Forschungsbericht 2026-22

Quality of Forecasts for Signal Timings at Traffic Lights

Maik Halbach, Robert Oertel,
Robert Markowski, Robert Alms

Deutsches Zentrum für Luft- und Raumfahrt
Institut für Verkehrssystemtechnik
Braunschweig/Berlin

Deutsches Zentrum
DLR für Luft- und Raumfahrt

Forschungsbericht 2026-22

Quality of Forecasts for Signal Timings at Traffic Lights

Maik Halbach, Robert Oertel,
Robert Markowski, Robert Alms

Deutsches Zentrum für Luft- und Raumfahrt
Institut für Verkehrssystemtechnik
Braunschweig/Berlin

64 Seiten
29 Bilder
1 Tabellen
11 Literaturstellen

Deutsches Zentrum
DLR für Luft- und Raumfahrt

Herausgeber:

Deutsches Zentrum
für Luft- und Raumfahrt e. V.
Wissenschaftliche Information
Linder Höhe
D-51147 Köln

ISSN 1434-8454
ISRN DLR-FB-2026-22
Erscheinungsjahr 2026
DOI: 10.57676/8dcs-g212

Erklärung des Herausgebers
Dieses Werk – ausgenommen anderweitig gekennzeichneter Teile – ist lizensiert unter den Bedingungen der Creative Commons Lizenz vom Typ Namensnennung 4.0 International (CC BY 4.0), abrufbar über https://creativecommons.org/licenses/by/4.0/legalcode

Lizenz



Creative Commons Attribution 4.0 International

Verkehrssteuerung, Lichtsignalanlage, Schaltzeitprognose, C-ITS, V2X, SPATEM

Maik Halbach
DLR, Institut für Verkehrssystemtechnik, Braunschweig
Robert Oertel, Robert Markowski, Robert Alms
DLR, Institut für Verkehrssystemtechnik, Berlin

Qualität von Schaltzeitprognosedaten an Lichtsignalanlagen

Der Forschungsbericht dokumentiert die Erarbeitung eines Konzepts zur Qualitätsbewertung von Schaltzeitprognosedaten an Lichtsignalanlagen.

Moderne Lichtsignalanlagen können heutzutage unter Nutzung von C-ITS-Technologien Informationen sowohl über ihren aktuellen Signalzustand, aber auch prognostizierte, zukünftige Signalzustände, mit anderen Verkehrsteilnehmern teilen. Um den Anforderungen an zukunfts- und leistungsfähige C-ITS-Services wie beispielsweise beim GLOSA-Service (Green Light Optimal Speed Advisory) gerecht zu werden, wurde ein Konzept zur Beurteilung der Prognosequalität derartiger Anlagen entwickelt. Es umfasst drei zentrale Bewertungsschritte: Ein Dynamikindex, der die Vergleichbarkeit einzelner Bewertungen über verschiedene Knotenpunkte und Steuerungen hinweg ermöglicht; eine separate Prüfung der Integrität und Plausibilität der Prognosedaten; und eine inhaltliche Beurteilung der Prognosegüte durch die gezielte Analyse von Schaltzeitdaten. Eine abschließende Gesamtwertung in Form von Qualitätsstufen kann aus diesen Einzelbewertungen abgeleitet werden.

Das Konzept liefert als Ergebnis den Projektpartnern und Stakeholdern aus Städten und Kommunen eine erste Grundlage für eine fundierte Qualitätsbewertung von Prognosedaten und -verfahren. Es soll zukünftig in praktischen C-ITS Anwendungen detaillierter erprobt und weiterentwickelt werden, um somit ein verlässliches Werkzeug zur Beurteilung von Schaltzeitprognosen bereitzustellen.

Traffic signal control, traffic light, forecasts signal timings, C-ITS, V2X, SPATEM
(Published in English)

Maik Halbach
German Aerospace Center (DLR), Institute of Transportation Systems, Braunschweig
Robert Oertel, Robert Markowski, Robert Alms,
German Aerospace Center (DLR), Institute of Transportation Systems, Berlin

Quality of Forecasts for Signal Timings at Traffic Lights

The report documents the development of a concept for the evaluation of forecast quality for signal timings at traffic lights.
Nowadays, modern traffic light systems are capable of using C-ITS technologies to share information with other road users not only about their current signal state but also about predicted future signal states. To meet the requirements for future-oriented and high-performance C-ITS services, such as the GLOSA service (Green Light Optimal Speed Advisory), a concept for evaluating the forecast quality of such prediction data has been developed. It comprises three central evaluation steps: a signal control dynamics index that enables the comparability of individual evaluations across different junctions and signal control systems; a separate check of the integrity and plausibility of the forecast data; and a substantive evaluation of forecast quality based on the analysis of signal timing data. A final overall rating in the form of quality levels can be derived from these individual evaluations.
The proposed concept provides project partners and stakeholders from cities and local authorities with an initial basis for a systematic assessment of the quality of forecast data and prediction methods. It is intended to be tested in greater detail and further developed in practical C-ITS applications in the future, thereby establishing a reliable tool for evaluating the forecasts of traffic signal timings.



Quality of Forecasts for Signal Timings at Traffic Lights

Final report

Quality of Forecasts for Signal Timings at Traffic Lights

Final report

Written by:

Deutsches Zentrum für Luft- und Raumfahrt e.V.
Institut für Verkehrssystemtechnik
Web: https://www.dlr.de/ts/
Maik Halbach, Robert Oertel, Robert Markowski & Robert Alms
Mail: maik.halbach@dlr.de

On behalf of and in cooperation with:

Landesbetrieb Straßen, Brücken und Gewässer Hamburg (LSBG)

LSBG
Landesbetrieb Straßen, Brücken und Gewässer Hamburg

Hamburg Verkehrsanlagen GmbH (HHVA)

Hamburg Verkehrsanlagen

As part of the project:

MODI A LEAP TOWARDS SAE L4 AUTOMATED DRIVING FEATURES

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Document identification:

Title	Quality of Forecasts for Signal Timings at Traffic Lights
Topic	Report
Authors	Maik Halbach, Robert Oertel, Robert Markowski & Robert Alms
Version 1.0	25-02-2026

Contents

1 Introduction **1**
- 1.1 Motivation . . . 3
- 1.2 Approach . . . 3

2 Requirements definition **5**
- 2.1 Data sources . . . 5
- 2.2 Expert interviews . . . 10
- 2.3 Application-related configurability . . . 14
- 2.4 Core requirements . . . 15

3 Evaluation concept **17**
- 3.1 Signal control dynamics . . . 18
- 3.2 Message integrity and plausibility . . . 22
- 3.3 Forecast quality . . . 25
- 3.4 Overall rating . . . 28

4 Evaluation metric **31**
- 4.1 Signal control dynamics . . . 31
- 4.2 Message integrity and plausibility . . . 34
- 4.3 Forecast quality . . . 36
- 4.4 Overview of the evaluation metric . . . 42

5 Application **44**
- 5.1 Application example . . . 44
- 5.2 Configuration and application notes . . . 52

6 Conclusion **55**

References **57**

Glossary **59**

Appendix **61**

1 Introduction

Today, modern traffic signal controls (TSC) increasingly offer technical capabilities for transmitting information to road users in their vicinity. Signal timings are a great example for that, using C-ITS technologies such as vehicle-to-infrastructure communication (V2X). In addition to the currently signalled state at a junction, this also includes forecast data about future signal states. Road users can receive this information using appropriate communication technology and adapt their own driving behaviour if necessary, e. g. to avoid unnecessary deceleration, acceleration and waiting. Especially in the field of automated and connected driving, forecasts of signal timings are an important building block for future developments, as in certain situations infrastructural support may be necessary to optimise processes.

Various methods are available and in use today for providing forecasts signal timings. Depending on the type and characteristics of a junction and the TSC algorithm deployed, the forecast can present a challenge. This applies in particular to forecasts within actuated TSC algorithms [1]. These TSCs continuously monitor the current traffic situation via local detection technology and can adapt their signal states at short notice depending on traffic demand. This can, in turn, have a direct impact on the forecast. A high-quality forecast should be able to deal with such events appropriately, as an abrupt change in the forecast can lead to a lack of acceptance of the provided information and C-ITS services based on it. The GLOSA service (Green Light Optimal Speed Advisory), which can support vehicles on the approach to a signalised junction, should be mentioned first and foremost. It is schematically shown in Figure 1.1 and already described in current, practice-relevant guidance documents [2], [3]. Especially in the case of automated connected driving, leaps in prediction could lead to a loss of confidence and comfort, or even to dangerous situations due to abrupt braking manoeuvres.

In this context, the forecast of signal timings should be of high quality and reliable. This applies both from the perspective of a city or a traffic light operator as the entity responsible for providing the forecast, as well as for all other stakeholders who use the forecast in their services. Requirements and transparent evaluation methods for forecasts have not yet been defined in guidelines or standards. Cities, traffic light operators and other stakeholders therefore find it difficult, or even impossible, to evaluate forecast quality in a clear and



Figure 1.1: **Schematic illustration of a roadside unit (RSU) transmitting a forecast for signal timing to connected vehicles as part of a GLOSA service.**

comparable way due to the lack of a standardised evaluation metric. This creates a pressing need for action and innovation in this area.

The Free and Hanseatic City of Hamburg currently has the highest number of modern TSCs with C-ITS equipment that support the transmission of forecast data in Germany with about 300 (as of 01/2026). Various new C-ITS services are currently being trialled and rolled out there, for which high quality and reliable forecast data are required. The number of systems with standardised support for C-ITS services will continue to increase in the urban area. In terms of promoting innovation, planning and procurement aspects could therefore also come into focus in the future. A forecast quality requirement within the framework of traffic signal planning and the use of forecast methods is conceivable. The quality evaluation of forecast data is therefore of fundamental importance for the Free and Hanseatic City of Hamburg. Against this background, the State Office for Roads, Bridges and Waterways (LSBG), together with Hamburg Verkehrsanlagen GmbH (HHVA) and the German Aerospace Centre (DLR), has developed a concept for quality evaluation of forecast data. In the future, this concept should enable a transparent and comparative quality evaluation of forecast methods and data at junctions in practice based on historical data.

1.1 Motivation

The transparent and comparative quality evaluation of forecast methods and data has so far been a challenge for practitioners. It is not readily apparent to them which indicators should be used for this purpose, whether a combination of indicators is useful in the evaluation and which values represent which level of quality. Forecast providers generally offer their own tools for visualising and manually analysing specific indicators, which allow access to historical data. However, a great deal of expert knowledge is usually required to interpret the individual indicators retrieved in this way, making it difficult to draw direct conclusions about the forecast quality.

In initial evaluation approaches based on this, predicted and actual phase changes are compared over a certain time horizon and a percentage agreement is computed. The degree of agreement is correlated with the reliability of the forecast. However, the information content is limited, as this usually results in a large number of individual values that have to be interpreted with a great deal of expertise. Furthermore, attempts are often made to assess the forecast quality with the help of the deviation between the predicted and actual phase changes. Various error measures are used for this purpose, but these do not take sufficient account of time dependencies. In practice, the qualitative evaluation of forecasts also depends heavily on other variables such as the time and forecast horizon as well as the spatial conditions of a junction. These and other constraints have not yet been adequately addressed in the quality evaluation of forecast methods.

The aim of the project was therefore to develop a concept for a comprehensive evaluation metric based on real-world forecast data and current C-ITS standards, which takes into account all practical boundary conditions regardless of the manufacturer. The metric should make it possible to check forecasts transparently and with as little effort as possible so that they can be used in practice with a high degree of reliability in the future. The quality of the forecast is a key prerequisite for the practical usability and ultimately the acceptance of new types of C-ITS services. The standardised evaluation and comparability of forecast methods could significantly support both future C-ITS services and the further development of forecasts and TSC algorithms.

1.2 Approach

The approach for developing the concept can be divided into two phases. In the first phase, the requirements for forecasts were analysed. The aim was to collect all the key features

that are important for a practical quality evaluation. The requirements analysis included the following steps:

I. Firstly, the relevant data sources were analysed, i. e. in which data format forecasts actually exist or are available, what information the datasets contain and to what extent they comply with current C-ITS standards.

II. This was followed by a requirements analysis with regard to the forecast from both the practitioners' and end users' perspective. For this purpose, expert interviews were conducted with various stakeholders from all relevant domains.

Building on this, an evaluation concept was developed in the second phase, which incorporates the requirements collected in the first phase and represents them in suitable indicators and metrics for quality assessment. The concept was developed, tested and verified in an iterative process based on real-world forecast data. The following steps were carried out:

III. First, all relevant and suitable indicators were selected according to the requirements analysis, conceptually described and their applicability was analysed with regard to the desired evaluation. Interactions and dependencies between the individual indicators as well as weighting proportions were considered with regard to a combined overall rating.

IV. Based on this selection of indicators and their combination, the corresponding calculation bases for the quality evaluation were established.

Based on this approach, this report is structured as follows: Chapter 2 summarises the requirements for an evaluation metric of forecast quality. Chapter 3 sets out a concept for such a metric, while Chapter 4 discusses concrete proposals for implementing the evaluation metric in the form of mathematical descriptions and configurations. Chapter 5 illustrates the possible results of such an evaluation metric in a practical example. Finally, chapter 6 summarises the collected findings on the evaluation concept, discusses open questions and provides an outlook for future research activities.

2 Requirements definition

The following chapter summarises the basic requirements resulting from both the data sources and the findings of the expert interviews.

2.1 Data sources

At a connected TSC, information from a junction can be shared with other road users via vehicle-to-infrastructure communication (V2X). Two V2X message types are particularly relevant for this: the *Map Extended Message* (MAPEM) and the *Signal Phase And Timing Extended Message* (SPATEM). These two types are closely linked in order to be able to send geographical information in combination with the current signal status of a junction. In Europe, these V2X messages are defined and specified by the ETSI organisation and the C-Roads platform in order to harmonise the development and introduction of message standards within the European group. Figure 2.1 provides an overview of the standardisation levels involved as well as the message types and data sources used in this work. They are described in more detail below.

2.1.1 Standardisation

ETSI: The European Telecommunications Standards Institute (ETSI) is a leading European standardisation organisation that develops standards for information and communication technologies. The V2X message types MAPEM and SPATEM relevant to the project are specified in detail by the ETSI.

C-Roads: Via the so-called C-Roads platform, authorities and road infrastructure operators have joined forces in a European initiative to support the introduction of cooperative

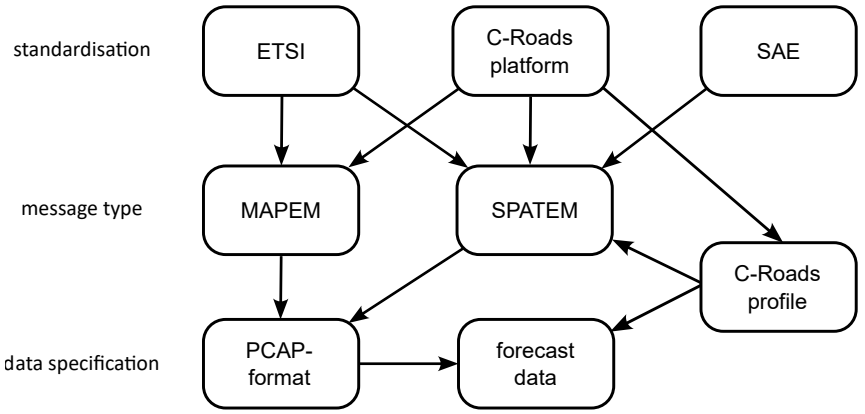


Figure 2.1: **Overview of standardisation levels, V2X message types and data sources.**

intelligent transport systems (C-ITS) and to advance these developments by harmonising functionalities and communication systems. This also includes V2X-supported services in road traffic, such as the forecast for signal timings at traffic lights addressed in the project. Based on the ETSI standard, the C-Roads initiative specifies how the data fields of certain V2X message types must be filled depending on the application purpose. C-Roads does not define any new message standards of its own, but specifies usage concepts for C-ITS services, so-called Facility Layer Services (FLS), which refer to the standards of the European Telecommunications Standards Institute (ETSI) or the International Organisation for Standardisation (ISO).

SAE: The J2735 [4] standard from the Society of Automotive Engineers (SAE) is another standard for V2X message specifications, which is mainly used in North America. Compared to ETSI, this is an international organisation that operates beyond the EU and is strongly supported by the automotive industry. In parts, both ETSI and C-Roads refer to the ISO standards of the SAE. The SAE standard specifies how the data fields of a SPATEM message should be filled for forecasts. However, as this standard differs in part from the ETSI standard established in the EU and is rarely used in practice in Europe, this report only takes into account the ETSI standard and the associated C-Roads profile.

2.1.2 V2X message types

MAPEM: The MAPEM message (MAP (topology) Extended Message) describes the topology of a junction and is defined in the ETSI standard [5]. In particular, it records the spatial layout of the lanes and the logical structure of the traffic movements within the junction.

SPATEM: The SPATEM message (Signal Phase And Timing Extended Message) describes the current signal states of a traffic signal and can, if applicable, also contain forecasts of the signal states. SPATEM is also defined in the ETSI standard [5]. Typically, a MAPEM message references the lanes of a junction, whose signal states are included accordingly in a SPATEM. The C-Roads Initiative handbook on MAPEM and SPATEM creation [6] specifically addresses this interconnection between the two message types and provides a detailed description of all message elements that are relevant for later applications (see C-Roads profile). The forecast in a SPATEM is not mandatory according to the intended application, which is why the corresponding data fields are not always populated.

2.1.3 Data specification

C-Roads profile: The C-Roads platform specifies the technical requirements for certain message types and their content in detail, structured according to functional Facility Layer Services (FLS) [7]. These FLS define the necessary message contents to, for example, provide warnings about safety-relevant events, information about roadworks as well as notifications of speed or traffic sign changes. Forecasts can be found in the FLS "Traffic Light Maneuver (TLM) and Road and Lane Topology (RLT) Service (TLM FLS and RLT FLS)". In this report, this FLS from the "Message Profiles" of the C-Roads Initiative [7] (Release Version 2.2.0) is referred to as the "C-Roads Profile".

Figure 2.2 illustrates the logical relationship between the required data fields of a SPATEM message for such a C-Roads profile, which are relevant for forecasts. The core information for predicting an upcoming phase change, defined as `movementEvent`, can be derived from the values `minEndTime`, `maxEndTime`, `confidence` and `likelyTime`. The signal state itself is represented in the SPATEM by the `eventState`.

The terms derived from the C-Roads profile are described in more detail below:

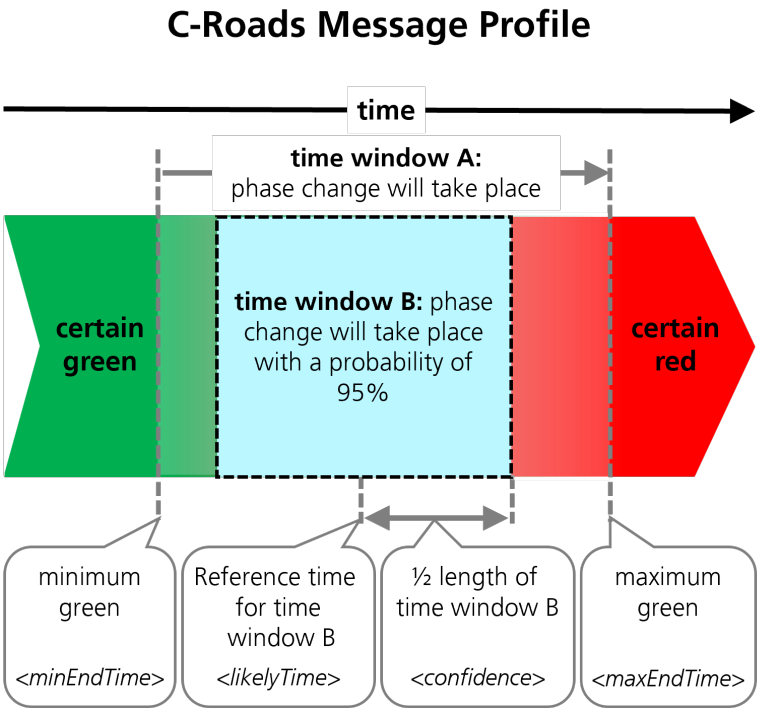


Figure 2.2: **Illustration of the C-Roads message profile with the four key values of forecasts for a phase change: minEndTime, likelyTime, confidence and maxEndTime.**

- **time window A:** Describes the interval in which a phase change is certain to take place. This range is defined by the values `minEndTime` and `maxEndTime` for the start and end of the time window.

- **minEndTime:** Describes the earliest phase change of the current `eventState`. Before this time, the current `eventState` is still definitely active.

- **maxEndTime:** Describes the latest phase change of the current `eventState`. After this time, the current `eventState` is definitely no longer active.

- **time window B:** Describes the time interval in which the phase change of the current `eventState` will occur with a probability of at least 95%. The window B results from the values of `likelyTime` and `confidence`, i. e. it extends from `likelyTime` − `confidence` to `likelyTime` + `confidence` and thus corresponds to twice the respective `confidence` value. Note: `confidence` values are interpreted as time values (see table 2.1).

- **likelyTime:** Reference time for time window B, which ideally represents the most probable phase change.

- **confidence:** The `confidence` defines the width of time window B around the `likelyTime`. The actual value of the `confidence` is specified as a time interval class between 0 and 15 (see Table 2.1). The smaller the value of `confidence` is, the larger the time window B is.

Table 2.1: **Confidence: Mapping between class and time interval**

Value	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Time interval [s]	>15.0	13.5	12.0	10.5	9.0	7.5	6.5	5.5	4.5	3.5	2.5	2.0	1.5	1.0	0.5	0

Recording format PCAP: The C-Roads platform has established a format for recording V2X communication, which also contains the V2X messages sent or received, called Packet Capture (PCAP) [8] to improve interoperability and cross-manufacturer harmonisation. It is generally used to record V2X messages in a standardised format and to exchange log files accordingly. This project based its evaluation on SPATEM messages recorded in PCAP format.

Real-world data: The project used real-world signal and forecast data. This data was collected at five signalled junctions in Hamburg over a period of 24 hours and recorded in the form of SPATEM messages as PCAP files. The forecast evaluation was carried out exclusively on the basis of the `movementEvents` corresponding to the current signal status within a SPATEM message. All junctions adhered to the C-Roads profile. The site selection took into account different traffic engineering characteristics, e. g. with regard to geometric and control complexity, in order to cover the widest possible range of boundary conditions in the data analysis and to support the development of the evaluation metric. Apart from the information contained in the SPATEM messages, no other data sources were used. This applies in particular to recordings from TSCs, which are not always immediately available and do not follow any generally defined standard in terms of recording format, which makes it difficult to incorporate such data in this concept.

2.2 Expert interviews

Guideline-based expert interviews were conducted as part of the development of the evaluation metric. A total of 16 German and European experts from various stakeholder groups were interviewed, including representatives of cities and municipalities, infrastructure and public transport operators, traffic light manufacturers, the automotive industry, research institutions and individual forecasting providers.

2.2.1 Objective and approach

The aim of the interviews was to record requirements for forecasts and a possible evaluation. A questionnaire with 18 questions served as the basis (see appendix questionnaire). As the stakeholders come from different domains, a glossary and a schematic diagram were additionally prepared for the interviews, which served as a shared terminology basis and foundation for discussion.

2.2.2 Key topics and findings of the expert survey

The catalogue of questions was designed to cover both requirements for the provision and usage of forecasts and an evaluation metric based on these requirements. In terms of content, the questions can be categorised into four key topics. The analysis of the interviews shows a heterogeneous, but overall consistent picture of the usage, evaluation and requirements of forecasts. The most important findings on the four key topics are summarised below:

(1) Application of forecasts: How are forecasts used today and which user groups, i. e. road users who receive the forecast information as end users (e. g. car traffic, public transport (PT), cycling), should benefit from them?

- Forecasts are primarily used to provide them via V2X messages (SPATEM) for vehicle-based services. The focus is on GLOSA-like applications for car traffic; cyclists and public transport play a secondary role.

- The use of the C-Roads profile dominates among stakeholders; the SAE standard is used in some cases. In the practical use of forecasts, the focus is particularly on `likelyTime` and `confidence`; `minEndTime` and `maxEndTime` are used much less frequently.

(2) Evaluation of forecasts: In what form are forecasts currently evaluated, which methods, tools or indicators are used for this purpose, and where are the current deficits?

- A systematic, data-based evaluation of forecast quality has not yet been carried out. There is no standardised approach or established evaluation tool; in some cases, manufacturer-specific technical reports exist or simple visualisations are used to assess key indicators.

- The evaluation of forecast quality is often based on subjective expertise, meaning that a reproducible evaluation is currently hardly possible.

(3) Expectations of forecasts: What functional expectations do practitioners and users have of forecasts, and what criteria can be derived from them for analyses based on existing real data?

- The specifications defined in a profile for the values `minEndTime`, `maxEndTime`, `confidence` and `likelyTime` should be reliably adhered to.

- Forecasts should be continuously available and become more accurate as the actual switching time approaches, i. e. the deviation should decrease and the probability of occurrence should increase. At the same time, under changing traffic conditions and resulting control interventions (due to, e. g., traffic-actuated control, public transport priority, or emergency vehicle priority), updates are required without abrupt changes occurring outside such situations.

(4) Requirements for an evaluation metric: What properties must an evaluation metric have in order to meet practitioners' needs, and in which application context should it be usable?

- An evaluation metric is expected to provide several meaningful characteristic values and to present them in an understandable way using school grades or levels. The

metric should be machine-readable and enable evaluation by time interval or signal groups.

- The application contexts mentioned include checking the forecast quality, the comparability of forecasts and forecast providers, support for decisions on the usage or location selection of forecasts, the optimisation of TSCs and the formulation of requirements in tender documents.

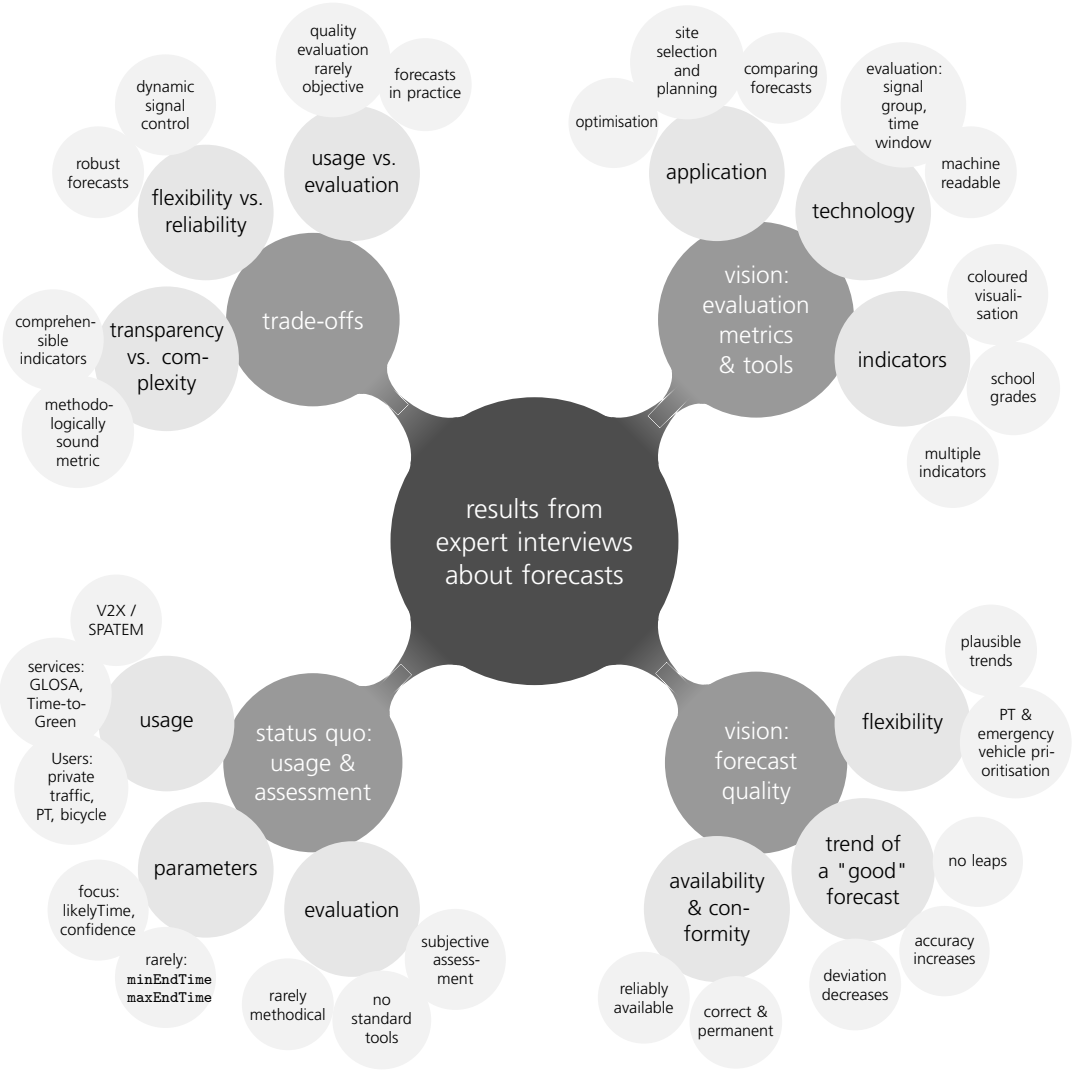


Figure 2.3: **Schematic summary of expert interviews on the status quo, requirements and trade-offs in forecasts and their quality evaluation.**

Figure 2.3 summarises the most important answers from the expert interviews in a mind map. Overall, the expert interviews provided important insights into the topic of forecasts for signal timings at traffic lights as well as their usage and current evaluation. This includes, for example, which behaviours a supposedly good forecast should have, or which criteria could be derived in order to use them for the evaluation of forecasts. Furthermore, initial aspects of how an evaluation metric itself could be structured were uncovered. Some trade-offs that have not yet been addressed in the assessment and evaluation of forecasts were also identified. After evaluating the expert interviews, the following requirements for forecasts and an evaluation metric can be derived.

- **Profile conformity:** Forecasts and their data should fulfil the requirements defined in the standard and profile. Only standard and profile-compliant data enables the usage of forecasts.

- **Availability:** If forecasts are provided, they should be continuously available. This is the only way to provide road users with permanent support when approaching and passing a traffic light.

- **Reliability:** Forecasts should become more accurate and reliable the shorter the corresponding forecast horizon is.

- **Comparability:** An evaluation metric should enable an evaluation and manufacturer-independent comparison of forecast methods and forecasts.

- **Configurability:** The use of forecasts in different scenarios requires a configurable evaluation metric. In particular, the metric should allow the selection and weighting of time periods and signal groups that could be of particular interest.

- **Comprehensibility and transparency:** The evaluation metric should provide the evaluation of forecasts with several informative indicators and present them in an understandable way in the form of school grades or levels.

- **Technical implementation:** An evaluation metric should be machine-readable and designed for automated analysis.

2.3 Application-related configurability

Depending on the application purpose, there can be very different requirements for forecasts. A categorisation according to end users, services and practitioners is shown below using examples of how the forecast requirements can differ accordingly:

I. **End users**: End users are road users of specific traffic modes, for example in private motorised traffic, heavy goods vehicles, public transport or walking and cycling, who use services based on forecasts during their journeys in the vicinity of a traffic signal.

   Example of the relevance of the forecast horizon: The speed of a road user strongly affects the time and travel distance available to react to forecasts in time and adapt driving behaviour accordingly.

II. **Services:** Services are applications that use forecasts to support end users when approaching and passing a traffic light, e. g. GLOSA, Time-to-Green, Time-to-Red and driver assistance systems (ACC, Traffic Light Assist, Brake Assist).

   Example of the relevance of the message content of forecasts: A Time-to-Green service requires precise predictions of when green begins. The start time of red is irrelevant for this.

III. **Practitioners:** Practitioners are stakeholders who provide forecasts or use them for their own services, for example municipalities and cities, companies in the automotive sector, transport companies, traffic light operators, traffic light and forecast manufacturers or app providers.

   Example of the focus on traffic type-related forecasting: transport companies are primarily interested in the forecast of signal groups for public transport; the other signal groups are of little relevance.

In order to meet this variety of sometimes even contradictory requirements for forecasts, the evaluation metric should be configurable. The concept should allow the focus of the evaluation to be set deliberately. This could include, for example, specific message contents, signal states, signal groups, or forecast horizons that are of particular interest depending on the application purpose for an end user, service, or practitioner.

2.4 Core requirements

Based on the technical context summarised in section 2.1 and the differing trade-offs and expectations regarding the forecast of signal timings identified in the expert interviews in section 2.2, the following key requirements for a quality evaluation concept are derived:

[A] Establishing comparability: Forecasting methods and TSCs can differ greatly from one another. The question therefore arises as to how much TSCs actually differ in their signal control dynamic and how challenging a prediction is in relation to the TSC. Highly dynamic, actuated TSCs are much more difficult to predict than more static TSCs.

⇒ A measure for evaluating signal control dynamics is required to establish comparability between forecasts across manufacturers.

[B] Checking messages and data for availability and internal consistency: A clear expectation was expressed that, if forecasts are to be used in practice, they should be continuously and persistently available in accordance with the SPATEM definition and consistent with the C-Roads profile.

⇒ From these stated requirements regarding the availability, reliability, and profile conformity of forecasts, it follows that a systematic validation of the messages for integrity and plausibility is required.

[C] Evaluating the forecast quality: The reliability and accuracy of forecasts as a function of their forecast horizon were identified as a particular challenge and requirement. Essentially, shorter forecasts are expected to be more accurate than forecasts with very long forecast horizons. In line with this expectation, the `confidence` should also be higher for shorter horizons than for longer ones. A reliable forecast is required to show only the smallest possible deviations without abrupt changes.

⇒ To evaluate the actual forecast quality, a measure is required that accounts for the dependencies between forecast horizon, accuracy and `confidence`.

[D] Comprehensibility and simultaneous flexibility of an evaluation metric: Given the complexity of forecasts and the technical expertise required for their evaluation, the standardisation, simplification and comprehensibility of an evaluation metric was emphasised. In addition, section 2.3 presented the requirement for application-oriented configurability, which resulted from very different purposes of a quality evaluation between the very heterogeneous stakeholder groups.

⇒ To enable a manufacturer-independent evaluation that is also easily understandable for different practitioner groups through evaluation levels, the evaluation metric should map complex individual indicators onto a few clearly defined quality classes. These classes should be configurable for different application profiles, thereby providing the necessary flexibility.

The following evaluation concept in chapter 3 is fundamentally derived from these key requirements. Additional detailed requirements from the expert interviews, and in part from the experience gained with real-world data during the data analysis, are incorporated into specific aspects of the overall evaluation and are included in chapter 4. Chapter 5 specifically addresses the configurability of the individual evaluations.

3 Evaluation concept

The project to develop an evaluation concept aims to derive an overall assessment analogous to a school grading system, similar to the approach used for classifying the evaluation levels of traffic flow in the Manual for the Design of Road Traffic Facilities (HBS) [9, Part S4.2]. The measure for evaluating forecast quality is composed of several aspects and tasks, as illustrated in Figure 3.1.

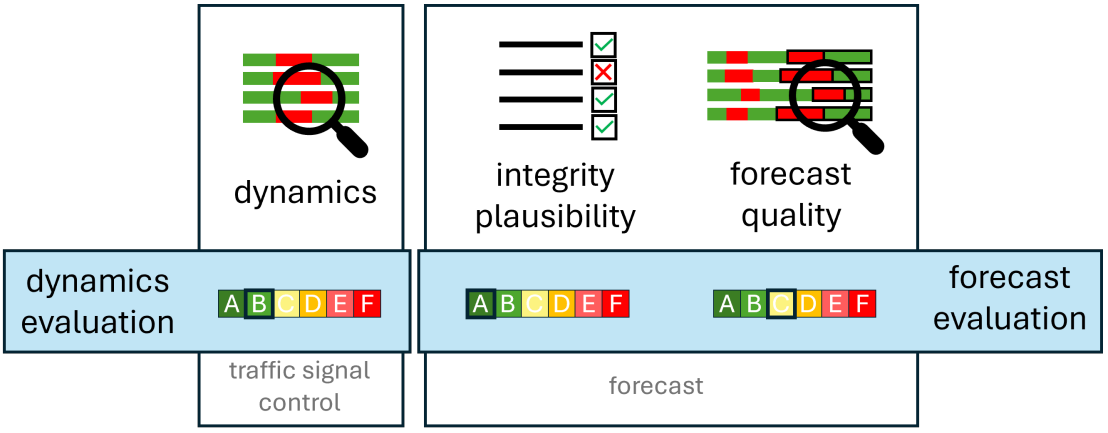


Figure 3.1: **Illustration of the concept for evaluating forecast data. Comparability between TSCs is ensured by evaluating *signal control dynamics*; the quality of forecasts is evaluated based on the aspects *integrity and plausibility* as well as *forecast quality*.**

To enable a targeted evaluation for a specific application, such as a signal group, an appropriate selection of the characteristic values to be considered must first be made, based on the specific characteristics of the respective signal group. For example, if the initial goal is only to assess the basic suitability of a dataset for a later, more detailed analysis, examining individual characteristic values as part of a plausibility check may already be sufficient to obtain a first evaluation of data quality in this case. Such an assessment of individual characteristic values is accommodated in the evaluation concept through a flexible configuration of the evaluation metric. In contrast, a much more comprehensive analysis, such as comparing two junctions, requires a broader, targeted selection of signal

groups, signal states, and characteristic values of the forecast, including specific weightings of these characteristic values within the dataset. Configuring the evaluation metric in this way therefore requires a clear definition of the analysis objectives to match the application context and the complexity of the evaluation. Chapter 5 discusses an example configuration for a possible application scenario. Regardless of the chosen configuration, the following steps are fundamentally required for a proper evaluation:

I. In a first step, the **signal control dynamics** of the TSC under consideration is determined, as this has a significant influence on the predictability of TSCs.

II. In the second step, the **integrity** and **plausibility** of the data in a message or dataset are checked. This includes verifying formal minimum requirements for the forecast that must follow logically from the message contents. The availability of the forecast itself is also analysed.

III. In step three, the **forecast quality** is determined by comparing the content of the forecast with actual signal timings.

IV. In the last step, the signal control dynamics are reported as an independent characteristic value. Separately, the results for message integrity and plausibility as well as the forecast quality are combined into an **overall rating**, using predefined weightings for their respective aspects.

In the following sections 3.1–3.4, the relevant indicators are explained in detail, defined and discussed in more detail using examples.

3.1 Signal control dynamics

The occurrence of random events in a system under consideration significantly influences its predictability. For example, the signal states of a fixed-time TSC (shown schematically in Figure 3.2 A) are almost trivial to predict. An actuated, slightly more dynamic TSC (see Figure 3.2 B) may still be easy to predict, while in comparison actuated, highly dynamic TSCs (see Figure 3.2 C) may not show a clear pattern due to the adaptation to random traffic. The traffic demand has an important influence on the degree of observable signal control dynamics. For example, a highly actuated TSC can actually behave very dynamically in times of higher traffic demand, while rigid behaviour can occur during periods of very low traffic demand or at saturation. In the latter cases, minimum or maximum defined green

durations are applied, which can be equated with low signal control dynamics. Various other influencing factors are conceivable here, such as frequent prioritisation of public transport or heavy pedestrian and bicycle traffic flows, which are flexibly taken into account by corresponding requirements within a dynamic TSC.

Therefore it only makes sense to compare the forecasts from two TSCs if both controllers exhibit similar signal control dynamics and, consequently, a similar level of predictability under comparable traffic demand. The signal control dynamics of a TSC is therefore described below as an index that characterises the randomness its observed behaviour. This signal control dynamics index serves as an indicator of the comparability of different TSCs with regard to their forecast quality.

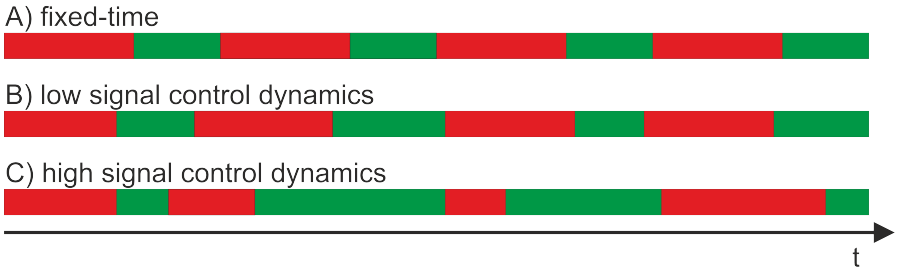


Figure 3.2: **Schematic representation of various dynamic characteristics of TSCs, ranging from A) with precisely repeatable switching operations, to B) with minor fluctuations in phase durations, to C) with significant fluctuations in phase durations.**

When evaluating signal control dynamics, the focus is on dynamic signal states. Dynamic signal states are signal states whose phase duration can vary, for example by extending a phase. Typically, this only affects the signal states `red` and `green`. Static signal states, on the other hand, have fixed duration; these primarily include `clearance (amber)` and `preparation (red-amber)`. As static signal states do not contribute anything to the signal control dynamic, they are not considered further below.

Otto et al. [10] have already investigated signal control dynamics on the basis of signal states from real-world data. The authors analysed the start, end and duration of green phases as well as their distributions in order to determine stability indicators. These stability indicators can be used to assess the predictability, and e. g. the suitability of a TSC for GLOSA services. Following this approach, the signal control dynamics are evaluated using the start and end of an `eventState`, as outlined below.

In SPATEM messages, on which the following analysis is based, signal states are represented `eventStates`. In the further analysis, signal state and `eventState` can therefore be treated

as equivalent. A special feature is the additional differentiation of the `eventState` for `clearance` and `green` in *protected* and *permissive*; this differentiation cannot be recognised from the signal states alone.

Dynamic states: Several aspects are considered to evaluate the signal control dynamic. Figure 3.3 illustrates these factors as follows:

- The observed phase durations, i. e. the length of the respective time intervals, are analysed (see Figure 3.3 A). Greater variability in phase durations and a more irregular distribution make the timing of the intervals harder to forecast.
- Coordination timing plans can constrain the behaviour of a TSC and thus create recurring patterns. These patterns usually become visible as fixed phase changes in relation to the cycle time of the TSC. For this reason, the time gap between the start and end of each signal state, i. e. the phase duration (see Figure 3.3 B and C), are also analysed.

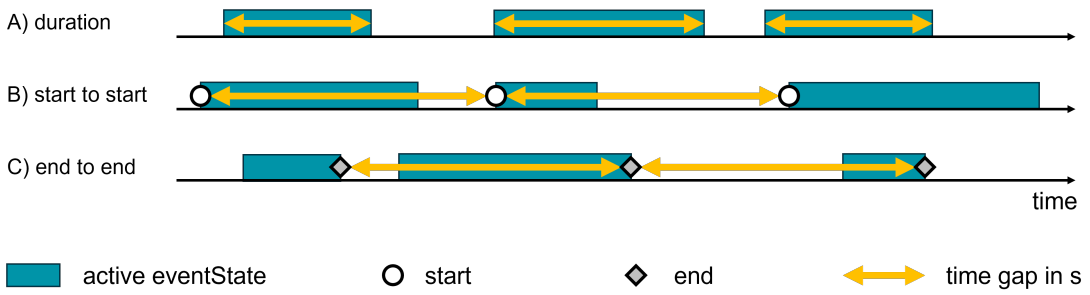


Figure 3.3: **Illustration of the main aspects for analysing signal control dynamics: A) Length of active event states, B) Start-to-start relationship of event states, and C) End-to-end relationship of *eventStates*.**

Information entropy: The information entropy introduced by Shannon [11] is used to determine the randomness of the observed values; the detailed calculation steps are described in section 4.1. It is a measure of the unpredictability of the observed time intervals. The information entropy tends to increase with the number of observed time intervals, provided that these occur frequently and thus contribute to the overall uncertainty. On the other hand, the information entropy decreases with fewer observed time intervals or a high

probability for a few time intervals. It is therefore suitable for quantifying the intervals at the beginning and end of the signal state (see Figure 3.3 B and C).

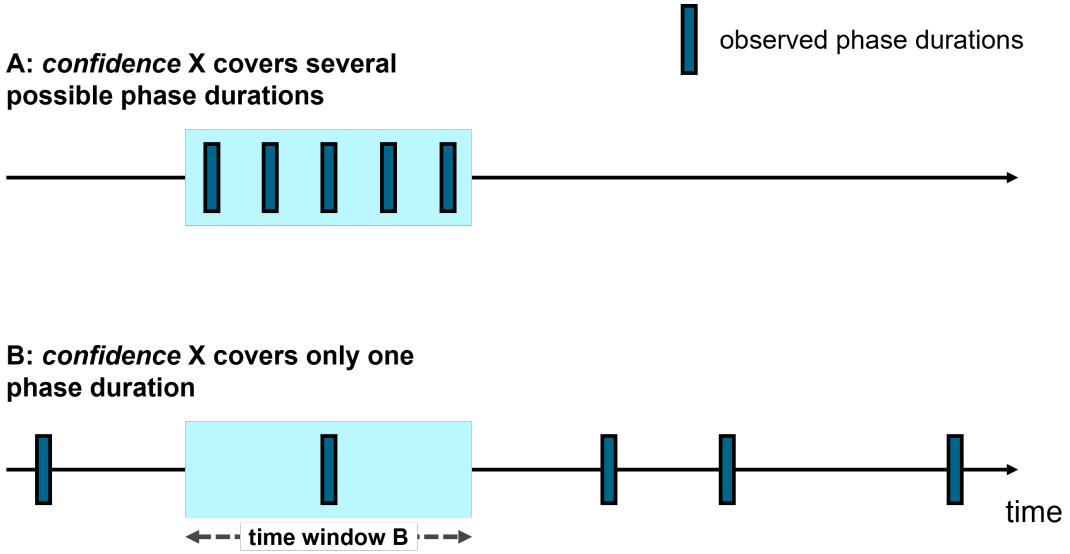


Figure 3.4: **The *confidence* allows for multiple phase durations to be covered by time window B if the phase durations are close together.**

***Confidence*:** In addition, the `confidence` can be included when analysing the phase durations for the evaluation of the signal control dynamics (see Figure 3.4). If the observed phase durations in a dataset are within a relatively narrow range, this range may be completely covered by a suitable `likelyTime–confidence` combination, as shown in Figure 3.4 A. If, on the other hand, larger gaps occur in the set of observed phase durations, the same `likelyTime–confidence` combination can no longer cover all phase durations (Figure 3.4 B). For this reason, Shannon's information entropy is adapted so that the time gaps in the observed phase durations contribute to a further increase in entropy, thereby better reflecting the unpredictability of the TSC in such cases.

3.2 Message integrity and plausibility

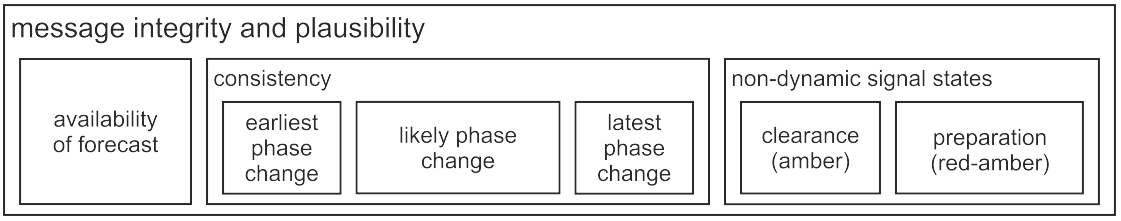


Figure 3.5: **Overview and structure of the plausibility and integrity evaluation.**

To assess the data integrity and plausibility, messages are evaluated without reference to the actual signal timings. Figure 3.5 provides an overview of this approach. A key prerequisite is the availability of the forecast, since high availability is required to deploy forecasts in various applications (e. g. GLOSA or Time-to-Green). The C-Roads profile and constraints from traffic engineering also require certain dependencies between the message contents, in particular consistency relationships of the signal timings and a fixed duration of non-dynamic signal states. The individual aspects are explained in more detail below.

Availability: The availability of the forecast is determined by the proportion of SPATEM messages that contain a profile-compliant forecast. In addition, at least one SPATEM message per second is expected over the observation period, since TSCs typically operate on this fixed time interval and apply state changes accordingly.

Shift of earliest and latest phase change: The values `minEndTime` and `maxEndTime` in an `eventState` describe the earliest and latest possible phase change, respectively. For consecutive messages, an earliest time `minEndTime` may remain unchanged or only shift later. Likewise, a latest time `maxEndTime` may remain unchanged or only shift earlier. Figure 3.6 illustrates this plausibility check schematically for both cases. Any shifts other than those described above are considered implausible.

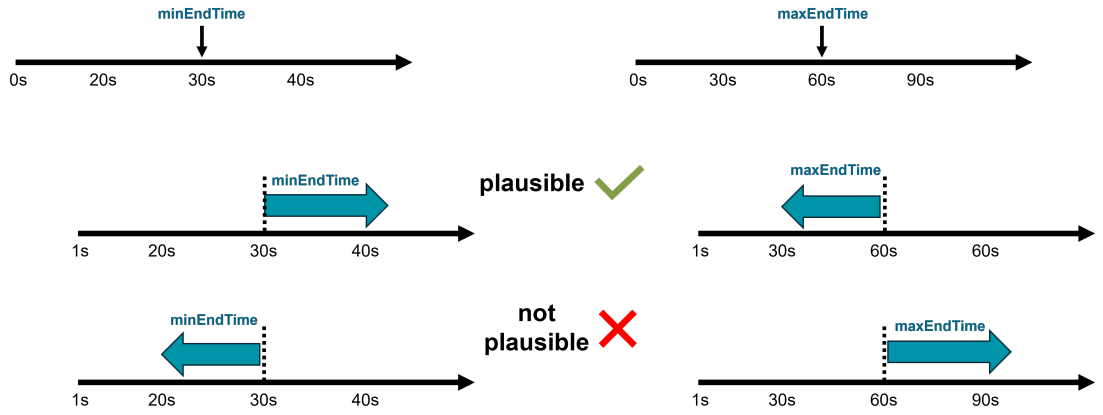
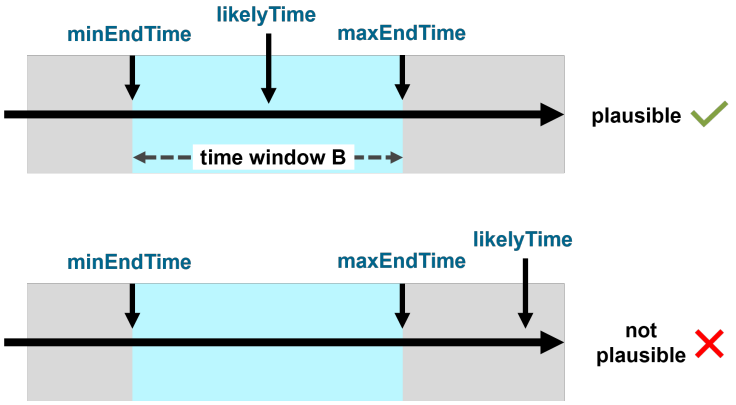


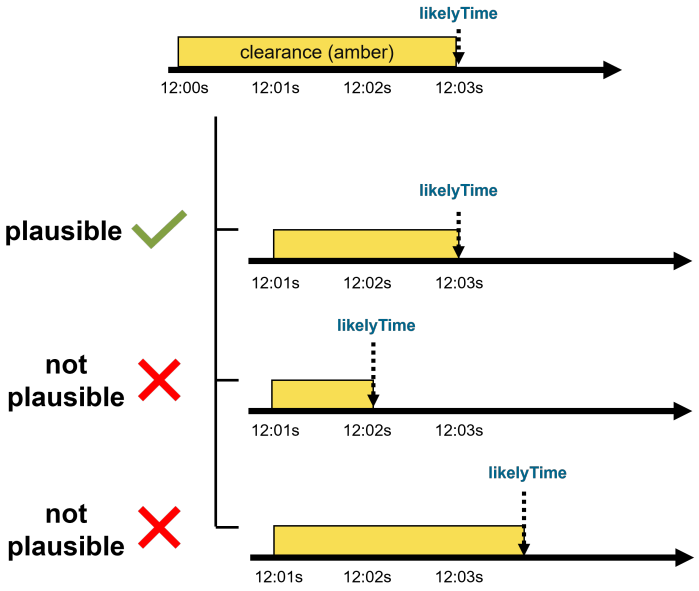
Figure 3.6: **Example of a plausibility check for the values *minEndTime* and *maxEndTime*.**

Logical consistency constraints: In addition, it is checked whether the values `minEndTime`, `maxEndTime`, and `likelyTime` are in a logically consistent temporal relationship. Figure 3.7a illustrates this plausibility check for `likelyTime`, which must lie between `minEndTime` and `maxEndTime`. Furthermore, it must be verified that `minEndTime` is less than `maxEndTime`.

Phase duration: Non-dynamic signal states, in particular `clearance (amber)` and `preparation (red-amber)`, have a fixed phase duration and therefore a well-defined, immutable phase change. If the predicted phase change of a non-dynamic signal state shifts during the forecast, i. e., while the system is currently in that non-dynamic signal state, this is considered implausible behaviour. Figure 3.7b illustrates such a plausibility check using `clearance (amber)` as an example.



(a) **Plausibility check of the *likelyTime* forecast in relation to *minEndTime* and *maxEndTime*.**



(b) **Checking the plausibility of clearance based on different *likelyTime* shifts.**

Figure 3.7: **Illustrative examples for plausibility checks of the *likelyTime*.**

3.3 Forecast quality

As no established measure for the evaluation of forecasts is developed yet, a concept for a data-driven method is described below. In particular, the interaction of forecast horizon, forecast accuracy and the associated `confidence` is analysed, since these three aspects strongly influence each other, as Figure 3.8 illustrates. Firstly, the maximum, C-Roads-compliant forecast horizon is determined. Within this scope, the average `confidence` values are assessed and weighted according to their associated forecast horizon. The individual steps are explained below.

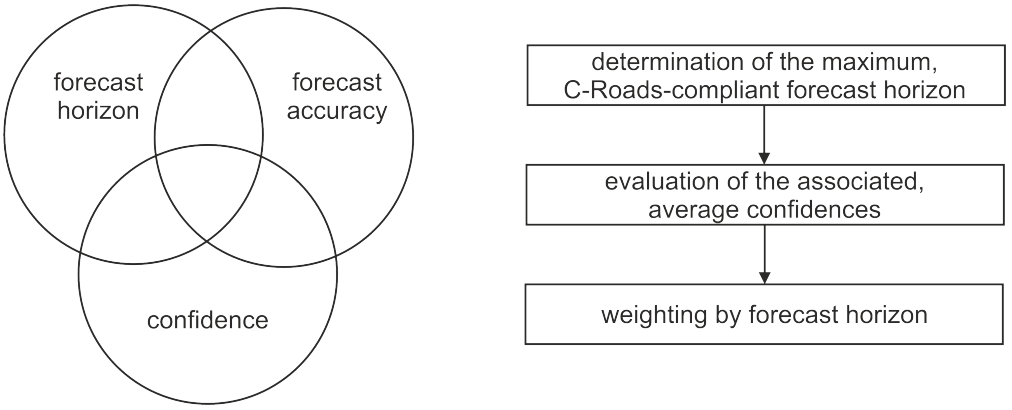


Figure 3.8: **Overview of the relationships between forecast horizon, forecast accuracy and *confidence* as well as the processing steps for determining the forecast quality.**

Deviation of the forecast accuracy: In a first step, the available data is used to determine the forecast horizons up to which the forecasts achieve the minimum accuracy of 95% required by the C-Roads profile. The forecast horizon is represented by the associated `likelyTimes` of the forecasts. Figure 3.9 shows two examples of such a dataset analysis with regard to forecast accuracy. For each of the messages, it is determined whether the actual phase changes are within the acceptable time window B, which is defined by the `likelyTime` and `confidence`. If more than 95% of the actual phase changes lie within the time window, the requirements of the C-Roads profile are met for these `likelyTimes` (see Figure 3.9 above). Conversely, `likelyTimes` outside this acceptance range do not fulfil the requirements of the C-Roads profile (see Figure 3.9 bottom).

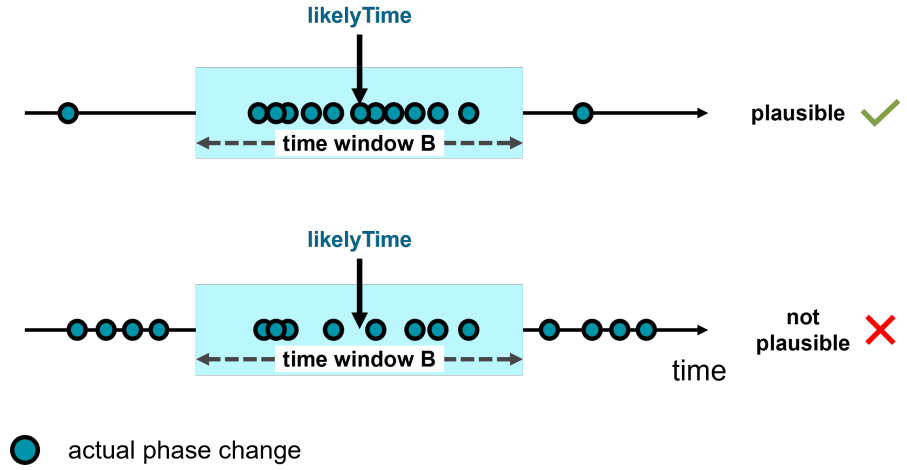


Figure 3.9: **Permitted and prohibited deviation in forecast accuracy. Top: 95% of the actual phase changes are within their forecast time window. Bottom: More than 5% of the actual phase changes are outside their forecast time windows.**

Forecast accuracy versus informativeness: The evaluation of forecasts based on accuracy alone is not particularly informative. This will be illustrated and discussed with the help of Figure 3.10: Subfigure A1 illustrates a good forecast, as in this case the `likelyTime` is close to the actual phase change with simultaneously high `confidence`. Such a forecast would be very useful for a practitioner. Subfigure A2 shows a case that illustrates a significantly better forecast compared to subfigure B2 shown later, as the `likelyTime` is closer to the actual phase change. However, an end user would not know the actual phase change in a live situation; it is therefore not possible to tell from the visible data pattern alone whether the corresponding `likelyTime` is actually accurate or inaccurate. The interpretation of A2 is therefore similarly difficult as with subfigure B2. Conversely, subfigure B1 shows an incorrect forecast: The `likelyTime` deviates significantly from the actual phase change, although the `confidence` is high. The forecast initially appears to indicate trust, but its content is incorrect and therefore problematic for practitioners. Finally, subfigure B2 shows a forecast in which the `likelyTime` is very far away from the actual phase change. With a simultaneously low `confidence`, the time window for a possible phase change is so wide that the actual phase change still falls within this potential time window. Formally, such a forecast would still be correct. Such a case can be perfectly acceptable for forecasts with a long forecast horizon, but is not very helpful for forecasts in the near future.

These considerations on Figure 3.10 are based on a phenomenon observed when analysing the real-world data provided, which can be summarised as follows:

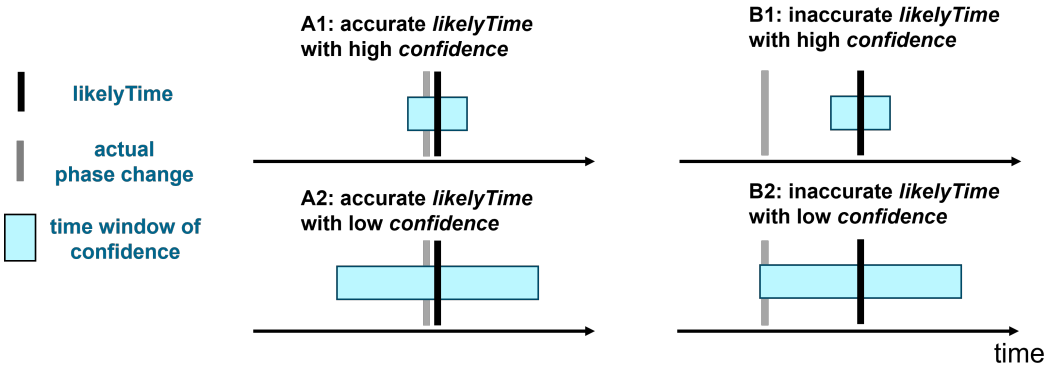


Figure 3.10: **Interaction of *likelyTime* and *confidence*: A1) the actual phase change is predicted with high *confidence*, A2) the actual phase change is predicted with low *confidence* and therefore has limited informative value, B1) an incorrect forecast and B2) a forecast with inaccurate *likelyTime*, but nevertheless correct due to a low *confidence*.**

It was frequently observed that the `confidence` decreases significantly with decreasing `likelyTime`, i. e. for forecasts in the near future. Contrary to users' expectations, the forecasts become increasingly uncertain as the predicted phase change approaches. In these cases, the accuracy of the respective forecasts appears to increase at the same time, because the wider forecast time window automatically means that more phase changes fall within this time window.

Give that the user survey suggests forecasts are expected to become more reliable with a shorter the forecast horizon, accuracy, `likelyTime` and `confidence` should be evaluated jointly, not separately.

Forecast quality as a weighted measure:

Based on the considerations outlined above and the requirements of the expert surveys, a weighted measure is introduced for the evaluation of forecast quality:

The forecast quality is described as a weighted sum of an aggregated `confidence` rating per `likelyTime`.

The evaluation of `confidence` is based on the width of the forecast time window: the wider the time window, the lower the rating. For this purpose, an evaluation function

is used that flattens out, as the usability no longer decreases significantly from a certain window width. However, such a function drops relatively sharply at the beginning. This is a circumstance that may require further optimisation of the evaluation methodology in the future.

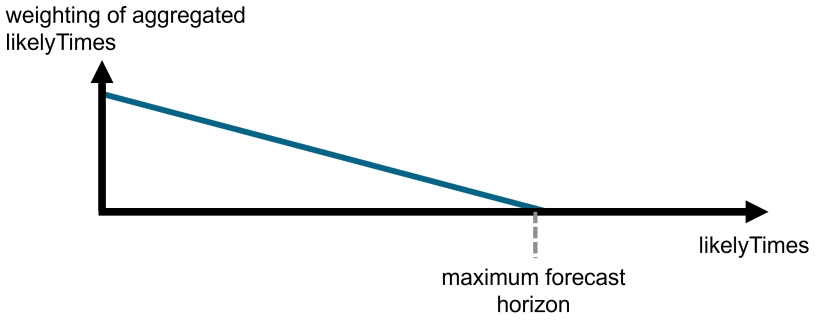


Figure 3.11: **The weighting of *likelyTime* decreases continuously up to the maximum forecast horizon in order to give greater weight to forecasts for the near future.**

Figure 3.11 illustrates how such a weighting per `likelyTime` is to be interpreted for a large number of data points. The higher the relative `likelyTime`, the further in the future the associated predicted phase change is. The weighting is performed by aggregating the evaluation of the `confidence` per associated relative `likelyTime` and weighting them with the underlying function (in this case a linear function). In addition, a search is made for the threshold in the forecast horizon above which the forecast accuracy no longer fulfils the 95% accuracy according to C-Roads. All evaluations of `confidence` above this threshold are set to zero to ensure C-Roads conformity. This means that the closer the predicted phase change approaches, the more heavily the associated aggregated `confidence` is weighted. A low `confidence` in the near future is therefore penalised more than a low `confidence` in the distant future. In addition, a violation of the required accuracy from the C-Roads profile leads to significant losses in the quality evaluation.

3.4 Overall rating

Based on the evaluation levels of traffic flow from the Manual for the Design of Road Traffic Facilities (HBS), the evaluation concept follows the principle of evaluation levels according to a quality index. Such a quality index results from the assignment of individual scores depending on the respective criterion under consideration. Figure 3.12 illustrates the

principle of such a quality evaluation in six levels. In order to achieve a certain evaluation level, a threshold value must be exceeded or fallen short of depending on the criterion considered. The evaluation levels must be defined specifically and appropriately in terms of their significance and threshold scaling. An evaluation according to such standardised levels thus enables comparability across different TSCs.

The overall rating of a forecast can combine several aspects, with each subindex contributing to the overall rating.

evaluation level	rating	value
A	excellent	...
B	very good	...
C	good	...
D	average	...
E	poor	...
F	very poor	...

Figure 3.12: **The quality index comprises 6 levels from *A - excellent* to *F - very poor*. The threshold values for achieving a level depend on the respective criterion and must be defined in an application-specific manner.**

In this case, the evaluation level of a forecast is derived from the three aspects described previously in Sections 3.1–3.3:

1. Index of signal control dynamics

2. Index of message integrity and plausibility

3. Index of forecast quality

The index of signal control dynamics is shown as an independent characteristic value and is primarily used to reasonably compare forecasts only between signal groups or intersections with comparable signal control dynamics. The subindices of message integrity and plausibility as well as forecast quality, on the other hand, can be combined with defined weightings to form a quality index that describes the actual forecast quality. In order to be able to represent different use cases, it is also possible to weight certain aspects more heavily or to limit the evaluation to selected aspects. For example, only a quality index for `green` may be of interest for a residual green display, so that red states do not have to be taken into account.

The HBS uses defined forms [9, Part S4.7] to formalise the calculation of the evaluation

levels. These tabular calculation templates are used to calculate indicators and evaluation levels for individual traffic flows and to summarise these results for each arm of a junction. Based on the HBS forms, a similar summary is provided for a junction and corresponding forecast for the overall rating of the forecast quality. Such a tabular formalisation can be implemented for all three aspects of the evaluation concept. Chapter 5 lists an application example for the quality evaluation of the three aspects based on the HBS forms.

4 Evaluation metric

This chapter presents the implementation of the evaluation concept as an evaluation metric. The following sections outline the calculation methods and discuss how the proposed equations are applied to datasets, which were tested during development using sample data. The chapter builds on the concepts and descriptions introduced in chapter 3.

4.1 Signal control dynamics

The dynamic index of a TSC is composed of the dynamic indices of the start and end times of phases, as well as the subindex for phase durations. To determine this, the time gap between the start and end times of the phases are first analysed for each `eventState`. This is done individually for each signal group, whereby the index of a signal group shows the weighted sum of all evaluated `eventStates`. The weighting is based on the proportion of `eventStates` in the observation period. To evaluate an entire junction, the signal control dynamic indices of the individual signal groups are combined at the end as an average value to form a total index. The subindices of the overall signal control dynamics are described individually below.

Subindex of phase start times: The subindex of phase start times describes the randomness of the intervals between phase starts. Clear patterns in these time gaps can indicate the presence of coordination timing plans, which can increase the predictability of the TSC.

For this subindex, all time gaps between consecutive phase starts within the observation period are determined. This allows to compute a probability for each of these intervals, which is then combined to form the entropy H_e:

$$H_e(X) = -\sum_i p_{i,e} \cdot \log_2 p_{i,e} \tag{4.1}$$

Here H_e describes Shannon's information entropy of an `eventStates` e, X the total of all time gaps between the start times $t_{i,e}$ and $p_{i,e}$ declares the probability for a distance $\Delta t_{i,e}$ between two start times. For the subindex D_s of a signal group s, the entropies of the individual `eventStates` are then added up weighted according to their total duration at the signal group:

$$D_s(X) = \sum_e a_{e,s} \cdot H_{e,s}(X) \tag{4.2}$$

Here, $a_{e,s}$ describes the share of the `eventState` e in the total duration of all `eventStates` from signal group s under consideration.

An important note: As the index value D_s is not limited, here the value $\beta = 4$ is assumed as the maximum for normalisation. The assumption is that the maximum value is chosen high enough that the predictability of the TSC no longer changes significantly with even higher values and that these TSCs therefore remain comparable. This results in the normalised index $D_{s,n}$ as follows:

$$D_{s,n} = \frac{1}{\beta} \cdot \min(D_s, \beta) \tag{4.3}$$

Subindex of phases end times: This subindex describes the randomness of the intervals between the phase end times. Similar to the subindex of start times, clear patterns can also indicate the existence of coordination timing plans, which can increase the predictability of TSC.

This subindex is calculated the same way as the subindex for the start times (see equation 4.3), with the difference that $\Delta t_{i,e}$ describes the time gaps between the end times in this case.

Subindex of phase durations: The subindex of phase durations describes the randomness of the observed durations of phases. A great variability in the observed phase durations indicates a more difficult forecasting. In this case, limiting the data to fewer data points for phase durations could make forecasting easier.

For this subindex, the time gaps between the observed phase durations of the `eventStates` are considered. Therefore, the original entropy formula is extended by an additional weighting factor w_e:

$$H_e(X) = -\sum_i w_e(i) \cdot p_{i,e} \cdot \log_2 p_{i,e} \tag{4.4}$$

The weighting factor $w_e(i)$ depends on the spacing to the adjacent observed phase durations, i. e. the gaps to the next smaller and next larger value. Based on the initial data analysis, the following weighting factor was found to work well in practice:

$$w_e(i) = \frac{1}{2} \cdot \log_{10}\left(\Delta^- x_{i,e} + \alpha\right) + \frac{1}{2} \cdot \log_{10}\left(\Delta^+ x_{i,e} + \alpha\right) \tag{4.5}$$

where $\Delta^- x_{i,e} = x_{i,e} - x_{i-1,e}$ and $\Delta^+ x_{i,e} = x_{i+1,e} - x_{i,e}$ represent the backward or forward difference to the next smaller or larger observed phase duration of the `eventStates` e. The addition of the value $\alpha = 9$ within the logarithm function results in a weighting factor of $w_e = 1$ for a minimum time interval of 1s. For larger time gaps, w_e increases accordingly. In principle, however, other implementations of this weighting are also conceivable.

The remaining accumulation of the observed phase durations is carried out analogue to the previously described subindices.

Index of signal control dynamics: Firstly, the dynamics index $I_{D,s}$ of the signal group is calculated from the sum of the normalised subindices for phase start times $D_{\text{start},s,n}$, phase end times $D_{\text{end},s,n}$ and phase durations $D_{\text{interval},s,n}$ with the associated weights w:

$$I_{D,s} = \frac{w_{\text{start},s} \cdot I_{\text{start},s,n} + w_{\text{end},s} \cdot I_{\text{end},s,n} + w_{\text{interval},s} \cdot I_{\text{interval},s,n}}{w_{\text{start},s} + w_{\text{end},s} + w_{\text{interval},s}} \tag{4.6}$$

In summary, the signal control dynamic index of a TSC I_D for a intersection is then the sum

of the subindices for the signal groups normalised to the number of signal groups N_s. It serves as an indicator of the comparability of the forecast of different TSCs:

$$I_D = \frac{1}{N_s} \cdot \sum_s I_{D,s} \tag{4.7}$$

4.2 Message integrity and plausibility

To check the integrity and plausibility of the message data, the messages must be analysed independently of the actual signal timings. It is examined whether the messages are consistent and fulfil the requirements of the C-Roads profile as well as the basic constraints and requirement of traffic engineering.

Availability: The availability of the forecast refers to the proportion of SPATEM messages that provide forecasts at all. A high level of availability is the basic prerequisite for ensuring that forecasts can be used reliably in applications.

Formally, the availability results from the percentage of SPATEM messages in which the necessary data fields `minEndTime`, `maxEndTime`, `likelyTime` and `confidence` are filled according to the profile specifications. In addition, the `confidence` field of a time interval class must have a value > 0. Time interval classes with a value of zero are evaluated as unavailability of the forecast according to the C-Roads profile.

Plausibility: Certain requirements apply to the variables `minEndTime`, `likelyTime` and `maxEndTime`. The `minEndTime` may only shift further into the future in successive messages, the `maxEndTime` may not shift further into the future and the `likelyTime` must always lie between `minEndTime` and `maxEndTime`. Accordingly, the following criteria are evaluated separately

- The `minEndTime` of the current message is less than or equal to the `minEndTime` of the subsequent message of the same `eventState` e.

$$minEndTime_{t,e} \leq minEndTime_{t+1,e} \tag{4.8}$$

- The `maxEndTime` of the current message is greater than or equal to the `maxEndTime` of the subsequent message of the same `eventState` e.

$$maxEndTime_{t,e} \geq maxEndTime_{t+1,e} \tag{4.9}$$

- The `likelyTime` of a forecast lies between `minEndTime` and `maxEndTime` of this forecast.

$$minEndTime_t \leq likelyTime_t \leq maxEndTime_t \tag{4.10}$$

Based on this, the percentage of SPATEM messages with an available forecast that fulfil the respective criterion can be determined.

Check for static *eventStates*: Non-dynamic `eventStates` have a fixed phase duration and therefore fixed, immutable phase change. The `likelyTime` specifies this phase change. Therefore, the invariability of phase changes is checked by examining consecutive pairs of messages for a match according to the following criterion:

$$likelyTime_{t,e} = likelyTime_{t+1,e} \tag{4.11}$$

If this criterion is met, the corresponding messages are considered profile-compliant. The proportion of profile-compliant messages results in a percentage that is included in the subindex. Specifically, the `eventStates protected clearance (amber)`, `permissive clearance (amber)` and `preparation (red-amber)` are considered.

Index for message integrity and plausibility: The index for message integrity and plausibility F_s of a signal group s is defined as the normalised weighted sum of the subindices $F_{x,s}$:

$$F_s = \frac{1}{\sum_x w_x} \cdot \sum_x w_x \cdot F_{x,s} \tag{4.12}$$

The various subindices $F_{x,s}$ of a signal group s indexed by $x = 1, \ldots, 7$ are as follows:

"1" Percentage of availability,

"2" Percentage of profile-compliant consecutive `minEndTimes`,

"3" Percentage of consecutive `maxEndTimes` in accordance with the profile,

"4" Percentage of profile-compliant `minEndTime-likelyTime-maxEndTime-` relationships,

"5" Percentage of profile-compliant predictions for `protected clearance (amber)`,

"6" Percentage of profile-compliant predictions for `permissive clearance (amber)`,

"7" Percentage of profile-compliant predictions for `preparation (red-amber)`.

Total index The total index of a junction for message integrity and plausibility F is formed from the sum of the subindices F_s of all signal groups and normalised to the number of signal groups N_s:

$$F = \frac{1}{N_s} \cdot \sum_s F_s \tag{4.13}$$

4.3 Forecast quality

In order to evaluate the quality of forecasts, the interaction of forecast horizon, forecast accuracy and the `confidence` for the subindex of the forecast quality is analysed. For this purpose, the maximum C-Roads-compliant `likelyTime` is determined, the associated `confidence` is evaluated and the resulting score is weighted depending on the forecast horizon. Finally, the subindices for `minEndTime` and `maxEndTime` are determined and a total index is calculated for a junction from the normalised sum of all signal groups.

Determination of the maximum, profile-compliant *likelyTime*: The following condition must be met for a single forecast in a SPATEM message to be correct:

$$L_{p,T} - c_{p,T,\text{C-Roads}} \leq T \leq L_{p,T} + c_{p,T,\text{C-Roads}} \tag{4.14}$$

Here T denotes the phase change time of a signal group, $L_{p,T}$ the `likelyTime` issued by a specific forecast p for this time and $c_{p,T,\text{C-Roads}}$ the associated `confidence` according to the C-Roads profile.

This condition is checked for each forecast in the observation period so that the average accuracy $G_{L,e}$ of the corresponding forecasts can be determined for each `eventState` e and for each associated relative `likelyTime` L. Since the C-Roads profile stipulates that an overall forecast accuracy of 95% should be achieved, this condition is checked separately for each accuracy $G_{L,e}$. As a result, the maximum relative `likelyTime` $L_{\text{max},e}$ can be determined for each `eventState` e, for which the following condition applies:

$$\forall i \in \{1, \ldots, L_{\text{max},e}\} : \; G_{i,e} \geq 0.95 \tag{4.15}$$

This condition states that for all relative `likelyTimes` with index i in the range from 1 to $L_{\text{max},e}$ the average accuracy $G_{L,e}$ is at least 95%. $L_{\text{max},e}$ accordingly describes the maximum forecast horizon up to which the forecasts of an `eventState` e are C-Roads-compliant overall.

Evaluation of the *confidence*: For each accuracy $G_{L,e}$, a corresponding average `confidence` time interval class $c_{L,e}$ can also be determined. The associated time interval $c_{L,e,\text{C-Roads}}$ describes the width of the average forecast time window for the respective `likelyTime`. Ideally, $c_{L,e}$ has the value 15, which corresponds to a forecast time window of width zero according to table 2.1 in the C-Roads profile. From this, the evaluation $C_{L,e}$ is defined as:

$$C_{L,e} = \frac{1}{\log_\gamma(c_{L,e,\text{C-Roads}} + \gamma)} \tag{4.16}$$

The parameter γ was set to $\gamma = 4$ following empirical tuning on the sample datasets. The denominator $\log_\gamma(c_{L,e,\text{C-Roads}} + \gamma)$ increases logarithmically with the width of the forecast time window, i. e. $C_{L,e}$ becomes correspondingly smaller. At maximum `confidence` $C_{L,e} = 1$, in the worst case $C_{L,e} \approx 0.48$ (approximately $1/2$).

The maximum, profile-compliant `likelyTime` $L_{\max,e}$ has further influence on the evaluation. Figure 4.1 illustrates this influence on the `confidence` evaluation $C_{L,e}$ using two examples and taking $L_{\max,e}$ into account. In Figure 4.1 A, all scores of the average `confidence` time interval class $c_{L,e}$ are included. In contrast, Figure 4.1 B shows a case in which the ratings $C_{L,e}$ are disregarded for all `likelyTimes` whose associated `likelyTime` is greater than the maximum, profile-compliant `likelyTime`. In this context, "disregarded" means that the minimum rating of zero is assumed for these values. This means that only evaluations of `confidence` whose relative `likelyTimes` also fulfil the requirements of the C-Roads profile are included in the forecast quality.

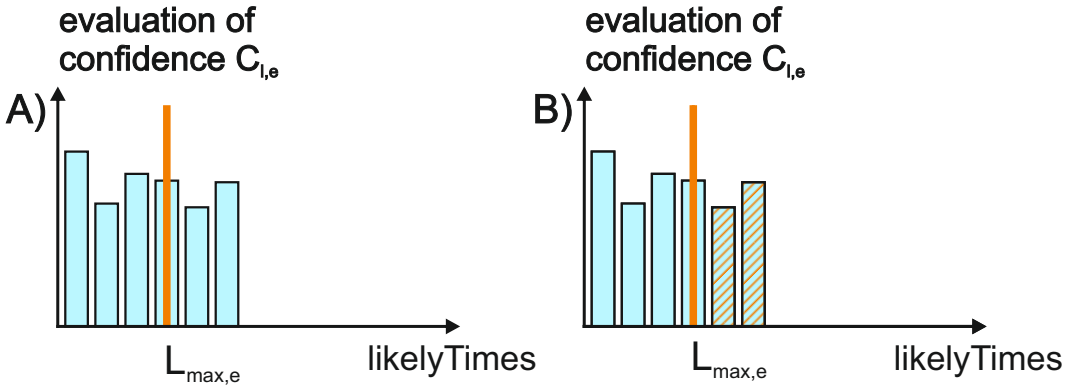


Figure 4.1: **Influence of the maximum, profile-compliant *likelyTime* $L_{\max,e}$ on the evaluation of the *confidence*; A: Shows all evaluations of the *confidence*. B: All scores that belong to a greater *likelyTime* than $L_{\max,e}$ are disregarded.**

Weighting of the *confidence* rating: Short-horizon forecasts are expected to have a higher `confidence` than forecasts further ahead. For this purpose, a weighting function is introduced that assumes its maximum at a relative `likelyTime` of 1 s and decreases linearly to the value zero up to the maximum forecast horizon $T_{P,\max}$. Figure 4.2 illustrates the implementation of such a weighting, taking into account the ratio $T_{P,\max}$ to $T_{S,\max,e}$. The maximum forecast horizon $T_{P,\max}$ describes the time limit up to which forecasts are evaluated; $T_{S,\max,e}$ is the maximum observed phase duration. For normalisation, the weighting function is scaled so that the sum of all partial weights equals 1; accordingly, the normalisation area shown in Figure 4.2 A has an area of 1. However, if $T_{P,\max}$ is greater than the maximum observed phase duration $T_{S,\max,e}$, only the sum of the weights up to this maximum observed phase duration $T_{S,\max,e}$ is normalised (see Figure 4.2 B). In this case, the weights for higher `likelyTimes` are not included in the normalisation, so that lower `likelyTimes` contribute more to the overall weighting.

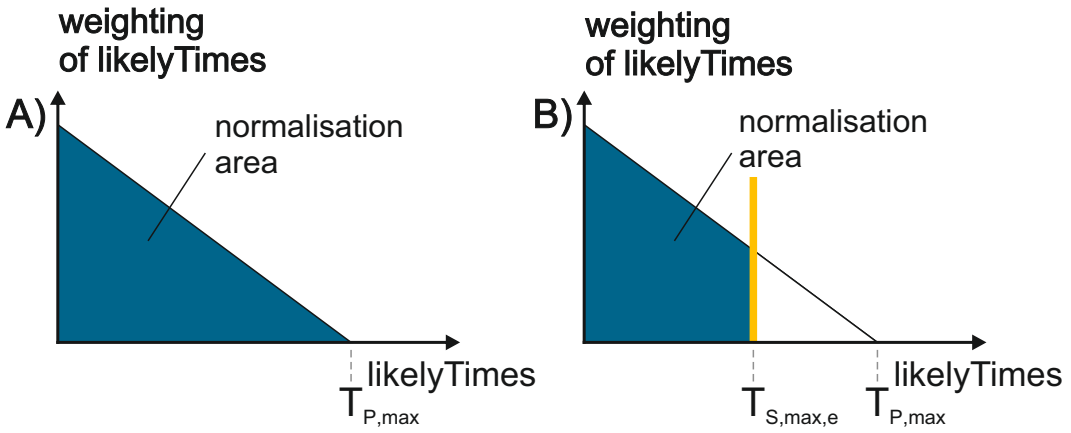


Figure 4.2: **Weighting of the *likelyTimes* that contribute to the *confidence* score. A: All weights up to the defined maximum forecast horizon $T_{P,\max}$ are evaluated. B: All weights that lie above the maximum observed phase duration $T_{S,\max,e}$ are excluded from the normalisation.**

Subindex of the forecast: The forecast subindex describes the evaluation of the interaction between the forecast horizon, forecast accuracy and `confidence`. The evaluation of the `confidence` serves as the basis for the index. The forecast accuracy contributes to the maximum, C-Roads-compliant `likelyTime` $L_{\max,e}$. The forecast horizon and its weighting is also represented by $L_{\max,e}$, which is explained below.

For this subindex $I_{P,e,s}$ of an `eventState` e and a signal group s, the evaluation of the `confidence` is combined with the weighting. Figure 4.3 illustrates this approach. The ratings of the `confidence` $C_{L,e,s}$ are multiplied by the associated weightings $w_{L,e,s}$ of the relative `likelyTimes` L and accumulated:

$$I_{P,e,s} = \sum_{L} w_{L,e,s} \cdot C_{L,e,s} \tag{4.17}$$

However, please note the following: All evaluations $C_{L,e}$ with `likelyTimes` greater than the maximum, profile-compliant `likelyTime` $L_{\max,e}$ are disregarded and assumed to be zero (see hatching in Figure 4.3). In this case, normalisation is performed up to the smaller value of the maximum observed phase duration $T_{S,\max,e}$ and the maximum forecast horizon $T_{P,\max}$, regardless of how large $L_{\max,e}$ is. This type of evaluation penalises the case where the accuracy of the forecast is not sufficiently within the value range up to 95% required by the C-Roads profile.

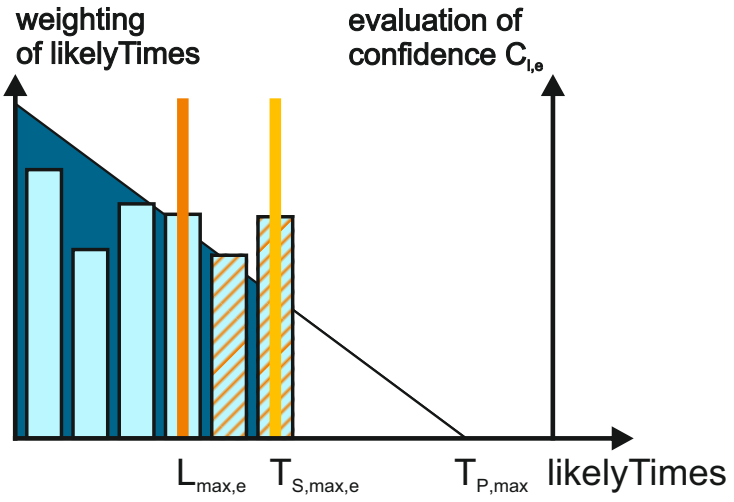


Figure 4.3: **Illustration of the interaction between normalisation, *confidence* evaluation and the maximum, profile-compliant *likelyTime* $L_{\max,e}$; the normalisation area is reduced because $T_{P,\max} < T_{S,\max,e}$. In addition, two *confidence* scores are disregarded because the maximum, profile-compliant *likelyTime* $L_{\max,e}$ is too low.**

The forecast subindex $I_{P,s}$ for a signal group s is obtained by aggregating the values across all `eventStates` e in the same manner as for the subindex of signal control dynamics. The aggregation is weighted by $a_{e,s}$, proportional to the share of the `eventStates` e in the total duration of the signal group:

$$I_{P,s} = \sum_{e} a_{e,s} \cdot I_{P,e,s} \tag{4.18}$$

The maximum score of the forecast subindex for an `eventState` e is obtained when the forecast is C-Roads-compliant up to the maximum forecast horizon or the maximum observed phase duration, and when `confidence` remains at its maximum throughout. A lower `confidence` has a strong negative impact on the forecast quality, especially in the near future, as does non-compliance with the C-Roads profile from a low relative `likelyTime`. Conversely, inaccuracies in the more distant future have significantly less negative effects on the forecast quality.

Subindices for *minEndTime* and *maxEndTime*: The subindices I_{minE} and I_{maxE} each record the percentage of forecasts for which the actual phase change T is consistent with the predicted `minEndTime` and `maxEndTime`. The predicted `minEndTime` $t_{\text{min},p,T}$ of a forecast p for a specific phase change T must therefore apply:

$$t_{\text{min},p,T} \leq T \tag{4.19}$$

Accordingly, the predicted `maxEndTime` $t_{\text{max},p,T}$ of a forecast p must apply to a specific phase change T:

$$t_{\text{max},p,T} \geq T \tag{4.20}$$

The subindices I_{minE} and I_{maxE} therefore correspond to the percentage of forecasts for which the respective condition is fulfilled.

Total index of the forecast quality: Firstly, the forecast quality Q_s of a signal group s is calculated from the subindex of the forecast $I_{P,s}$ and the subindices $I_{\text{minE},s}$ and $I_{\text{maxE},s}$ for `minEndTime` and `maxEndTime` with the associated weights w:

$$Q_s = \frac{w_{P,s} \cdot I_{P,s} + w_{\text{minE},s} \cdot I_{\text{minE},s} + w_{\text{maxE},s} \cdot I_{\text{maxE},s}}{w_{P,s} + w_{\text{minE},s} + w_{\text{maxE},s}} \tag{4.21}$$

Lastly, the overall forecast quality I_P for a junction is calculated as the sum of the forecast qualities Q_s of the signal groups s normalised to the number of signal groups N_s

$$I_P = \frac{1}{N_s} \sum_s Q_s \tag{4.22}$$

Subsequently, the values of the subindices calculated in the sections 4.1–4.3 can be summarised in a table according to the evaluation concept and presented as evaluation levels.

4.4 Overview of the evaluation metric

To conclude this chapter, a general overview of the evaluation concept described in section 3 and the evaluation metric in section 4 is provided.

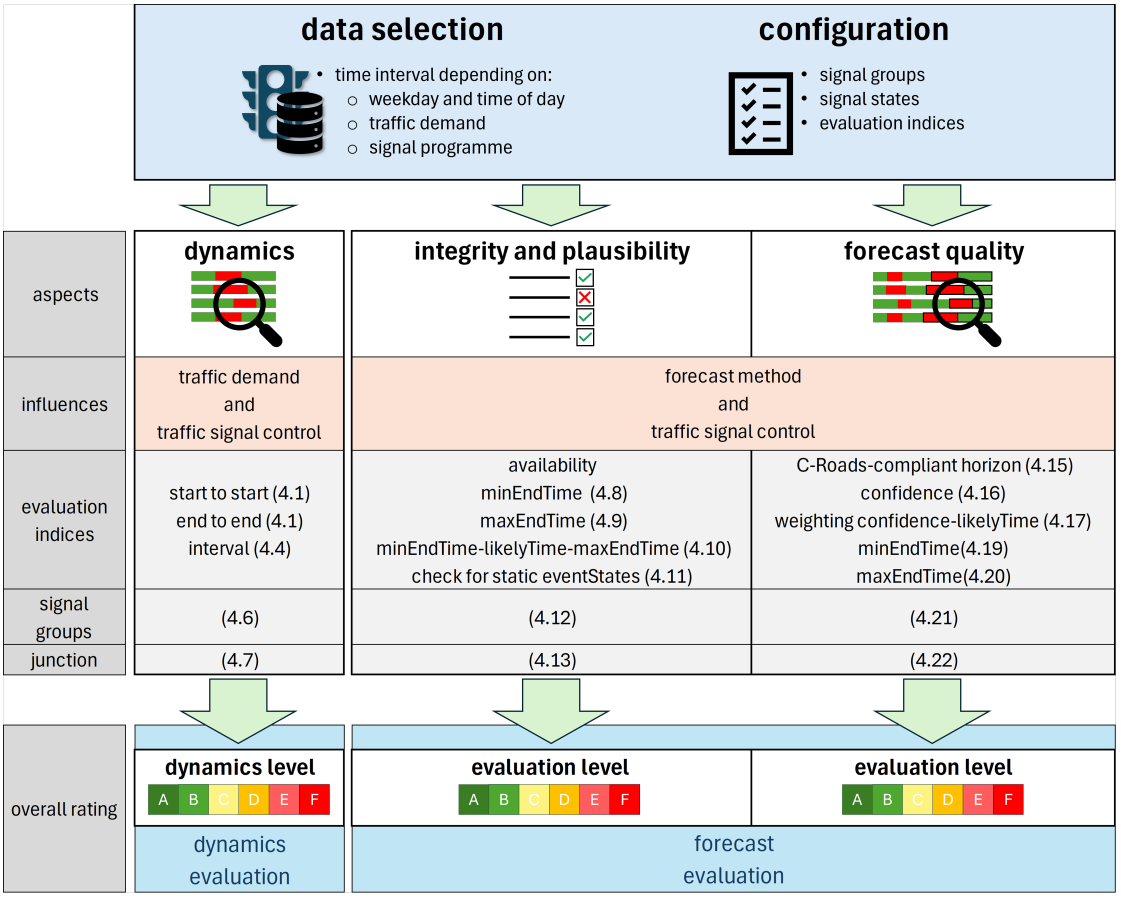


Figure 4.4: **Overview of the evaluation concept for determining an overall rating for a junction. The information in brackets (4.1 – 4.22) refers to the corresponding equations, which are listed in chapter 4.**

Figure 4.4 summarises the basic requirements, key aspects and processing steps graphically, and assigns the individual evaluation indices to their respective evaluation criteria. Together, these indices yield an overall rating. Based on the structure presented, the determination of the evaluation metric begins with the selection of suitable **data sources** and a proper **configuration**. First, the relevant time periods, traffic demand and traffic signal programmes

are chosen for the analysis, as well as which signal groups, signal states and evaluation indices are to be taken into account for the respective use case. Based on these choices, the evaluation metric is configured to meet the analysis objectives while ensuring the desired comparability with other junctions.

After this initial preparation, the **signal control dynamic** of the TSC under consideration is determined. For this purpose, the start and end times as well as the phase durations are analysed and combined to an index of signal control dynamics, from which a dynamic level for the junction can be derived. This characteristic value describes how much the TSC changes over time and how challenging it is to make a reliable forecast.

Separately, the **integrity and plausibility** of the available forecast data must be checked. In particular, the focus is on the availability and profile conformity of the forecasts as well as the temporal consistency of `minEndTime`, `likelyTime` and `maxEndTime`. The results of these checks are summarised in a subindex. Most importantly, the **forecast quality** is evaluated by comparing the profile-compliant `likelyTimes` and their associated `confidence` with the actually observed phase change, using a weighting across the forecast horizon. This evaluation results in a subindex of the forecast quality, which describes the effective quality of the forecast content.

Lastly, the subindices are combined to form an evaluation of the signal control dynamics as well as a forecast evaluation, from which a compact overall rating can be derived in the form of evaluation levels.

5 Application

The quality evaluation metric makes it possible to analyse forecast methods and data from different perspectives for a range of use cases. For example, it enables a comparative evaluation of available forecasting methods for potential deployment at a junction. It also allows the forecast quality of the same method to be analysed across multiple junctions. In addition, it can be used to assess the suitability of transmitted forecast data for the planned implementation of C-ITS services at a junction. The range of possible applications is broad, but reliable forecast evaluation always depends on an appropriate data basis. The selected dataset should represent a typical traffic demand situation and cover a sufficiently long time period. Furthermore, the evaluation metric must be configured in accordance with the analysis objective. Only then can a precise and valuable quality evaluation be achieved.

In the following, the usage of the developed quality evaluation metric is demonstrated using a fictitious application example. In addition, general guidance is provided that may be helpful for applying the evaluation metric. This includes, for example, recommendations for selecting the analysis period, configuring the metric, and defining weightings and threshold values for the evaluation levels.

5.1 Application example

The application example illustrates how the configuration and the resulting forecast quality evaluation are carried out. The objective of this example is to use the quality evaluation metric to determine, for a fictitious actuated TSC according to the layout shown in Figure 5.1, whether the deployment of a GLOSA application could be reliably implemented for motorised traffic in the east–west direction. It is assumed that an appropriate data basis is available, i. e., SPATEM messages covering a sufficiently long period that is representative in terms of traffic demand. The focus is less on the specific evaluation levels obtained and more on the procedure and practical use of the evaluation metric. Against this background, the fictitious junction has been deliberately simplified, and not all aspects are evaluated comprehensively.

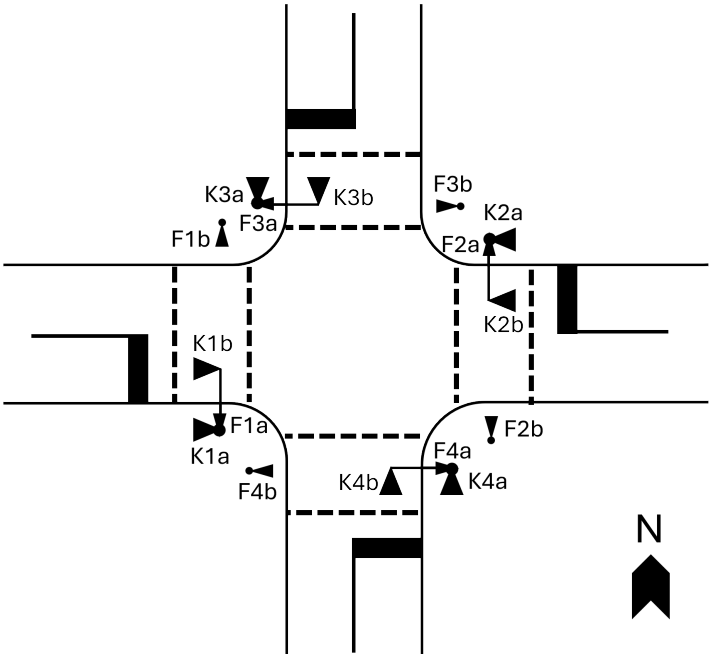


Figure 5.1: **Site plan for the application example: Testing the possible use of a GLOSA application at a symmetrical junction.**

5.1.1 Example configuration

Before an evaluation can be carried out, the metric must first be configured. This means specifying which characteristics and properties the junction has and which indicators within the individual subindices are included in the evaluation. Figure 5.2 summarises the corresponding configuration for the example in tabular form. A value of "1" indicates that a feature or property is selected, whereas "0" indicates that it is not selected. The key configuration parameters are as follows:

- The vehicle signal groups K1 and K2 are to be analysed for the potential deployment of a GLOSA service. These groups are therefore set to "1"; the remaining signal groups are not considered and not selected, as no GLOSA application is to be implemented for them. This is fully shown in Figure 5.2 in the section *signal group*.

- For the GLOSA application in this example, the state `permissive green` is not relevant, since the application is intended only to support west–east movements. When

this state is active, straight-moving GLOSA users have right of way. Therefore, this state is deselected in the configuration in the section *state* in Figure 5.2.

- For the evaluation of the three subindices, signal control dynamics, message plausibility and integrity, and forecast quality, a hypothetical standard configuration is used. All features are selected and the maximum forecast horizon to be analysed is set to 15 s. This is specified in the corresponding sections of Figure 5.2.

signal group	selection
K1 (car)	1
K2 (car)	1
K3 (car)	0
K4 (car)	0
F1 (pedestrians)	0
F2 (pedestrians)	0
F3 (pedestrians)	0
F4 (pedestrians)	0

state	selection
protected green	1
permissive green	0
red	1

subindex signal dynamics	selection
duration	1
start to start	1
end to end	1

subindex integrity/plausibility	selection
availability	1
minEndTimes	1
maxEndTimes	1
minEndTime–likelyTime–maxEndTime	1
protected clearance (amber)	1
permissive clearance (amber)	1
preparation (red-amber)	1

subindex forecast quality	selection
confidence likelyTime	1
minEndTimes	1
maxEndTimes	1
maximum forecast horizon [s]	15

Figure 5.2: **Example configuration for the fictitious junction of the application example.**

5.1.2 Example evaluation of the signal control dynamics

Figure 5.3 shows an example evaluation of the signal control dynamics based on the form and calculation sheet of the evaluation levels of the traffic flow of the HBS. The individual columns of the table should to be interpreted as follows.

signal group	selection	state [event state]	selection	percentage of observation period	subindex	selection	subindex value	state value	state level	signal group value $[I_{D,s}]$	signal group level	total value $[I_D]$	total level
0	1	2	3	4	5	6	7	8	9	10	11	12	13
K1	1	protected green	1	40%	duration	1	0,8	0,9	B	0,7	C	0,7	C
					start to start	1	0,9						
					end to end	1	1,0						
		red	1	55%	duration	1	0,5	0,4	E				
					start to start	1	0,3						
					end to end	1	0,4						
		permissive green	0	-	-	-	-	-	-				
K2	1	protected green	1	60%	duration	1	1,0	1,0	A	0,8	B		
					start to start	1	0,9						
					end to end	1	1,0						
		red	1	35%	duration	1	0,6	0,7	C				
					start to start	1	0,7						
					end to end	1	0,7						
		permissive green	0	-	-	-	-	-	-				
K3	0	-	-	-	-	-	-	-	-	-	-		
K4	0	-	-	-	-	-	-	-	-	-	-		
F1	0	-	-	-	-	-	-	-	-	-	-		
F2	0	-	-	-	-	-	-	-	-	-	-		
F3	0	-	-	-	-	-	-	-	-	-	-		
F4	0	-	-	-	-	-	-	-	-	-	-		

Figure 5.3: **Tabular representation of an overall rating from subindices of signal control dynamics.**

The signal groups are shown in column 0; column 1 shows that the signal groups K3, K4 and all pedestrian signal groups are not selected and are therefore not included in the evaluation. For the states in column 2, only the `protected green` and `red` are included in the evaluation; the state `permissive green` is not relevant for the use case and was deselected accordingly in column 3.

For the values of subindices of the signal control dynamics, fictitious values in the range [0; 1]

were assumed in column 7. In principle, these subindices are determined using equations 4.1 and 4.4. This value describes the entropy and predictability of the respective analysed subindex. The values of the individual subindices (column 7) are first aggregated within each state in column 8 as a normalised sum. Thus, column 8 represents the signal control dynamics of each individual state.

Subsequently, the state values are weighted using equation 4.6 according to the proportion of the state in the total duration of a signal group and combined into a single value for each signal group in column 10. Therefore, column 10 contains a signal control dynamics value for each signal group. The values of the individual signal groups from column 10 are finally aggregated according to equation 4.7 into an overall value of signal control dynamics in column 12. The evaluation levels in columns 9, 11, and 13 are then determined based on the fictitious assignment of threshold values.

Figure 5.4 illustrates such a specification. The signal control dynamics of the example junction achieve an overall rating of level **C**.

dynamics level	rating	value
A	fully flexible	≤ 1,0
B	very flexible	≤ 0,9
C	flexible	≤ 0,7
D	partially static	≤ 0,5
E	static	≤ 0,3
F	very static	≤ 0,1

Figure 5.4: **Exemplary threshold value assignment to the signal control dynamics levels.**

5.1.3 Example evaluation of integrity and plausibility

Figure 5.5 shows an evaluation of the message integrity and plausibility. In this example, the calculation is also based on the HBS form and calculation sheet. According to the configuration (see section 5.1.1), the signal groups and subindices to be evaluated are selected in columns 1 and 3.

signal group	selection	subindex	selection	subindex value	signal group value $[F_s]$	signal group level	total value $[F]$	total level
0	1	2	3	4	5	6	7	8
K1	1	availability	1	0,9	0,9	B	0,9	B
		minEndTimes	1	0,8				
		maxEndTimes	1	0,7				
		minEndTime–likelyTime–maxEndTime	1	1,0				
		protected clearance (amber)	1	1,0				
		permissive clearance (amber)	1	1,0				
		preparation (red-amber)	1	1,0				
K2	1	availability	1	0,8	0,9	B		
		minEndTimes	1	0,9				
		maxEndTimes	1	0,8				
		minEndTime–likelyTime–maxEndTime	1	1,0				
		protected clearance (amber)	1	1,0				
		permissive clearance (amber)	1	1,0				
		preparation (red-amber)	1	1,0				
K3	0	-	-	-	-	-		
K4	0	-	-	-	-	-		
F1	0	-	-	-	-	-		
F2	0	-	-	-	-	-		
F3	0	-	-	-	-	-		
F4	0	-	-	-	-	-		

Figure 5.5: **Tabular presentation of an overall rating from the subindices of integrity and plausibility.**

For the subindices of message integrity and plausibility (column 2), fictitious values in the interval $[0; 1]$ were assumed as an example (column 4). The values of the subindices are first merged within a signal group using the equation 4.12 in column 5. This means that column 5 contains a value for integrity and plausibility for each signal group. The equation 4.13 is then used to form the sum of the values of all signal groups and normalised to the number of signal groups, so that an overall rating for the junction is obtained in column 7. The evaluation levels in columns 6 and 8 then result from the assignment to fictitious threshold values, as shown for illustration in Figure 5.6. The message integrity and plausibility of the example junction achieves the evaluation level **B** in the overall rating.

evaluation level	rating	value
A	excellent	≤ 1,0
B	very good	≤ 0,9
C	good	≤ 0,7
D	average	≤ 0,5
E	poor	≤ 0,3
F	very poor	≤ 0,1

Figure 5.6: **Exemplary threshold value assignment to the evaluation levels.**

5.1.4 Example evaluation of the forecast quality

Figure 5.7 shows an example of the evaluation of the forecast quality based on the HBS form and calculation sheet. In columns 1, 3 and 6, the signal groups, states and subindices to be evaluated are selected according to the configuration from section 5.1.1.

For the subindices `minEndTime` and `maxEndTime`, fictitious values in the interval [0; 1] are assumed in column 8 as an example, which are generally determined using the equations 4.19 and 4.20. The value of the subindex `confidence-likelyTime` in column 8 is generally determined using equation 4.18 and taking into account the maximum forecast horizon from column 7. Fictitious values in the interval [0; 1] were also defined for these subindices. The values of the individual subindices (column 8) are within each state in column 9 as a normalised sum. The values of the states (column 9) are then combined within a signal group with the equation 4.21 in column 11. This means that column 11 contains a value for the forecast quality for each signal group. Eventually, equation 4.22 is used to form a sum of the values of all signal groups from column 11 and normalised to the number of signal groups in order to finally obtain an overall rating of the forecast quality for the junction in column 13.

The evaluation levels in columns 10, 12 and 14 result from the fictitious threshold values according to Figure 5.6. The example junction achieves an overall rating of evaluation level **C** in terms of forecast quality.

signal group	selection	state [event state]	selection	percentage of observation period	subindex	selection	maximum forecast horizon [s]	subindex value	state value	state level	signal group value $[Q_s]$	signal group level	total value $[I_P]$	total level
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
K1	1	protected green	1	40%	confidence likelyTime	1	15	0,4	0,7	C	0,7	C	0,8	C
					minEndTimes	1	-	0,9						
					maxEndTimes	1	-	0,7						
		red	1	55%	confidence likelyTime	1	15	0,5	0,7	C				
					minEndTimes	1	-	0,8						
					maxEndTimes	1	-	0,9						
		permissive green	0	-	-	-	-	-	-	-				
K2	1	protected green	1	60%	confidence likelyTime	1	15	0,6	0,7	C	0,8	C		
					minEndTimes	1	-	0,7						
					maxEndTimes	1	-	0,9						
		red	1	35%	confidence likelyTime	1	15	0,8	0,9	C				
					minEndTimes	1	-	0,9						
					maxEndTimes	1	-	0,9						
		permissive green	0	-	-	-	-	-	-	-				
K3	0	-	-	-	-	-	-	-	-	-	-	-		
K4	0	-	-	-	-	-	-	-	-	-	-	-		
F1	0	-	-	-	-	-	-	-	-	-	-	-		
F2	0	-	-	-	-	-	-	-	-	-	-	-		
F3	0	-	-	-	-	-	-	-	-	-	-	-		
F4	0	-	-	-	-	-	-	-	-	-	-	-		

Figure 5.7: **Tabular representation of an overall rating from the subindices of forecast quality.**

5.1.5 Example of the overall rating

The three aspects analysed for the possible deployment of a GLOSA application are then combined into an overall rating in accordance with Figure 5.8. Within this overall rating, the dynamics of the TSC are categorised as "flexible" with the level **C**, the message integrity and plausibility with the evaluation level **B** as "very good" and the forecast quality with the evaluation level **C** as "good". This yields a comprehensible and comparable evaluation result. Based on this, it could be decided whether a GLOSA application is feasible at this junction under the existing constraints.

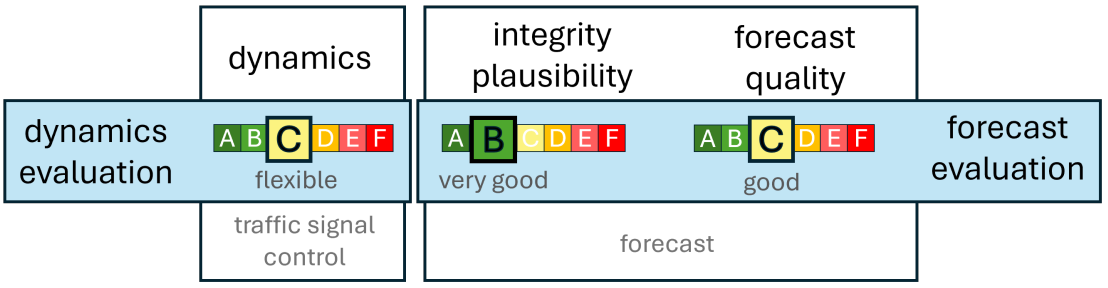


Figure 5.8: **Presentation of the overall rating of the application example.**

5.2 Configuration and application notes

Before a quality evaluation metric can be applied, an appropriate data basis must be available that represents a characteristic traffic demand and covers a sufficiently long analysis period. A demand-oriented configuration must be carried out for all sub-indices of the metric, and suitable threshold values for the evaluation levels should be used. The purpose of these specifications is to ensure adequate comparability when analysing different forecasts. Accordingly, some guidance is provided based on experience with the real-world data used, which should be taken into account when preparing, configuring, and applying the quality evaluation metric.

5.2.1 Configuration

Selection and weighting: The primary purpose of the configurability of the evaluation metric is to specifically reflect certain requirements for a forecast while maintaining compara-

bility with other TSCs. As described in the previous application example in chapter 5.1, one possible use case could be to evaluate and compare the forecast only for motorised traffic. In this case, all pedestrian and bicycle signal groups can be excluded from the evaluation. Such an approach can be applied to other application scenarios and junctions as well. However, applying different weightings, for example between main and minor directions, is significantly more difficult to transfer while preserving comparability and should be avoided if possible. In principle, it is also possible to analyse only individual traffic flows or signal groups and their forecasts.

Similar considerations apply to the inclusion and exclusion of `eventStates`. For a Time-to-Green application, it may be sensible, for instance, to focus on forecasting the end of red phases and not consider the other `eventStates`. Furthermore, it should be noted that a finer weighting, which does not fully exclude other `eventStates`, may lead to unexpected effects when comparing different TSCs and could thus compromise comparability.

Comparability: For a suitable comparison of two forecasts, the associated signal control dynamics must be fundamentally similar regardless of the analysis period. The analysis period itself can have a strong impact on the signal control dynamics and its predictability. Therefore, the selected data basis should cover comparable observation periods, configurations should be set as similar as possible and similar threshold values should be defined for the scaling of the dynamic and evaluation levels.

5.2.2 Analysis period from the perspective of practitioner and end users

Selection of the analysis period: To compare two forecasts, the corresponding signal control dynamics must generally be similar, independent of the analysis period. At the same time, the analysis period can strongly influence signal control dynamics and, therefore, predictability. The selected dataset should cover comparable observation periods. Configurations should be set as consistently as possible, and similar threshold values for scaling the signal control dynamics and evaluation levels should be defined properly.

Practitioners versus end users: From a practitioner's perspective, it can be useful to analyse different traffic signal programmes separately, since typically different forecasts apply to each programme. In a research context in particular, a forecast may only need to perform well under specific conditions for a research application, e. g., during low traffic demand in a particular programme. In such cases, it may be useful to extract the corresponding

time periods and analyse them either individually or as a group. Furthermore, the SPATEM message type used as a data basis does not provide a means to assess traffic demand. If the forecast of signal timings is to be analysed with respect to a certain traffic demand, the practitioners must select the appropriate analysis periods themselves.

From the perspective of end users, however, features such as the degree of traffic actuation of a TSC, the chosen programme, or the current traffic demand are not relevant for the subjectively perceived quality of a forecast. End users expect a reliable forecast whenever they pass a signalised, forecast-capable junction.

6 Conclusion

As part of the project "Quality of forecasts for signal timings at traffic lights", a concept for the evaluation of forecast quality was developed. Several tasks were carried out, which are described in the individual chapters of this report. Chapter 1 first introduces the topic, outlining the challenges and the objectives of the project. Chapter 2 summarises the technical background, the results of expert interviews and a requirements analyses, as well as key requirements for a quality evaluation concept. Chapter 3 describes the developed evaluation concept with its individual subindices, while chapter 4 translates the concept into a concrete evaluation metric based on specific calculation steps. Chapter 5 complements these explanations with a practical example of the concept applied to a fictitious junction, and lastly offers a configuration and application guidance derived from working with real-world forecast data.

The concept comprises three central evaluation steps: (i) a signal control dynamics index that captures the predictability of TSCs and enables the comparability of evaluations across different junctions and signal control algorithms, (ii) a separate evaluation of the integrity and plausibility of the available forecast data, and (iii) a substantive evaluation of the forecast quality based on the analysis of signal timing data. An overall rating can be derived from these three individual evaluations.

A key outcome of the concept development is that, based on the developed, machine-readable metric, an automated evaluation of forecast quality would, in principle, be possible. The developed evaluation concept provides the foundation for this. At the same time, it has become clear that configuring the metric still requires a certain level of domain expertise and expert knowledge, for example to judge whether forecasts of TSCs with differing signal control dynamics can be compared at all. In designing the concept, numerous, in some cases unfavourable, characteristics observed in real-world forecast data were taken into account, which has increased the overall complexity of the evaluation metric. Some implementations within the metric are based on justified assumptions and could be further refined in the future if required. Examples include the consideration of specific signal control characteristics and correlations between signal groups in the dynamics analysis. However, this would have to be based on a data foundation extending beyond the SPATEM messages utilised in this project. A wide range of extensions is conceivable, which could be integrated

modularly into the developed evaluation metric. Overall, the experience gained in the application of the metric provides an initial insight into the practical suitability of the metric, while also highlighting its further development potential.

It remains an open question how the effort currently required for manual configuration of the metric can be reduced in the future or improved through suitable standard configurations and support tools. This applies in particular to the appropriate selection and specification of analysis characteristics. Further research is needed to test the metric using representative real-world datasets in different scenarios, in order to derive robust standard parameter sets and enable more systematic comparisons across different junctions. Ideally, typical and recurring use cases could thus be defined, such as testing an actuated TSC for suitability for a GLOSA application. The range of possible use cases is broad. However, the availability of corresponding standard parameter sets would considerably simplify the application of the metric for such use cases.

In this context, defining appropriate threshold values for the grade-like classification of signal control dynamics and forecast quality is of further interest. The concept provides for linking quality levels with specific threshold values. The qualitative interpretation of a level must therefore be translated into a nominal value. Depending on the application and use case, combinations of level interpretation and corresponding threshold value may vary. For establishing robust links between level and threshold, incorporating practical experience from end users and practitioners will therefore be of great importance in the future.

Against this background, the evaluation metric should be tested more extensively in practical applications and further developed in collaboration with different stakeholders. In this way, the metric could become established as a reliable tool for evaluating the forecast of signal timings and be applied in practice as a common quality standard.

References

[1] R. Hoyer and L. Schneegans, "Prognose von Schaltzeiten verkehrsabhängiger Lichtsignalanlagen in Deutschland – Herausforderungen und Ansätze," *at - Automatisierungstechnik*, vol. 69, pp. 931–939, Nov. 2021. DOI: 10.1515/auto-2021-0097

[2] M. Gay et al., *Applicability of V2X for transit signal priority / Nutzung der C2X-basierten ÖV-Priorisierung an signalisierten Knotenpunkten* (Berichte der Bundesanstalt für Straßen- und Verkehrswesen, Reihe V: Verkehrstechnik 353). Bremen: Fachverlag NW in der Carl Ed. Schünemann KG, 2022.

[3] *Hinweise für den Technologiewechsel C-ITS an Lichtsignalanlagen* (FGSV), 2024th ed. Köln, Germany: FGSV, 2024, vol. 321,4.

[4] SAE International, "V2X Communications Message Set Dictionary," SAE International, SAE Standard J2735_202211, Nov. 2022, Revised November 2022, Issued December 2006. DOI: 10.4271/J2735_202211

[5] ETSI, "Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Facilities layer protocols and communication requirements for infrastructure services," European Telecommunications Standards Institute, Technical Specification TS 103 301 V1.3.1, 2020, Version 1.3.1 (2020-02).

[6] C-Roads Platform, Working Group 2 Technical Aspects, "C-ITS European Handbook for MAPEM and SPATEM Creation," C-Roads Platform, Technical Report Release 2.0.9, 2023, Working Group 2 – Technical Aspects.

[7] C-ROADS Platform, Working Group 2 Technical Aspects, Taskforce 3 Infrastructure Communication, "C-its message profiles," C-ROADS Platform, Technical Report Release 2.2.0, Jul. 2024.

[8] C-ROADS Platform, Working Group 2 Technical Aspects, Taskforce 5 Cross-Testin and Validation, "C-ITS Cross-Border Testing: PCAP Exchange Specification," C-ROADS Platform, Technical Report Release 2.0.3, Apr. 2022.

[9] FGSV, Ed., *Manual for the Design of Road Traffic Facilities / Handbuch für die Bemessung von Straßenverkehrsanlagen: HBS 2015* (FGSV W1 - Wissensdokumente FGSV 299 B), 2015th ed. Köln, Germany: FGSV-Verl., 2015.

[10] T. Otto, I. Partzsch, and M. Klöppel-Gersdorf, "Intersection with highly adaptive traffic lights can still be suitable for c-its service glosa," in *Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems - Volume 1: VEHITS*, INSTICC, SciTePress, 2025, pp. 148–156. DOI: 10.5220/0013354500003941

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x

Glossary

Term	Description
actual phase change	Time at which the actual change in state of a signal group occurs
C-ITS	Cooperative Intelligent Transport System
confidence	time window in which a phase change of a signal group takes place with a probability of 95%
dynamic signal state	signal states with different phase durations
ETSI	European Telecommunications Standards Institute
eventState	state of a signal group in a SPATEM message
forecast	describes the predicted phase change of a signal group
forecast horizon	time interval (in seconds) over which a forecast is made
GLOSA	Green Light Optimal Speed Advisory
likelyTime	most probable time of phase change of a signalGroup; reference time for the confidence
MAPEM	MAP (topology) Extended Message
maxEndTime	latest possible time for a phase change of a signalGroup
maximum forecast horizon	time interval (in seconds) up to which a forecast is evaluated
minEndTime	earliest possible time for a phase change of a signalGroup
non-dynamic signal state	signal state with a fixed phase duration
signalGroup	ID of a signal group in a SPATEM message

Term	Description
signal group	grouping of synchronised signals at a junction; one or more signals that control certain traffic flows together and always show the same signal state
SPATEM	Signal Phase And Timing Extended Message
TSC	traffic signal control
V2X	Vehicle-to-everything; denotes e. g. vehicle-to-infrastructure communication

Appendix: questionnaire

As part of the project, expert interviews were conducted with representatives of various stakeholder groups in order to identify key requirements for the evaluation of forecast. As stakeholders from different domains were involved, a glossary and an illustration were also prepared for this purpose, which served as a common terminology and basis for discussion. The glossary used and the 18 questions are listed below.

Questionnaire

1. For which application(s) do you use the forecast?

2. On which specific value within the forecast is your application focused on?

3. Is the forecast only relevant at a certain point in time or in a certain time interval before the signal timing?

4. Is the signal state change from red to green or from green to red relevant for your use case?

5. Which definition (definition according to C-Roads "C-ITS Message Profiles" / definition according to SAE J2735) do you think makes more sense? (see glossary)

6. What challenges do you see using the forecast?

7. Who uses your application?

8. How do you determine the quality of forecasts?

9. What are the minimum requirements that forecasts have to fulfil in order to be usable? What are the characteristics?

10. May a usable forecast change over time? And if so, in what characteristic?

11. What characteristics must changes within a signal time forecast have so that this forecast is no longer useful?

12. What forecast horizon (in seconds) do you need for your application?

13. Which criteria must forecasts fulfil depending on the forecast horizon? (e. g. max. deviation, frequency of deviation, accuracy)

14. What distinguishes a "good" from a "bad" forecast in your opinion?

15. Do you currently assess the quality of forecasts? If so, which tool do you use?

16. Which additional characteristics or parameters that could be broadcasted by traffic lights (in addition to SPaT) would be desirable to improve your application?

17. For which applications should a evaluation metric of forecast quality used in the future?

18. How should the evaluation of forecasts be depicted? For example, as a visual representation or as an output of individual or multiple characteristic indicators?

Glossary for expert interviews

Forecast: describes the future/forecasted signal state and the associated signal state switch

Forecast horizon: describes the time interval (in seconds) of forecasts

Forecast quality: describes how forecasts fulfils previously defined requirements with regards to various features and characteristics

Signal state switch: the actual status switch of a signal/signal group (e. g. red to red/amber to green)

Cycle: describes the run-through of all phases in a sequence so that all road users get green once; can be fixed or dynamic, e. g. with the omission of some phases

Cycle second: describes a point in time within a cycle, calculated from the start of a cycle

Point in time: defines a specific cycle second

Time interval: describes a period of time (from cycle second X to cycle second Y)

Characteristic: refers to how a requirement or forecast property is specifically expressed or realised

Change: describes the comparison of the forecast at two points in time and the deviations that occur in the process; describes the change within the forecast over time/over a defined time interval

Start of green: signal state switch from red to green (red/yellow is assigned to red for simplification)

End of green: signal state switch from green to red (yellow is assigned to red for simplification)

Example figure: Figure 6.1 illustrates the following aspects of forecasts: At the current point in time, there is a forecast for the signal state switch from green to red, based on a specific forecast horizon. Forecast "A" predicts the signal timing too early. Forecast "B" predicts the signal timing too late. The characteristic describes the change in the forecast over the various points in time/cycle seconds.

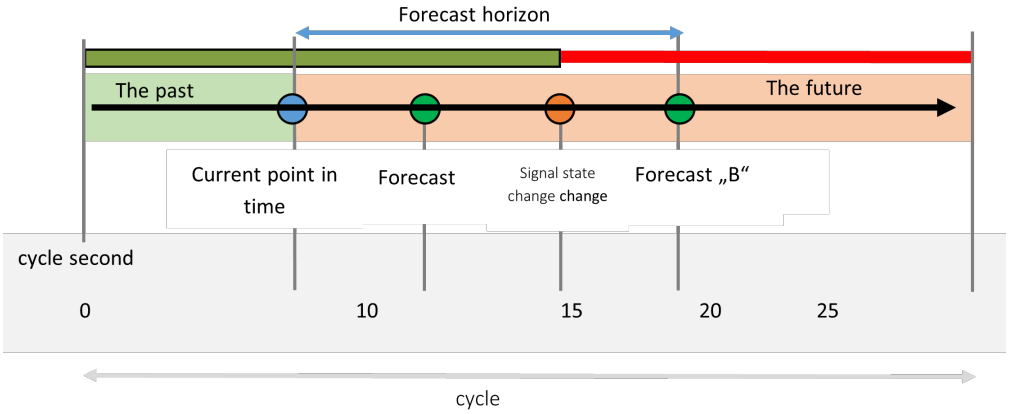


Figure 6.1: **Discussed forecast illustration used in expert interviews.**