Guidelines for TISA QBench Calculations

Version 1.0

26-January-2016

Document Control Sheet

Document Title:	Guidelines for TISA QBench Calculations
Document reference:	SP16002_TISA_QBench_Guidelines_v1.0.docx
Issues:	26-January-2016
Status:	Final Report
Editor:	Pieter Pauwels, Nikolaus Witte & Dyana Szibbo

Contents

1 Purpose of this document

This document describes the guidelines to be taken into account when using the Specification for the Traffic Flow TISA QBench Calculations. These describe the metric calculations for measuring the quality of Traffic Flow information.

The intended audience for these guidelines includes producers of traffic information, both public and private, buyers such as automobile manufacturers, personal navigation solution providers, mobile phone network operators and other media companies, or roadway network operators, as well as all intermediaries, third-party stakeholders and facilitators such as government agencies.

The complete TISA Value Chain can be seen as an audience for these guidelines:

The guidelines outline the purpose, limitations and expected outcomes of the Calculations. It is important to understand how the calculations should be performed and how the results should be interpreted. If not, incorrect conclusions can be derived from a calculated TISA QBench score.

As guidelines, the methods presented in this document leave room for interpretation and often balances general principles with formal rules. Only with correct usage and sufficient feedback it is possible to determine what constitute the best approach on a case-by-case basis. Thus the most important recommendation of all is that those reporting test results clearly document their methodology and its adherence to or start from the proposed guidelines.

It is therefore suggested that test results should be as transparent as possible in order to be recognized as legitimate benchmarks.

2 Terms and Definitions

For the purposes of this *guideline*, the following terms and definitions apply.

Definitions:

Travel Times	Expressed in seconds
Lengths and Distances	Expressed in meters
Speeds	Expressed in meters per seconds

Abbreviations & Parameters:

Symbol	Suggested Value	Description
d		Travelled distance
Cap	10%	Capping (in function of the Free Flow speed) to bound the measured speeds to limit the impact of single events
α	0.5	A factor used to reduce the impact on overestimating the travel time when calculating the Actual Benefit
V_{ff}	the maximum allowed speed for conditional access roads 80% of the maximum allowed speed for non-conditional access roads	Free Flow speed
T_{ff}		Free Flow travel time
V_{gt}		Ground Truth speed
T_{gt}		Ground Truth travel time
V_{rep}		Reported speed
T_{rep}		Reported travel time
V_{ss}	1m/s	Stand Still speed (lowest possible speed)
T_{ss}		Stand Still travel time
V_{ct}	50% of the Free Flow speed (V_{ff})	Congestion Threshold speed
T_{ct}		Congestion Threshold travel time
V_{lower}		Lower boundary of the Tolerance around the Ground Truth
V_{upper}		Upper boundary of the Tolerance around the Ground Truth

3 General

3.1 Purpose of TISA QBench

TISA QBench is intended for a B2B communication on the quality of traffic information. It will help the partners by:

- improving the consistency and fairness of evaluations;
- lowering their overall costs by eliminating duplication of efforts;
- establishing recognition for true value-added which will pull quality upward.

From a service provider point of view, it is possible to calculate a quality-indication inside the Value Chain. It can be done at different stages in the chain to compare the reality with the situation reported by the provider. It will show the impact of algorithms, assumptions, calculations, translations, latency...

A navigation devices manufacturer (located at the end of the Value Chain) can gather logs from the navigation devices and compare them with the reported information received on the devices (taking into account that reception was available in that area). It gives the possibility to compare different technologies (RDS/TMC, connected services, DAB/HD Radio) or compare different service providers.

The TISA QBench method is purely based on measurements of travel times. The purpose is to evaluate the accuracy of reported delays versus actual experienced delays by road users. No freeflow measurements are taken into account. Nor is the accuracy of freeflow travel times considered.

The TISA QBench result will show the benefit of the reporting service in comparison to the freeflow status.

It is strongly advised that you test your TISA QBench implementation. To this end TISA provides test data sets covering all aspects of the TISA QBench calculation. The test data sets and how to apply them are explained in a separate document.

3.2 TISA QBench in a Nutshell

TISA QBench is a measure for delay accuracy. Essentially, TISA QBench is the aggregated delay estimate of the reporting source relative to the aggregated delay of the Ground Truth. However, since over and under estimation of the reported delay would cancel each other out, delay is replaced by "benefit" (with respect to reporting free-flow condition): the benefit of the Ground Truth is its delay. For the benefit of the source the difference of its delay to the Ground Truth delay is always "counted negative", that is

- in case the source underestimates the delay, the benefit of the source is just its delay,
- otherwise the benefit of the source is the delay of the Ground Truth minus the delay difference of source and Ground Truth.

So the benefit of the source is always smaller or equals the benefit of the Ground Truth, and the absolute difference of the benefits equals the absolute difference of the delays.
The TISA QBench is then defined as the sum of the source benefits divided by the sum of the Ground Truth benefits. Inherent noise of road traffic and extreme measurements (outliers) are reduced by including extra parameters which can be set for conditional access and non conditional access roads:

- Bound Ground Truth speed and source speed from above and below.
- Set a tolerance: source delays different from the Ground Truth delay but within the tolerance are not penalized.
- Give different weights to over- and underestimation of delays: A falsely reported congestion might be less of a problem for a customer than a missing a congestion.

On a more conceptual level TISA QBench consists of three steps:

- Ideal Benefit as a benefit function of the Ground Truth.
- Actual Benefit as Ideal Benefit minus a penalty.
- TISA QBench as the quotient of sum of Actual Benefits and sum of Ideal Benefits.

This places TISA QBench in a much broader context of different quality metricies: e.g. for different benefit and penalty functions one can obtain Bogenberger's quality metric for traffic incidents (QKZ analysis, see [2003KBG]).

3.3 Limitations

When performing an analysis on quality, one must always include factors as timeliness (i.e. frequency of the provided information), ease of use (i.e. delivery format), and add subjective variables on top of that. These guidelines do not attempt to capture any of those elements. However, it is important to have sufficient knowledge of the matter and keep these factors in mind.

The earlier in the Value Chain, the better the quality. Any performed calculations, aggregations, technical translations,… in the chain will probably decrease the quality of the information. When comparing services, this must be taken into account.

The calculations and these guidelines do not contain any statement regarding map-matching, routing-specifications or any other pre-processing steps.

The calculations to derive a TISA QBench value are purely based on velocity-information (speed/travel time). It does not take into account vehicle intensity measurements (although they are of great use for public authorities and traffic managers).

The TISA QBench has no focus on the reporting of events such as accidents or road works. Although these are of huge interest for the partners in the Value Chain, they are not taken into account for the calculation of the TISA QBench.

The investigated road network has a big impact on the calculations of the TISA QBench. Only measurements on the selected network are taken into account. If for example a detour is recommended on basis of a reported delay on the planned route, this is not reflected in the TISA QBench if the detour is not in the investigated road network.

Lastly, a TISA QBench score is only a proxy measurement for the "real" user benefit, and as any technical method, has limitations. The TISA QBench method shall be used with the intention to optimize/maximize the "real" user benefit, rather than just the TISA QBench score. In particular, strategies to adjust the reported speeds solely for the benefit of an improved TISA QBench score (i.e. not for the purposes of reporting accuracy) shall be avoided. Such manipulation is considered an unacceptable misuse of the TISA QBench method and its technical limitations. Organizations applying the TISA QBench method hence are required to do so in good faith and keep their services optimized towards "real" user benefit.

4 Ground Truth

The 'Ground Truth', or testing data set, provides an independent view of the traffic information. It is used to assess how a provider's predictions generalize to an independent data set, giving an idea of their performance in practice.

It is important to note that the Ground Truth must be gathered and processed in an trustworthy manner; while ground truth is rarely without any error, it will serve as a reference to compare other traffic information sources. If there is bias in the collection or calculation of the ground truth, then it will not provide an accurate representation of performance. This document avoids setting formal standards for the Ground Truth, but does provide practical recommendations on how to collect reference data. Recommendations for test-drive and floating car methods are in the following two sections, respectively.

Important: when comparing the travel time on a driven section, the Ground Truth is defined by a measurement at *exiting* a driven section. It must be compared with Reported Information at *entering* the section.

4.1 Test Drives

Ground Truth information can be gathered with test-drives.

The earlier QFCD method has shown the feasibility of using drive tests to collect Ground Truth information [2009QFCD] as was deployed first in the USA for quality measurements of traffic congestion reports [2007SAE]. This method of collecting Ground Truth is effort-intensive and requires test drivers collecting ground truth information.

The ground truth can be gathered with a data logger which gathers at least a timestamp and gps position. With sufficient accuracy and update frequency for map matching and later TISA QBench score calculation, see also further recommendations in [2010NATWG].

The gathered logs must be aggregated to meet the spatial resolution of the Reported Information. This can be granular offsets or complete TMC Links. The aggregated logs should become a one-on-one comparison (travel time) with the Reported Information. The calculation of the travel time must estimate the exact time the vehicle has entered the segment and the exact time it has exited the segment.

For representative testing, care should be taken in the route selection, i.e. which roads to drive during which times of the day. Also test drivers must mimic the experience of the "average" driver, i.e. follow the bulk of the traffic, not always the slowest or fastest lane. To follow the traffic with the majority of the other drivers present on the road at that time, but with respect to maximum allowed speed.

This guideline does not want to encourage any dangerous behavior on the road. It is important to maintain the highest safety for both the test driver and the other participants in traffic. In reference [2010NATWG] [2011MDOT] recommendations are given for route selection, driving behavior and test equipment needed.

Test drives and comparison with reported travel times will produce a test result, i.e. a TISA QBench score. The question is what this TISA QBench score resulting from a limited test has to say on the "real" quality as experienced on a day-to-day level. The first aspect is that the test shall be representative, see earlier. The second aspect is to understand the relative accuracy of the test result.

The statistical accuracy of a test result is expressed with a confidence interval. A 95% confidence interval around a measured value indicates that a 95% level of confidence can be had that the "true" value ("true" TISA QBench score) lies within this interval. With a relatively small set of observations resulting from test drives, such a confidence interval for the mean quality value can be estimated easily with "bootstrapping", a well-known statistical method based on resampling [1979EFB], see also section 7.1.

Enough test driving should be done, such that the 95% confidence interval for the TISA QBench score is relatively small. A small confidence interval makes it easier to prove that with 95% certainty a service exceeds a minimum TISA QBench level (i.e. exceeds the minimum level of the confidence interval). If this is not the case, the consequence is that either more test drives have to performed (to reduce the width of the confidence interval), or the service quality itself needs to be improved.

The amount of test driving needed to produce such a small confidence interval depends on the level of congestion in an area. In very congested areas, a limited amount of test driving suffices to collect enough (congestion) ground truth. In rather uncongested areas, much more time needs to be spent to drive an equivalent time in congestion.

4.2 Automated using Probes

Another proposed method for collecting Ground Truth data against which to compare Reported Information is to use floating cars, or "probes". This method uses trips that are recorded by vehicles outfitted with a GPS, ideally a fleet of vehicles representing the "average driver" (or end users the performance is being measured for). The traffic conditions met by those probes are considered to be the Ground Truth. Each probe's observed travel speed is compared with the speed estimate from the traffic information provider. The calculation of the travel time for each probe must estimate the exact time the vehicle has entered the segment and the exact time it has exited the segment.

These probes must be independent from any probes used to generate the traffic information provider's estimates. The probes can be seen as multiple test-drives, and with a sufficient test fleet size, this method is able to overcome the sample size limitations of the Test-drive method. For example, a fleet of probes will be much better suited to estimate a network-wide score, as they can cover more space and time than Test-drives.

The processing of these test-probes into the Ground Truth speed estimates can be more complex. Map-matching, routing, link speed calculations and other pre-processing must be performed to obtain the ground truth, and those calculations may introduce some error. Some approaches to measure and/or minimize this error include:

- Measuring and minimizing the error rate of map matching.
- Measuring and minimizing the error rate of routing.
- When a static window is used (instead of a rolling window), only probes that cover the static window sufficiently should be used (suggested: 90% coverage).
- Filtering out probes that do not represent the average driver's behaviour.
- The used probes should represent a sufficient coverage in space and time to provide a relevant score for the chosen network. For example: measurements on a small area can not be seen as a global representation of a bigger area.

Note: The current iteration of this document considers a *single vehicle* driving a *single run* on a present route.

4.3 Other technologies

More technologies are available for collecting samples from individual vehicles. These can be toll-tag readers, automated number plate recognition readers, Bluetooth/WIFI mac address readers,.... They can provide a Ground Truth for one specific fixed part of the road network. The vehicles are been detected on site A where an unique ID is been merged with a timestamp. On site B the same unique ID is again been detected in a later time where the time difference is been used for the fixed road length between site A and site B. Take in account that Bluetooth/Wifi mac address readers will detect around 30%-40% of all vehicles on one specific part of the road network in Europe.

It is important to reach the spatial resolution of the Reported Information as good as possible. The gathered information should become a one-on-one comparison (travel time) with the Reported Information.

5 Reported Information

The Reported Information can be derived on different stages in the Value Chain. It can be derived directly in the traffic information system of the provider where its quality should be the highest (minimum latency, best granularity, least modifications).

A good stage would be directly after all needed algorithms and transformations. For example, the transformation of the location-description into a TMC-location description is quite common and will have an influence on the encountered quality. This would be the stage immediately before sending the information to the partner or broadcaster.

The final stage would be inside an end-user-device as this is the information seen at the end of the Value Chain.

Important: when comparing the travel time on a driven section, the Reported Information is defined by a measurement at entering a section. It must be compared with the Ground Truth at exiting the driven section.

It is important to reach the same spatial resolution between the Ground Truth and the Reported Information. The gathered information should become a one-on-one comparison (travel time) between the two.

It might also be of interest to compare two different methods of gathering flow information. For this, one method can be considered as the Ground Truth while the other is seen as the Reported Information.

6 Distinctions

During the calculations, it is suggested to make a distinction based on time and location. This will result in multiple scores but will provide a better overview on the relevance of the available Traffic Information.

6.1 Distinction on Time

It is recommended to make a distinction in the periods of time for performing tests. Depending on the goal of testing, some periods in time might be of higher interest for the tester.

The definition of the periods of time can be seen as country-dependant. These periods can be agreed upon between the two partners that would like to calculate TISA QBench scores. As a default, the following periods are defined:

* Morning Rush Hour
  On weekdays (no holidays) between 06h00 and 10h00
* Day
  On weekdays between 10h00 and 16h00
  During the weekend and holidays between 06h00 and 20h00
* Evening Rush Hour
  On weekdays (no holidays) between 16h00 and 20h00
* Night
  Between 20h00 and 06h00

A TISA QBench value can only be calculated when a congestion is ongoing or reported. For some periods of time it might be difficult to get enough data to perform a good calculation.

6.2 Distinction on Location

It is recommended to make a distinction in the classes of roads taken into account of the tests. Two types of roads are defined:

* Conditional Access roads:
  Centrally divided expressways with no intersections. Only connections via ramps, no at-grade crossings.
* Non Conditional Access roads:
  Intersections (roundabouts, traffic lights,…) allowed.

During testing it is best to deal with these two types separately. Test results may vary hard between the two types of roads and also the relevance will not be the same.

An extra location-based division is also possible for a specific region, for example a capital city or an important region (industrial, airport, special events,…). This can be agreed upon between the partners.

7 Reliability of TISA QBench Score

In the TISA QBench Calculations a clear description of the mathematics is given. Using this will result in a quality-value for the provided Ground Truth and Reported Information. To establish a trustworthy value, it is recommended to also provide some measure of reliability, i.e. is the used ground truth a representative sample. In particular one should pay attention to network coverage, total amount of data and distribution of the data over the relevant period of time. One standard method to provide a confidence interval is bootstrapping but alternative techniques to provide reliability scores may be applied as well. Bootstrapping and some alternatives are presented in the subsequent chapters.

7.1 Confidence Interval

Given a sample and some measure on the sample, bootstrapping provides a method to construct a confidence interval for the measure. In the case of TISA QBench the sample is a set of travel times comparisons, where a travel times comparison is a pair of Ground Truth travel time measurement and the corresponding travel time reported by the provider. Let N be the size of the sample, then a new sample is constructed by drawing N-times from the set of comparisons "with putting back", that is some comparisons may appear more than once in the new sample and some not at all. This process of re-sampling is repeated 1000 times and the resulting set of TISA QBench scores is ordered by their value. Now, to construct the 90% confidence interval the 5th and 95th percentile of the sorted TISA QBench scores are the lower respectively upper bound of the interval.

Bootstrapping is recommended for small data sets, e.g. from a test-drive. In case of large data sets like automatically collected floating car GPS data, bootstrapping can be time consuming and memory intensive - after all the re-sampling requires the original data set. With the use of histograms and sparse matrix implementation bootstrapping is also doable for large data sets. However, with large datasets the confidence interval will be most likely insignificant.

7.2 Scores Stability for Large Test Sets

In the case of a large test data set, such as with the automated probe method, the stability of a QBench score could be studied using simple means other than bootstrapping. For example 2nd standard deviation of the mean is an appropriate stability test for large data sets.

Alternatively, in studying a QBench network score for a given period, one possibility is to observe the QBench score as each measurement comes into the system: the QBench score is calculated using 'all data points until now'. If the Ground Truth is sufficiently large, the QBench score will stabilize with increasing amount of aggregated BActual values.

A smoothed version of this can also be plotted, and the smoothed subtracted from the unsmoothed may be plotted on an additional graph to give an idea of the score's 'noise', and to view how it diminishes over the given period.

For an idea of the network coverage of the score, the network coverage across the day could be viewed as well.

8 References

[1979EFB] Efron, B. (1979). Bootstrap methods: Another look at jackknife. Ann. Stat. 7, 1-26.

[2003KBG] Bogenberger, K. The Quality of Traffic Information. (2003).

[2007SAE] Dance, F., Gawley, D., Hein, R., and Kates, R., Enhancing Navigation Systems with Quality Controlled Traffic Data, SAE Technical Paper 2008-01-0200, 2008,

[2009QFCD] Bogenberger, K. and Hauschild, M. 2009. QFCD – A Microscopic Model for Measuring the Individual Quality of Traffic Information. Proceedings of the 16th ITS World Congress and Exhibition on Intelligent Transport Systems and Services, 21-29 September 2009, Stockholm, Sweden.

[2010NATWG] NATWG, Traffic Information Benchmarking Guideline (TIBG) Version 1.0 RELEASE 204-22-101.

[2011MDOT] TTI, Virginia DOT et al., GUIDELINES FOR EVALUATING THE ACCURACY OF TRAVEL TIME AND SPEED, Version 1.0, june 2011