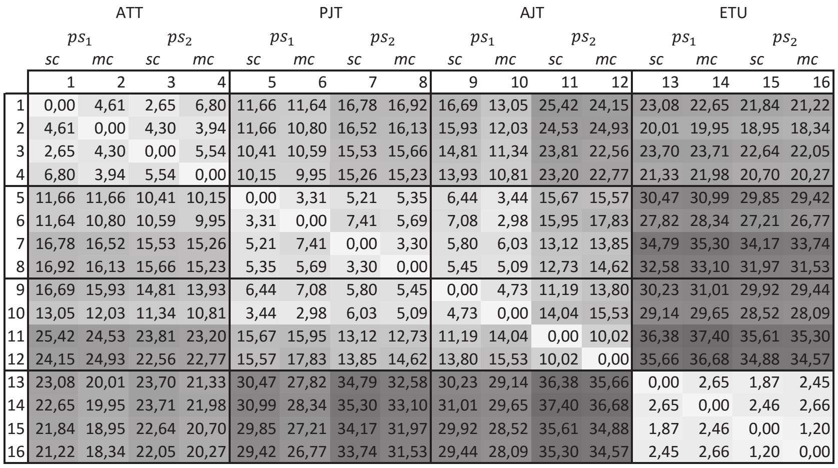
E Heat maps

The tables in this section show the pairwise inconsistencies of two evaluation functions. For better comprehensibility, low values are indicated by a light shading and high values by a dark shading. All values are multiplied with 100 for better depiction.



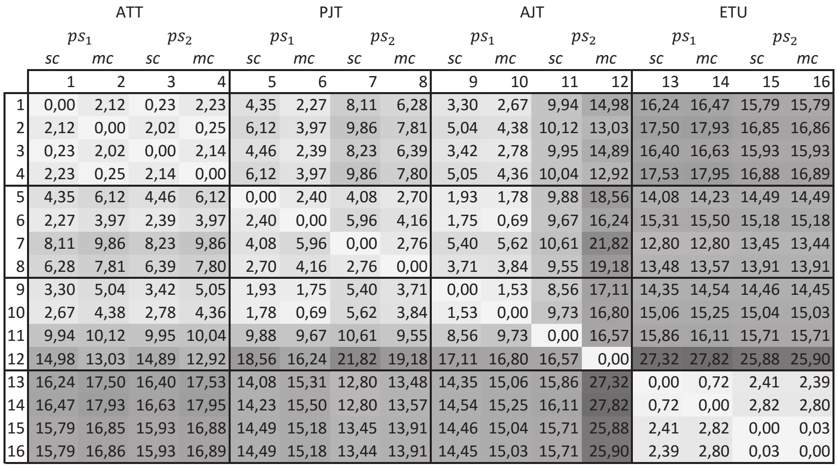
	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0,00	5,52	0,65	6,08	4,76	11,24	15,33	21,43	6,58	12,12	18,70	10,22	26,68	22,90	29,89	25,66
2	5,52	0,00	5,94	1,26	1,28	6,35	10,73	16,83	1,75	7,49	15,51	7,95	31,11	28,42	32,37	28,44
3	0,65	5,94	0,00	6,50	5,30	11,67	15,87	21,97	7,01	12,62	18,90	10,53	26,80	22,48	30,01	25,13
4	6,08	1,26	6,50	0,00	1,77	5,16	9,47	15,57	0,75	6,22	15,02	7,14	32,37	28,98	32,60	29,00
5	4,76	1,28	5,30	1,77	0,00	6,93	10,57	16,67	2,27	7,81	16,68	7,75	31,27	27,21	32,57	27,54
6	11,24	6,35	11,67	5,16	6,93	0,00	4,38	10,48	4,66	1,13	13,05	7,22	37,46	34,14	37,51	34,17
7	15,33	10,73	15,87	9,47	10,57	4,38	0,00	6,10	8,98	3,25	12,71	10,06	41,84	37,31	41,89	37,33
8	21,43	16,83	21,97	15,57	16,67	10,48	6,10	0,00	15,08	9,35	10,94	15,14	45,92	41,04	45,97	41,06
9	6,58	1,75	7,01	0,75	2,27	4,66	8,98	15,08	0,00	5,74	14,84	7,04	32,86	29,49	33,33	29,51
10	12,12	7,49	12,62	6,22	7,81	1,13	3,25	9,35	5,74	0,00	12,72	7,77	38,59	35,02	38,65	35,05
11	18,70	15,51	18,90	15,02	16,68	13,05	12,71	10,94	14,84	12,72	0,00	12,60	38,10	39,85	38,16	39,87
12	10,22	7,95	10,53	7,14	7,75	7,22	10,06	15,14	7,04	7,77	12,60	0,00	32,79	27,25	32,85	27,27
13	26,68	31,11	26,80	32,37	31,27	37,46	41,84	45,92	32,86	38,59	38,10	32,79	0,00	14,64	4,23	17,80
14	22,90	28,42	22,48	28,98	27,21	34,14	37,31	41,04	29,49	35,02	39,85	27,25	14,64	0,00	17,85	4,87
15	29,89	32,37	30,01	32,60	32,57	37,51	41,89	45,97	33,33	38,65	38,16	32,85	4,23	17,85	0,00	21,01
16	25,66	28,44	25,13	29,00	27,54	34,17	37,33	41,06	29,51	35,05	39,87	27,27	17,80	4,87	21,01	0,00

Fig. 7. Heat map showing inconsistencies in the normalized value $i_\varphi(f_1, f_2)$ in NS case study



	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0,00	4,61	2,65	6,80	11,66	11,64	16,78	16,92	16,69	13,05	25,42	24,15	23,08	22,65	21,84	21,22
2	4,61	0,00	4,30	3,94	11,66	10,80	16,52	16,13	15,93	12,03	24,53	24,93	20,01	19,95	18,95	18,34
3	2,65	4,30	0,00	5,54	10,41	10,59	15,53	15,66	14,81	11,34	23,81	22,56	23,70	23,71	22,64	22,05
4	6,80	3,94	5,54	0,00	10,15	9,95	15,26	15,23	13,93	10,81	23,20	22,77	21,33	21,98	20,70	20,27
5	11,66	11,66	10,41	10,15	0,00	3,31	5,21	5,35	6,44	3,44	15,67	15,57	30,47	30,99	29,85	29,42
6	11,64	10,80	10,59	9,95	3,31	0,00	7,41	5,69	7,08	2,98	15,95	17,83	27,82	28,34	27,21	26,77
7	16,78	16,52	15,53	15,26	5,21	7,41	0,00	3,30	5,80	6,03	13,12	13,85	34,79	35,30	34,17	33,74
8	16,92	16,13	15,66	15,23	5,35	5,69	3,30	0,00	5,45	5,09	12,73	14,62	32,58	33,10	31,97	31,53
9	16,69	15,93	14,81	13,93	6,44	7,08	5,80	5,45	0,00	4,73	11,19	13,80	30,23	31,01	29,92	29,44
10	13,05	12,03	11,34	10,81	3,44	2,98	6,03	5,09	4,73	0,00	14,04	15,53	29,14	29,65	28,52	28,09
11	25,42	24,53	23,81	23,20	15,67	15,95	13,12	12,73	11,19	14,04	0,00	10,02	36,38	37,40	35,61	35,30
12	24,15	24,93	22,56	22,77	15,57	17,83	13,85	14,62	13,80	15,53	10,02	0,00	35,66	36,68	34,88	34,57
13	23,08	20,01	23,70	21,33	30,47	27,82	34,79	32,58	30,23	29,14	36,38	35,66	0,00	2,65	1,87	2,45
14	22,65	19,95	23,71	21,98	30,99	28,34	35,30	33,10	31,01	29,65	37,40	36,68	2,65	0,00	2,46	2,66
15	21,84	18,95	22,64	20,70	29,85	27,21	34,17	31,97	29,92	28,52	35,61	34,88	1,87	2,46	0,00	1,20
16	21,22	18,34	22,05	20,27	29,42	26,77	33,74	31,53	29,44	28,09	35,30	34,57	2,45	2,66	1,20	0,00

Fig. 8. Heat map showing inconsistencies in the normalized value $i_\varphi(f_1, f_2)$ GS case study



	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*	*sc*	*mc*
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0,00	2,12	0,23	2,23	4,35	2,27	8,11	6,28	3,30	2,67	9,94	14,98	16,24	16,47	15,79	15,79
2	2,12	0,00	2,02	0,25	6,12	3,97	9,86	7,81	5,04	4,38	10,12	13,03	17,50	17,93	16,85	16,86
3	0,23	2,02	0,00	2,14	4,46	2,39	8,23	6,39	3,42	2,78	9,95	14,89	16,40	16,63	15,93	15,93
4	2,23	0,25	2,14	0,00	6,12	3,97	9,86	7,80	5,05	4,36	10,04	12,92	17,53	17,95	16,88	16,89
5	4,35	6,12	4,46	6,12	0,00	2,40	4,08	2,70	1,93	1,78	9,88	18,56	14,08	14,23	14,49	14,49
6	2,27	3,97	2,39	3,97	2,40	0,00	5,96	4,16	1,75	0,69	9,67	16,24	15,31	15,50	15,18	15,18
7	8,11	9,86	8,23	9,86	4,08	5,96	0,00	2,76	5,40	5,62	10,61	21,82	12,80	12,80	13,45	13,44
8	6,28	7,81	6,39	7,80	2,70	4,16	2,76	0,00	3,71	3,84	9,55	19,18	13,48	13,57	13,91	13,91
9	3,30	5,04	3,42	5,05	1,93	1,75	5,40	3,71	0,00	1,53	8,56	17,11	14,35	14,54	14,46	14,45
10	2,67	4,38	2,78	4,36	1,78	0,69	5,62	3,84	1,53	0,00	9,73	16,80	15,06	15,25	15,04	15,03
11	9,94	10,12	9,95	10,04	9,88	9,67	10,61	9,55	8,56	9,73	0,00	16,57	15,86	16,11	15,71	15,71
12	14,98	13,03	14,89	12,92	18,56	16,24	21,82	19,18	17,11	16,80	16,57	0,00	27,32	27,82	25,88	25,90
13	16,24	17,50	16,40	17,53	14,08	15,31	12,80	13,48	14,35	15,06	15,86	27,32	0,00	0,72	2,41	2,39
14	16,47	17,93	16,63	17,95	14,23	15,50	12,80	13,57	14,54	15,25	16,11	27,82	0,72	0,00	2,82	2,80
15	15,79	16,85	15,93	16,88	14,49	15,18	13,45	13,91	14,46	15,04	15,71	25,88	2,41	2,82	0,00	0,03
16	15,79	16,86	15,93	16,89	14,49	15,18	13,44	13,91	14,45	15,03	15,71	25,90	2,39	2,80	0,03	0,00

Fig. 9. Heat map showing inconsistencies in the normalized value $i_\varphi(f_1, f_2)$ in GL case study

F Mixed integer program for clustering problem

Let φ_s^f be the normalized evaluation values of a public transport service $s \in S$ with respect to evaluation function $f \in F$. Then, an optimal clustering of the set of evaluation functions F into k clusters can be found by solving the program

$$\begin{array}{llll}
\min & & \sum_{f \in F} d(f) & \\
\text{s.t.} \quad \sum_{j=1}^{k} b_{j,f} & = & 1 & \forall f \in F \\
d(f) & = & \sum_{j=1}^{k} d(m_j, f) \cdot b_{j,f} & \forall f \in F \\
d(m_j, f) & = & \frac{1}{|S|} \sum_{s \in S} |\varphi_s^f - m_{j,s}| & \forall f \in F,\ \forall j = 1, \dots, k \\
m_{j,s} & = & \frac{1}{\sum_{f \in F} b_{j,f}} \sum_{f \in F} \varphi_s^f \cdot b_{j,f} & \forall s \in S,\ \forall j = 1, \dots, k \\
m_{j,s} & \in & \mathbb{R} & \forall s \in S,\ \forall j = 1, \dots, k \\
b_{j,f} & \in & \{0, 1\} & \forall f \in F,\ \forall j = 1, \dots, k \\
d(m_j, f) & \in & \mathbb{R} & \forall f \in F,\ \forall j = 1, \dots, k \\
d(f) & \in & \mathbb{R} & \forall f \in F
\end{array}$$

The binary variable $b_{j,f}$ links the evaluation functions f to the clusters j and the first constraint ensures that each function is assigned to exactly one cluster. The second constraint assigns the distance of each evaluation function f to its cluster center m_j to the variable $d(f)$. The distance between the functions f and the cluster centers m_j are computed in the third constraint using the distance function $d(m, f)$ as defined in Equation 11. With the fourth constraint the cluster centers are computed as arithmetic mean of all evaluation functions that are assigned to the cluster. The objective is to minimize the total distance of all evaluation functions to their respective cluster center. We solve a linearized version of this clustering problem.

G Clusterings

The tables in this section show the clusterings of the 16 evaluation functions in each of the three case studies. The clusterings were found with the mixed integer program described in Appendix F. In the first column of each table is stated how many clusters are used. An asterisk indicates that the clustering is not proven to be optimal. The remaining columns contain the clusterings. The clusterings are separated by horizontal lines and in each row one cluster is represented by the ids of the evaluation functions contained in the cluster.

	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	1	2	3	4	5	6	7	8	9	10	11	12				
													13	14	15	16
3	1	2	3	4	5				9			12				
						6	7	8		10	11					
													13	14	15	16
4	1	2	3	4	5				9			12				
						6	7	8		10	11					
													13		15	
														14		16
5	1	2	3	4	5				9			12				
						6	7	8		10						
											11					
													13		15	
														14		16

Table 13: Optimal clustering of the set of evaluation functions F into k clusters in the NS case study

	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	1	2	3	4	5	6	7	8	9	10	11	12				
													13	14	15	16
3	1	2	3	4												
					5	6	7	8	9	10	11	12				
													13	14	15	16
4	1	2	3	4												
					5	6	7	8	9	10						
											11	12				
													13	14	15	16
5*	1	2	3	4												
					5	6	7	8	9	10						
											11					
												12				
													13	14	15	16

Table 14: Optimal clustering of the set of evaluation functions F into k clusters in the GS case study. The asterisk indicates that the clustering is not proven to be optimal.

	ATT				PJT				AJT				ETU			
	ps_1		ps_2		ps_1		ps_2		ps_1		ps_2		ps_1		ps_2	
	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc	sc	mc
k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	1	2	3	4	5	6	7	8	9	10	11	12				
													13	14	15	16
3	1	2	3	4	5	6	7	8	9	10	11					
												12				
													13	14	15	16
4	1	2	3	4	5	6	7	8	9	10						
											11					
												12				
													13	14	15	16
5	1	2	3	4												
					5	6	7	8	9	10						
											11					
												12				
													13	14	15	16

Table 15: Optimal clustering of the set of evaluation functions F into k clusters in the GL case study