FESTA Support Action
Field opErational teSt supporT Action

D2.4 - Data analysis and modelling

Grant agreement no.:	214853
Workpackage:	WP2.4
Deliverable n.:	D2.4
Document title:	Data analysis and modelling
Deliverable nature:	PUBLIC
Document preparation date:	May 21 2008

Authors: Sylvain Lassarre, Marco Dozza, Samantha Jamson, Frank Lai, Farida Saad, Anna Vadeby, Trent Victor, Rino Brower, Oliver Carsten, Alessandra Disilvestro, Philippus Feenstra, Jeroen Hogema, James Lenard, Claire Minett, Andrew Morris, Martijn van Noort, Jeroen Schrijver, Alessandro Taddei.

Consortium:
Centro Ricerche Fiat S.C.p.A., University of Leeds, BMW Forschung und Technik GmbH Daimler AG, Gie Recherches et etudes PSA Renault, Volvo Car Corporation, Volvo Technology Corporation, Robert Bosch GmbH, A.D.C. Automotive Distance Control Systems GmbH, Delphi France SAS, Loughborough University, Chalmers University of Technology, INRETS, TNO, BASt, VTT, INFOBLU SPA, Orange France, ERTICO, Universitaet zu Koeln

European Commission
Information Society and Media

Project funded by the European Commission
DG-Information Society and Media
in the 7th Framework Programme

Document Control Sheet

Project name:	FESTA
Workpackage, workpackage title:	WP2.4 Data analysis and modelling
Task, task title:	
Document title:	FESTA - D2.4 Data analysis and modelling

Authors
Sylvain Lassarre, Marco Dozza, Samantha Jamson, Frank Lai, Farida Saad, Anna Vadeby, Trent Victor, Rino Brower, Oliver Carsten, Alessandra Disilvestro, Philippus Feenstra, Jeroen Hogema, James Lenard, Claire Minett, Andrew Morris, Martijn van Noort, Jeroen Schrijver, Alessandro Taddei.

Reviewers
Herbert Baum, Torsten Geissler (Univ. Koeln)

Date of submission to consortium:	May 21 2008
Date of submission to European Commission:	date

Revision history:

VERSION	DATE	AUTHOR	SUMMARY OF CHANGES
V0.1	2008-03-29	S. Lassarre (INRETS)	document structure
V0.2	2008-05-06	S. Lassarre (INRETS) et al.	draft based on FESTA handbook
V0.3	2008-05-13	S. Lassarre (INRETS) et al.	Final draft for peer review
V0.4	2008-05-21	S. Lassarre (INRETS) et al.	Final version

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

Table of contents

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

Executive Summary

The chapter of the handbook and the deliverable on data analysis will provide guidance and general principles for

- pre-testing to check the usability of the system and the feasibility of the evaluation process,
- controlling the consistency of the chain and the precision with different sampling schemes,
- modelling the impact for each indicators and for an integrated evaluation including a systemic and multidisciplinary interpretation of the effects,
- integrating and controlling the quality of space-time data from various sources (numerical, video, questionnaires),
- selecting the appropriate statistical techniques for data processing, PI estimation and hypothesis testing in accordance to the list of indicators and experimental design,
- scaling up from experimental data and identified models to population and network level.

Experimentalists stress the role and importance of a preliminary field test in FOT. Three main objectives have been defined to make a preliminary diagnosis of usability of the systems and to check the relevance and feasibility of the evaluation process. These preliminary tests are very important for the practical deployment of the FOT as well as for the overall scientific evaluation process.

Recommendations about the monitoring of local and global consistency of the chain of operations from the database extraction to the hypothesis testing are given, especially to ensure the validation of the calculation of the Performance indicators.

Integration of the outputs of the different analysis and hypothesis testing requires a kind of meta-model and the competences of a multidisciplinary evaluation team, specially for interpretation of the system impact and secondary effects (behavioural adaptation, learning process, long-term retroaction, …).

In cooperation with WP2.2, methods for data quality control have been defined. Four types of checks have been defined to complement the information of the data base in order to prepare the data for the analysis.

Statistical methods have been described for three steps of the chain: data processing, PI calculation and hypothesis testing. They belong either to exploratory data analysis or to inferential analysis. Special attention has been given to the precision of the estimates of the effects or impacts of the system on the Performance indicators by stressing the importance of controlled randomisation and application of mixed regression models.

Scaling-up relies upon the potential to extrapolate from the PIs to estimates of the impact at an aggregated level. Three approaches have been defined to carry out the scaling up process from direct estimations to simulation models with the related assumptions. Models and

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

methodologies for scaling up results on traffic flow, environmental effects (e.g. PM10, CO2, Noise emissions in db) and traffic safety have been collected.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

1. Introduction

The strategy and the steps of data analysis need to be planned in order to provide an overall assessment of the impact of a system from the experimental data. Data analysis is not an automatic task limited to some calculations algorithms. It is the place where hypothesis, data and models are confronted. There are three main difficulties:

- the huge and complex amount of data coming from different sensors included questionnaires and video to be processed;
- the potential bias about the impact of the system(s) on behaviour which may arise coming from sampling issues including location of the study, the selection of a relatively small sample of drivers, etc.;
- the resort of auxiliary models such as simulation models to extrapolate from the behavioural effects estimated and tested within the sample to effects at the level of the whole transport system.

To be confident in the robustness of the outputs of the data analysis for the global evaluation, one has to follow some strategic rules in the process of data analysis and apply to the whole chain and to its five links (Figure 1) the required techniques such as applying appropriate statistical tests or using data mining to uncover hidden patterns in the data.

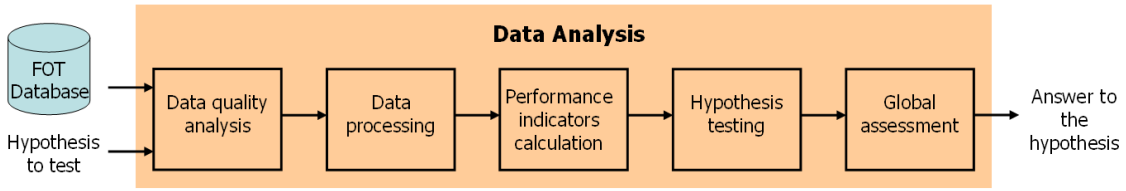


Figure 1 : Block diagram for the data analysis.

Some specific actions are required to tackle the difficulties mentioned above and to ensure the quality and robustness of the data analysis.

1. A pilot study is a prerequisite to check the feasibility of the chain of data collection and treatment and to achieve a pre-evaluation of the usefulness of the system. A lot of time can be wasted if this step is neglected because it is more difficult to restore the chain during the FOT.

2. As there will be a lot of computations from measurements to test of hypothesis through Performance Indicators estimations, the data flow has to be monitored in detail but also in the large. One of the strategic rules to follow is to ensure local and global consistency in the data processing and data handling and analysis. It is a loss to focus on a part of the chain of treatment if there is weak link. All the precisions gained from a particular step will be lost.

3. A lot of uncertainties will be part of the data because of the measurement and sampling errors. Stemming from the experiment design, the sources of variability and

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

bias of the PIs have to be identified, where feasible, in order to control them in the data analysis.

4. Many hypotheses have to be tested simultaneously. There is a crucial need for an integrative assessment process which could ideally combine within a meta-model information gathered on the usability, usefulness and acceptability of the system with the observed impacts of the system on behaviour. Furthermore, it is a multidisciplinary task. The estimated effects obtained from the sample of drivers and data have to be extrapolated using auxiliary models to scale them up.

5. Appropriate techniques have to be applied for each link of the chain : data quality, data processing, data mining and video analysis, PI calculation, hypothesis testing and global assessment. A brief description of them is provided. The techniques come from two set of statistical and informatics tools belonging to two main kinds of data analysis : exploratory (data mining) and confirmatory or inferential (statistical testing). The first one is useful to process signals and to identify sequences of events. The second is useful to test the impact by estimating the variances of the PIs' estimates according to the nested structure of the statistical units.

The development of theses five actions are presented in the following chapters 2 to 6 of the deliverable.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

2. Relevance of the evaluation process through preliminary field test

Conducting a pilot study is necessary to prepare the deployment of the FOT and to support the design of the relevant tools for the evaluation process. This task should be performed early in the evaluation process and should be carried out as soon as the first vehicles are available. These preliminary field tests represent an important step for the mobilisation and the dialogue between the various teams involved in the FOT and for promoting a common framework and consensus for the evaluation process. For being relevant, these tests should have an adequate duration. These preliminary field tests have to deal with three main levels of analysis with specific objectives.

1. Obviously, the first preliminary field tests have to check *the technical functioning of the data collection systems in real driving situations*. They should enable to identify potential problems of sensor calibration or drift and thus to establish the periodicity of maintenance procedures during the FOT. They should also permit to validate the data collection procedure from data acquisition, data transmission to data storage. The technical teams involved in the FOT should be in charge of these field tests.

2. The second level of preliminary field test deals mainly with the issue of *assessing the usability and usage of the systems under study and of identifying the main critical issues associated with their use in real driving situation*. This is particularly relevant for:

- Structuring the familiarisation phase of the drivers before their participation to the FOT;
- Contributing to the design of the questionnaires for the subjective assessment of the systems;
- Testing and/or improving the various tools developed for data processing, such as automatic identification of critical "use cases" and "scenarios" and video based identification of triggering events or categorisation of road and traffic contexts.
- Identifying a number of critical scenarios when using the systems, scenarios that could be investigated more extensively when the data gathered from the FOT are processed and analysed.

  This test requires the participation of a sufficient number of drivers (depending of the target population in the FOT) and should be performed in real driving situation. An experimental journey on the road could be designed for that purpose (depending on the hypotheses formulated). This level of analysis provide useful data for designing the relevant tools for the evaluation process as mentioned above, for estimating the time required for data processing and data analysis and thus calibrating these phases in the FOT. It may be seen also as an opportunity for training the team (s) in charge of data processing. Finally, it represents an important step for testing some of the hypotheses formulated in the FOT and/or for refining them.

  Psychologists, Ergonomists, and Human Factor experts should perform these tests in close cooperation with the teams in charge of statistical analyses as well as the team in charge of developing processing tools.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

3. The third level consists in *testing the feasibility of the overall evaluation process* from the selection of the participants to the data collection. It is a kind of final rehearsal before the deployment of the FOT. It enables in particular to check the communication process between the various teams involved in the practical deployment of the FOT and the robustness of the technical tools designed for data collection and transmission.

These preliminary tests are very important for the practical deployment of the FOT as well as for the overall scientific evaluation process.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

3. Consistency of the chain of data treatments

There will be a lot of computations and data flow starting from the measurements contained into the data base to the test of hypothesis through PIs estimations and to the global assessment. This process in the form of a chain of operations has to be monitored in detail but also overall. It is inefficient to focus on a part of the chain of treatment if there is weak link. All the precisions gained elsewhere will be lost.

There are five operations linked together in terms of data treatments: a data quality control, a data processing and mining, a PI calculation, a testing of hypothesis and a global assessment. It is a bottom-up process which takes as input the outputs of the previous operation. In addition, three kinds of models are needed as support to carry out the three top operations : probability models for justifying the calculations of the PIs, integration models to interpret in a systemic way the results of the test, auxiliary models to assess the effects on a larger scale (scaling up). Moving from the data to an overall assessment is not only a bottom-up process; it has also to include some feedbacks (Figure 2). There are two movements along this chain: a data flow going up and a control feedback loop from the top about the consistency of the evaluation process which mainly depends on the control of the uncertainty.

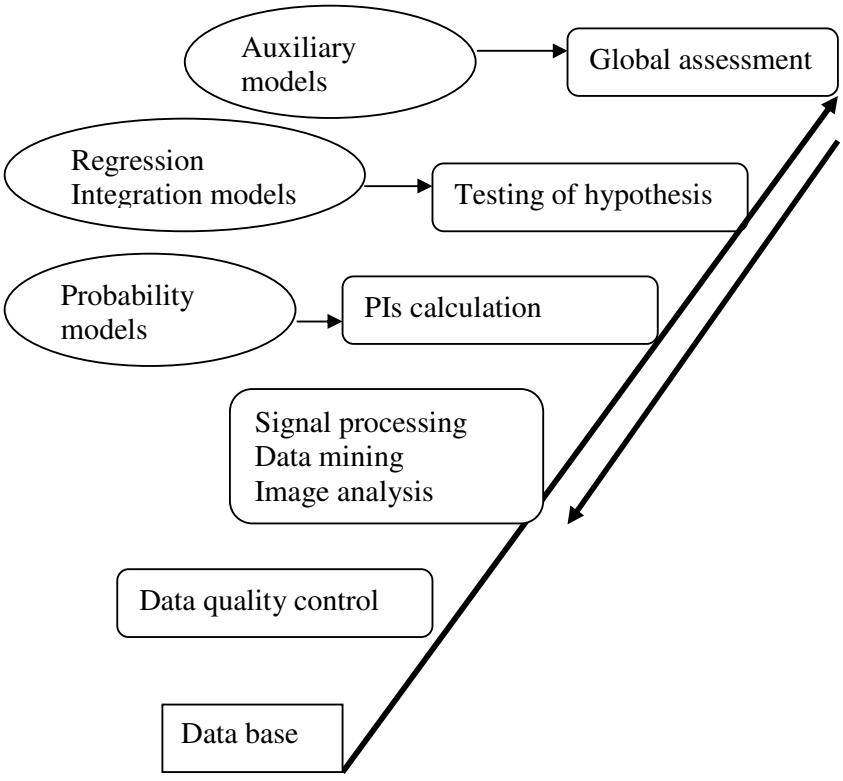


Figure 2 : Deployment of the chain with feedbacks and additional models.
(The bottom-up arrow correspond to the operational chain of data analysis and the top-down arrow to the reflexive process of data analysis, horizontal arrows to the plug-in of external models to the top links of the chain)

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

In moving up of the chain, the consistency of each operation can be checked locally according to the specifications which are governed by the nature of the PIs which correspond to a set of hypothesis related to the use case of the system. For each PI, there are some rules which give the domain of validity of the calculations procedures. For example, it is important to sample data which can change rapidly at a high data rate. The sampling rate must fit the variability of the variable. From a data base point of view, however, it may be easier to collect relatively static data at high frequency.

It is a mistake to measure with great accuracy a variable which is evolving smoothly. The sampling rate must fit the variability of the variable. The variability of the indicator will come from each of the three levels: driver, driving situation and measurement. The total variance of the estimate of a performance indicator on the sample breaks down into an inter-individual (between driver), intra-individual (between situation) and infra-situational variance (between measurement). If the inter-individual variance is strong, the increase in the size of the situations observed and in that of the measuring points per situation will not bring any precision gains. Playing on the size of the samples and on the quantum is the way to monitor that the correct scale is used and that the uncertainty of the indicator's estimate is correctly spread.

In complement to the local consistency, a global criterion is to have sufficient sample to get enough power to carry out the test of an hypothesis or to make an overall assessment with enough precision. This is a feedback loop coming from the top to control the uncertainty of the estimations. The chain of the calculations procedures must be reliable and smooth with an overall consistency. The precision required for measurements depends on the uncertainty of the auxiliary models, of the regression models and of the probability models. When combining different PIs related to acceptability, usability and utility inside an integration model or extrapolating effects by means of auxiliary models to get the global assessment, the uncertainty has to be equally distributed. It is not only a bottom-up process from data to the overall assessment, it has to include too some feedbacks.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

4. Precision in sampling

The design of an FOT should be undertaken, as far as possible, in the same way one would design a traditional laboratory experiment. The aim is to measure the effect of an intervention or treatment, which in the case of an FOT is the use of a system or systems, on a sample of subjects and in various driving situations while controlling for external conditions. From the sample, we have to infer the effect on the population by aggregating the values obtained through the sensors without and with the system , to get an estimate on the effect on the chosen performance indicator.

According to the experimental plan, there are four factors units:

- driver which is random factor,
- system which is a fixed factor with two modalities : without(neutral)/with the system,
- situation which is a random factor,
- measurement or space-time quantum which the ultimate random factor.

The hazard comes from the sample of drivers and driving situations which are taken at random from larger populations and from replications within the situation according to the frequency of measurement. The system use (without/with) is a fixed factor.

The factors are either embedded or nested:

- a combination <driver*system<situation<space-time quantum>>> if the drivers use all the systems included the neutral mode (normal driving without the system)
- a combination <driver<system<situation<space-time quantum>>> if there is a group of drivers for each system.

How to insure that this inference is valid, in other world, that the estimation is very near the true effect in the population ? The precision of the estimate depends on the bias and variance which could be combined to get a measure of the sampling error. To control the bias and variance, one has to rely on well defined sampling plan using appropriate randomisation at he different levels of sampling : driver, driving situation and measurement.

Consideration should be given to identifying the possible sources of (unintended) bias and variance in the sample and either attempt to minimise or account for these in the data analysis. This is one of the most fundamental principles of statistical methods. For example, the success of the analysis of variance technique depends on being able to isolate as many sources of variance as one can. The variance contributed by each of these isolated factors can then be assessed by comparing it to residual, or “uncontrolled” variance, which ideally should be as small as possible. The following sections describe the possible sources of bias and variance and the associated methods the researcher can use to account for them.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

4.1 Driver variation

The simple fact of the matter is that drivers vary. If drivers were identical in all respects, very few would be required to take part in a FOT as generalisations could easily be made. We can, therefore, never measure how all drivers would react to the system under consideration in the FOT. Instead we take a sample form the population (made up of all the other samples we could have taken at any time). We then infer from the responses/outcomes of the sample, combined with the variation within it, something about the effect of the system on any other sample we might have chosen. The accuracy of this inference depends on the size of the sample, the extent to which drivers vary and the efficiency in design and analysis of the FOT.

The range of behaviours that drivers exhibit (in terms speed selection, headway preference, overtaking behaviour) is immense, but fortunately obey to some probability laws and models. On one hand a FOT should attempt to include drivers who exhibit behaviours right across the spectrum. Therefore, drivers who prefer to travel at the speed limit should be included alongside those who prefer to travel in excess. The two types of drivers may exhibit very different reactions to a system or rate its acceptability in different ways. On the other hand, this range of driving behaviour is problematic in that it can affect the statistical analysis and lower the power of an FOT.

Strict randomisation procedures ensure that only the outcome that is being varied (or the outcome whose variation we are observing) is working systematically. It is then possible to ascribe the variance to the unknown competing variables. The smaller the variance, the more informative the FOT will be. However, strict randomisation is not usually possible or desirable[1] in an FOT, particularly when the sample sizes are relatively small.

The theoretical best method is to stratify the population of drivers according to some variables or factors related to the outcome and to sample proportionally to the size of the sub-population and to the a priori variance of the outcome e.g. speed choice). For practical reasons, a different sampling or selection procedure may be followed. In either case, it is important to be able to compare the sample to the overall driver population in order to identify what are the main discrepancies and to assess possible sources of bias. All a priori information about the variances of the outcome have to be collected to adjust the sample size in order to minimize the variance of the estimate.

[1] It may not be desirable, for example, to waste sample size by recruiting drivers who only drive small amounts each week. Many FOTs have for good reasons used quota sampling procedure, in which equal numbers of (say) males and females are recruited. This can create bias when scaling up the observed data to estimates of effects at a national or European scale.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

4.2 Driving situation variation

When designing an FOT it is sometimes necessary to select drivers who carry out their journeys on particular road types, times of day etc. in order to test a particular system effectively. However, even using data extraction techniques that identify the appropriate journeys, there will be variation within and between those journeys and the driving situations within these journeys. For example a particular journey may be affected by congestion part-way through, or weather conditions may change from day to day. This type of variation cannot be controlled and are considered as random. Again the same concern as to be applied in order to apply strict randomisation procedures. One has to check that the sample of driving situations cover the range of prevailing driving situations. The observation period should be sufficiently long to allow for these random effects. One example here is that seasonal effects should be considered. An a priori information about the variance of the outcome related to the driving situations will be useful to define the sample size.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

4.3 Measurement variation

Once in a driving situation, by means of the sensors, we get a series of measurements at a certain frequency. Their size is not fixed but varies. Each set of measurements within a driving situation constitute a sample of units taken from a cluster, according to the sampling theory. Usually, there is a correlation between the measured outcomes. The information coming from this sample of measurements is not so rich as expected from an independent sample. One such cluster is at the driver level – the data collected from one driver is not independent.

How to quantify the variance of the estimate of an outcome from the experimentation taking into account these three sources of variations? Let us pose the problem of the case in which an indicator takes the form of a quantitative variable. One assumes that measuring the variable on which the performance indicator depends is sampled in time and space. One obtains an interlocking of statistical units indexed by i, j, k: <driver<situation<space-time quantum>>>. If one takes the example of the ACC in car-following situations in urban areas, the driver i stemming from a sample of n drivers will be a certain number of times n_{ij} in that situation during his journeys, and the measurements will apply to a sample n_{ijk} of measuring points in time and space (space-time quantum).

The variability of the indicator will come from each of the three levels:
<driver< situation<space-time quantum>>>, with variances σ^2_C, σ^2_S, σ^2_Q, which are measures of the dispersion at each level. The variance of the average indicator on the sample is equal to

$$\text{var}(\bar{I})=\frac{\sigma_C^2}{n}+\frac{\sigma_S^2}{nm}+\frac{\sigma_Q^2}{nml}$$

where n is the size of the driver sample, m the average number of journeys*situations for one driver and 1 the average number of spatial-temporal measuring points in a situation. The precision depends crucially from the first term, i. e. from the variation between the drivers, as the other variances divided by their respective and important sample sizes are negligible.

The total variance of the average of the indicator on the sample breaks down into an inter-individual, intra-individual and infra-situational variance. If the inter-individual variance is strong, an increase in number of situations observed and in the measurement points per situation will not bring any precision gains (Särndahl and al., 1992). However, it may help to ensure a reduction in bias from, for example, seasonality.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

5. Requirements for Integration/ Scaling up

Having treated and aggregated the data by means of statistical models, there are two kinds of problems to solve related to first the synthesis of the outputs and second to the scaling up of the results from the sample to a larger population. Integration of the outputs of the different analysis and hypothesis testing requires a kind of meta-model and the competences of a multidisciplinary evaluation team (Saad, 2006). Scaling-up relies upon the potential to extrapolate from the PIs to estimates of the impact at an aggregated level.

Integration is probably one of the most critical tasks because of the large amount of data to process and the diversity of hypotheses and research questions to deal with. This synthesis has to deal in particular with the:

- *Direct effects* of a the system under study on the users and driving as well as its *indirect effects* (or behavioural adaptation) on the users and on non users (imitating effects or modification of interactions between user and non user);
- *Short and long-term effects* of system(s) used;
- And their impacts on safety, efficiency, environment, mobility, acceptance and adoption.

It has been observed in previous FOT that:

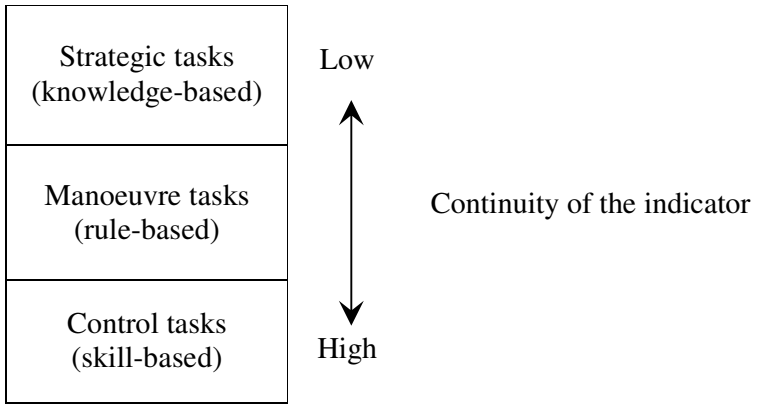
- drivers did not necessarily accept the most efficient system (in terms of speed limit abidance for example);
- drivers' acceptance of the system may vary with time of exposure (better use of and compliance with the system goals in the short term than in the long term);
- some indirect effects may counter-balance or compromise the positive Direct effects of the system;
- system use and efficiency may vary according to the situational context and so on. These are examples of the issues with have to deal with when making the synthesis of the results of previous FOTs.

The results obtained for each Performance Indicator (coming from objective and subjective data) have to be compared and weighted with reference to *well-established criteria* related to each impact domain. This includes three main tasks:

1) Identification of tests and models of impact for each indicator

Appropriate approaches (e.g. statistical tests) are required for investigating the effect of the indicator with reference to its nature; for example whether the indicator refers to the control, manoeuvre, or strategic elements of the driving task (Michon, 1985). The lower the task in the hierarchy, the more frequent the interactions with the rapidly changing environment around the driver and hence the degree of continuality of the indicator in question.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.



Discrete variables (e.g. choice of route) may be analysed in terms of frequency, while continuous variables (e.g. vehicle speed, headway, and lateral position etc) are prone to be studies by central tendency (e.g. mean, median etc). The effect of the indicators may be further investigated by comparison across categories of other temporal variables (e.g. time of day), spatial variables (e.g. type of road), or demographic variables (e.g. gender and age etc).

2) Harmonisation of models for a synthetic evaluation.

While it is often necessary to employ quantitative models from previous studies to investigate the effect of the indicator in question, it is worth noting that individual models were developed on different bases and assumptions. For example, some traffic models are based on rural roads, which will not be appropriate to be applied on indicators derived from urban environment. Similarly, an accident model developed for young drivers would be very different from that developed for mature drivers. However, in the absence of appropriate models available for the purpose of study, it is likely to employ a second best model with appropriate weighting or adjustment.

3) Interpretation of analysis results

It is important to bear the constraints, assumptions, and implications behind the design of study in mind when interpreting the analysis results. Behavioural adaptation may lead to side effects (i.e. indirect effect); for example, a lane departure warning system may inappropriately encourage a fatigued driver to carry on the trip. Behavioural adaptation may also result in prolonged learning process; for instance, an in-vehicle speed limiting device may cause the driver to alter his/her selection of route with reduced degree of engagement. Since an ADAS essentially alters the content of the driving task, a drivers' learning process of using a new ADAS system may or may not settle within such a short space of time. For example, driver behaviour observed from a 2-hour trial may contradict the results if the trial lasted two months.

Sometimes it is necessary to include individuals from different backgrounds and disciplines into the process of interpretation of analysis results, as a set of data may deliver

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

cross-disciplinary implications. For example, a piece of analysis result may be interpreted in favour of safety while against environmental benefits. It is also common practice to include stakeholders in the process of study to widen the impact of the research. In the event of the result from the FOT being related to policy making process, it would also be beneficial to recruit manufactures (e.g. vehicle, or nomadic devices), and relevant authorities to reach consensus on the use of research results.

Extrapolating from the sample to the population depends on the external validity of the experiment. The power of generalisation to the population of the estimates of impact is related to their precision which is composed of two parts bias and variance. We can use three approches :

1) if the required performance indicator measure PI is available in the sample (e. g. if journey time is an impact of choice for efficiency and journey time has been collected), the impact at the population level can be calculated directly, althoug sometimes a correction factor or othet form of extrapolation adjustment may have to be introduced (Cochran, 1977). Formely, the direct approach from the sample itself with a possible rectification (redressement) towards the population P could be written as E(PI(P))=E(PI(s)) with E meaning mathematical expectancy. The estimate based on the sample s gives a value which converge in expectancy to the population value if the sampling plan respects randomisation procedures. In sampling theory, we can adjust a posteriori the sample s to the margins of structural variables or by means of a stratification in order to extrapolate in a better way (Cochran, 1977).

2) If neither a performance indicator nor a proxy indicator are available, then it is necessary to adopt an indirect approach through individual/ aggregated models which provide an estimate of the output from the behavioural PIs estimated from the sample. Formely, E(PI'(P))=E(f(PI(s))) or E(PI'(P))=f(E(PI(s))) , with the function f () representing the model. Speed changes can be translated into changes in crash risk by applying statistically derived models from the literature which have investigated the relationship between mean speed, speed varaince or individual speed and crash risk. Emissions models can be used to calcualte the instantaneous emission of a car as a function of its recorded speed and gear selected.

3) Finally a macroscopic or microscopic traffic simulation model can be applied to translate the effects observed in the sample to a network or traffic populations effect. The outputs from such a simulation can for rexample, be used to calculate journey time effects or fuel consumtions effects at the network level. Combined with individual(/aggregated) models, it can be written E(PI(P))=E(PI(s')) and E(PI'(P))=E(f(PI(s'))) , s' being the simulated sample.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6. Appropriate techniques at the five links of data analysis

The five links follow the right branch of the development process of a FOT from data quality control to global assessment. Different techniques of data analysis and modelling which could be used at each step are presented here.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6.1 Step 1 : data quality analysis

Data quality analysis is aimed at making sure that data is consistent and appropriate for addressing the hypothesis of interest (FESTA D3, Chapter 4.5). Data quality analysis starts from the FOT database and determines whether the specific analysis that the experimenter intends to perform on the data to address a specific hypothesis is feasible. Data quality analysis can be performed by following the 4 sub-steps reported below (and shown in Figure 3) and provide, as a outcome, a report detailing the quality of the data to be used to test the hypothesis of interest.

Sub-steps for data quality analysis:

a. **Assessing and quantifying missing data** *(e.g. percentage of data actually collected compared to the potential total amount of data which was possible to collect).*
b. **Controlling data values are reasonable and units of measure are correct** *(e.g. a 6 Km/h mean speed value may be unreasonable unless speed was actually recorded in m/s instead of Km/h).*
c. **Checking that the data dynamic over time is appropriate for each kind of measure** *(e.g. if the minimum speed and the maximum speed of a journey would be the same, then the data may not have been correctly sampled).*
d. **Guaranteeing that measures features satisfy the requirements for the specific data analyses** *(e.g. in order to calculate a reliable value of standard deviation of lane offset, the lane offset measure should be at least 10s-long; further, this time length may depend on the sampling rate; AIDE D2.2.5, Chapter 3.2.4).*

Please, notice that the first three sub-steps refer to general quality checks; thus, if any of these fails, data analysis cannot proceed. If a failure is encountered, it should then be reported to the database responsible so that the possible technical error behind can be tracked down and solved. However, the last sub-step is related to the specific analysis or specific performance indicator considered in the following data analysis steps. As a consequence if step 4 fails, it may not be due to a technical issue that needs to be solved but to an intrinsic limit of the collected data.

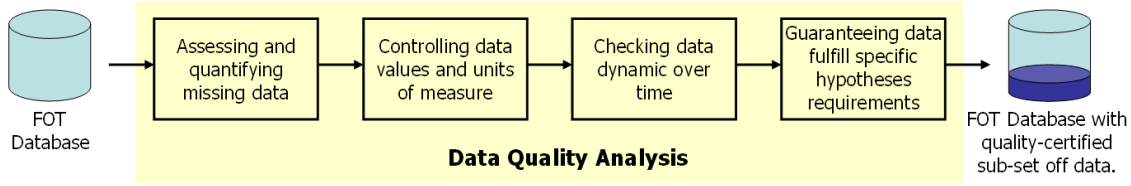


Figure 3 : Block diagram for the data quality analysis.

Data quality analysis implementation is reported (below) in distinguished paragraphs for data from on-vehicles sensors data (generally CAN data and video data) and subjective data (generally from questionnaires) due to the intrinsically different nature of these data.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

Data quality implementation: on-vehicle sensors data

- Assessing and quantifying missing data

No matter how well data collection and data storage process will be performed, some data is likely to be missing. This can be due to technical problems such as a broken sensor or a lose contact or to some human mistake during data downloading. The list of possible issues resulting in data loss is endless. In order to determine whether data loss is critical for data analysis, data loss needs to be quantified. The higher is data loss, the higher is the probability that the data set may be biased resulting in limited reliability of the whole data analysis process. It is recommendable to track down missing data in order to understand whether it may be related to a specific vehicle, driver, or technical issue and how this specific factor could bias data analysis. Depending on the factors causing data loss and on how much data is in play, different percent values of data loss can be more or less acceptable. As a reference, in the study RDCW (Road Departure Crash Warning Field Operational Test - University of Michigan – Transportation Research Institute) set to 5% the maximum acceptable percentage of data loss.

Statistics wise, data loss is not as crucial as far as it does not bias the measures. In fact, if the nature and the distribution of the data loss are known, different statistical models can be used to deal with missing values. However, the nature and distribution of the data loss are not always known. In this case, setting a maximum percentage for data loss is a more robust way to control for bias. In fact, in general the lower is data loss the less likely it is for the data set to be biased.

Depending on the hypothesis to be tested, different amount of data loss and biases can be accepted and different ways to deal with missing data may be acceptable. However, since data from an FOT is meant to be used in order to address several hypotheses if data is corrupted there will not be a standard and easy way to get around it. For this reason, in this step of data quality analysis the attention of the experimenter should be focused on verifying the quality of the data and possibly feedback this information to the database and data collection responsible personnel.

- Controlling data values are reasonable and units of measure are correct

Video data

For video data, in this sub-step the experimenter should check the frame integrity. This process may be hard to automate even if it may require only a fast visual inspection to a human. Fundamentally, a sample frame should be visualized and made sure the camera recorded what it was supposed to record. This task may seem simple and naive, however, this may be the only occasion to discover that a camera is partially obscured by some dirt, or it is moved, or tilted. Issues such as the ones just mentioned should be immediately reported to the technical responsible and solved as soon as possible since they may impair data analysis.

CAN data and external sensors data

Sensor data, both from CAN and external sensors, need to comply with the sensor datasheet and the general rules of physics. As a consequence, mean, max, and min values should be checked. For example a mean acceleration of $10 m/s^2$ is absurd and cannot be considered for data analysis. Instead, such a value should be notified to the technical responsible of the FOT so that the error can be traced down and solved. Further, accuracy, resolution, and frequency

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

of the measures should be checked and compared with data sheet values and physical limits (FESTA D2.1).

- Checking that the data dynamic over time is appropriate for each kind of measure
Video data for this sub-step is both a target and a tool. In fact, recording of video data needs to be verified but it is also the best tool to ascertain that data from CAN or other sensor is reasonable. Indeed, by comparing the scenes captured in the video data, with the values of sensor data such as speed, acceleration, yaw rate, rain sensors, wipers activity, etc… it is possible to cross-check the validity of the collected data. When video is not available, other measures can be crosschecked. For instance, yaw rate, steering wheel and GPS coordinate should be related. Data dynamic over time can be cross-checked also by analyzing the measures distribution. In fact different profiles can be expected for different measures and a distribution plot can capture in one plot, many hours of data which would not be possible for a human to check in a reasonable time.

- Guaranteeing that measures features satisfy the requirements for the specific data analyses
The main difference in between the data quality analysis described here and the one described in t FESTA D2.2 Chapter 4, is that at this stage of the process the data quality analysis procedure can take into account the specific requirement for the specific analysis to be performed on the data. These requirements may be strict and comes at different points in time during the FOT experimental protocol and data analysis design. Fro example, when calculating a specific performance indicator, specific requirements may need to be met by the measure. These requirements may be related to the data sample frequency, time length, granularity, accuracy, sensitivity, signal-to-noise ration, etc…. These requirements, which are performance-indicator- and/or analysis- dependent, need to be known and taken into account by the experimenter and verified before applying any algorithm to calculate performance indicator. It is worth notice that the algorithm or equation used to calculate the performance indicator (FESTA D2.1) may not take into account the performance indicator requirements and may return a value apparently valid even when applied to a data sample which does not fulfil the requirement of this performance indicator. Examples of these requirements are reported in the FESTA D2.1. As performance indicators may set requirements on data analysis, also the experimental protocol may do so. For example, the experimental protocol may set the requirements, in terms of age distribution, for the data analysis to applicable to different age groups. The experimental protocol may also set baseline and specific events which may then imply new requirements for data analysis. Furthermore, the definition of hypothesis and use case will set requirements on road geometry, weather condition, geography, etc…
In summary, many requirements on data analysis will be set during the design of the FOT (hypothesis formulation, use case definition, experimental protocol design, performance indicator calculation, etc…). Data quality analysis must keep track of these requirements and for each specific and different analysis determine the different level of reliability of the data.

Data quality implementation: subjective data

Subjective data is mainly data collected using questionnaires or interviews. If data is collected with personal interviews, the interviewer has an immediate control on data quality and can prevent data loss and out-of-scale value to be reported. If data is collected with a

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

questionnaire, especially a remote one such as a web-based questionnaire, data is more likely to be missing or unacceptable. For this reasons, the following paragraphs will mainly refer to data from questionnaires.

- Assessing and quantifying missing data
As stated above, no matter how well data collection and data storage process will be performed, some data is likely to be missing. For instance, some questions in the questionnaires may not have an answer or the answer may not be readable or, more simply, some questionnaires may be totally missing. In this case, quantifying data loss is very important to determine whether the data is biased. For example, it should be checked whether missing questionnaires result from a specific group or category of drivers and how this correlation may bias the whole analysis.

- Controlling data values are reasonable and units of measure are correct
Data quality analysis for subjective data is harder than for on-vehicle sensors data because, most of the time, there are not *right* or *wrong* answers and, as a consequence, there are not *right* or *wrong* values. However, when using a scale, values outside the scale are to be considered wrong and questions answered with out-of-scale values should be considered as missing data. Even if out-of-value data suggest that the questionnaire scale may not have been understood, there is no mean for the experimenter to know for sure whether the driver who filled in the questionnaire understood the scale or less. For this reason, extra caution should be used in the questionnaire design (FESTA D2.2, Chapter 3), showing example to clarify the questionnaire and making a pilot to assess the questionnaire clarity and completeness.

- Checking that the data dynamic over time is appropriate for each kind of measure
Same answering dynamics may make the experimenter suspicious. For instance, if a questionnaire asks to rate on a 1-10 scale one hundred statements depending on how right/wrong they are and all one hundred statements get the very same rate from one subjects, chances are that that data is not reliable. However, there is no scientifically sounding way for the experimenter to determine whether for instance, the subjects was just trying to rush throw the questionnaire. Nevertheless, attention while compiling the questionnaire can be enhanced by changing the direction of the positive and negative answers. For example in the Van der Laan's scale (Van der Laan, J.D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. Transportation Research - Part C: Emerging Technologies, 5, 1-10.) some positive attributes are to the right and other to the left, thus trying to keep up the reader attention. However, as said before, these symptoms of unreliability can only be prevented by making sure subjects are given enough information on how to complete the questionnaire and confidence that their data will be secret and used in a good way. This is the reason why, a personal contact with all subjects filling in the questionnaires is desirable even if its feasibility depends, most of the time, on the subjects' sample (FESTA – D2.2 Chapter 3.1.3).

- Guaranteeing that measures features satisfy the requirements for the specific data analyses
As for on-vehicle sensor data, analysis of subjective data should take into account all requirements set by the FOT during hypothesis formulation, use case definition, experimental protocol design, performance indicator calculation, etc…). Data quality analysis must keep track of these requirements along the process from the definition of the hypothesis up to the

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

last statistical analysis and, for each specific and different data analysis to be performed, determine the different level of reliability of the data.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6.2 Step 2 : Data processing

Once data quality has been established, the next step in data analysis is data processing. Data processing aims to "prepare" the data for addressing specific hypothesis which will be tested in the following steps of data analysis. Data processing includes the following sub-steps: filtering, deriving new signals from the raw data, event annotation, and reorganization of the data according to different time scale (Figure 4). Not all the above-mentioned sub-steps of signal processing are necessarily needed for all analyses. However, at least some of them are normally crucial.

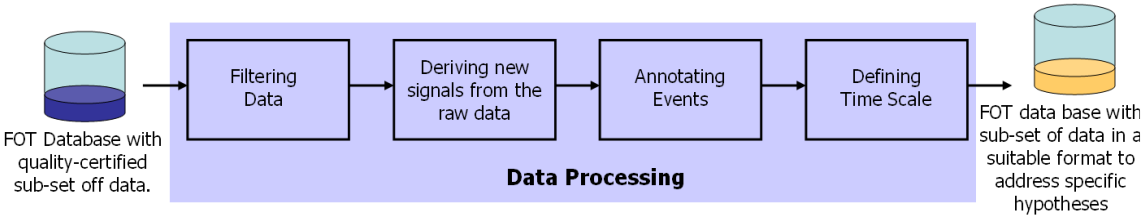


Figure 4 : Block diagram for the data processing procedure.

Filtering Data
This sub-step is aimed at selecting part of signals which are relevant for addressing the specific hypothesis to be tested. Data filtering can involve a simple frequency filter, e.g. a low-pass filter to eliminate noise, but also any kind of algorithm aimed at selecting specific parts of the signals. For example, for a specific hypothesis data may be relevant only when the vehicle is driving on a freeway. In this case, GPS data can be used to determine the time in which the vehicle was actually on a freeway and the other data of interest can be filtered accordingly.

Deriving new signals from the raw data
This sub-step is aimed at elaborating and/or combining one or more signals in order to obtain a new signal more suitable for the hypothesis to be tested. For example, if the analysis is interested in jerks, in this process, the acceleration will be derived to obtain the new signal jerk. Data analysis can also be interested, for instance, in the distribution velocity vs acceleration. In this later case, in this process, acceleration and velocity will be combined in order to obtain this new variable.

Annotating Events
This sub-step is aimed at marking specific time indexes in the data so that event of interest can be recognized. Event of interest can mark specific use cases (FESTA D3 Chapter 3) to be considered in the analysis as well as other events of interest for the hypothesis to be tested, such as crashes, accidents, and overtaking (as defined in FESTA D2.1). Especially if data is collected continuously (instead of on trigger), this process is fundamental to individuate the part of data which should be analyzed and to discard the part of data which is not of interest. Ideally, an algorithm should be used to go through all FOT data and mark the event of interest. However, the American FOT (such as the 100-Car Naturalistic Driving Study) showed that, especially when the data to be annotated is from a video and requires the

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

understanding of the traffic situation, writing a robust algorithm can be very challenging and manual annotation from an operator may be preferable.

Defining Time Scale

This sub-step is aimed at re-organizing data in the most suitable time scale for the specific hypothesis to be addressed in the following steps of the data analysis. In an FOT, data can be collected for several months. As a consequence, the experimenter can look at the data using very different time scales: minutes, days, weeks, months, maybe years. Depending on the different time scale a different data re-organization may be needed in order to cluster the data so that, for example, the trend of driver behaviour in the first week while using a new ADAS can be compared with the same trend over a 3-month time scale.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6.3 Step 3 : PI calculation

Whereas the Performance Indicator (PI) deliverable D2.1 and PI chapter of the FESTA handbook aim at *describing and compiling* performance indicators, this section is aimed at addressing issues related to the *implementation of the calculation* of performance indicators (PI) during the analysis process. The scale of the dataset and the uncontrolled variation in driving situations that occurs from driving freely with vehicles become a seriously limiting factor unless efficient calculation methodology is implemented. The choice of which PIs and hypotheses to calculate is clearly dependent on the amount of effort required. The large amount of PIs precludes a detailed discussion regarding the specifics of particular PI calculation. Therefore, this section will concentrate on key generic issues and recommendations which are common across a number of PIs. First some practical issues will be described, then statistical analysis considerations will be described.

Efficient calculation of PIs

Budget and other limitations of FOT projects limit the choice of PIs and hypotheses because the amount of effort and difficulty associated with calculating different PIs varies. Thus, methods for efficient calculation of PIs are crucial. Efficient methods are required to automatically sort through large amounts of data and calculate continuous PIs or find discrete PIs (e.g. near-crash and incident types of interest for the analysis of a safety system). For example, a PI calculation may want to find, for each different weather and road type, all relevant hard-braking incidents associated with and without a Forward Collision Warning (FCW) system, but yet excluding false alarms or hard-braking incidents that are irrelevant for FCW.

Although many PIs may be simply described and understood in a high level description, they can be quite complicated to calculate in implementation. Recall that five different types of measures are identified as components that are needed to calculate PIs (see PI handbook chapter and D2.1 for more details): Direct Measures, Indirect Measures, Events, Self-Reported Measures and Situational Variables. The above PI calculation example illustrates that calculation may involve an algorithm which specifies how to identify different traffic situations, may use a kinematic threshold to find a hard braking event, and may require video review to validate an event or situation.

PI calculation is typically implemented in a calculation algorithm or “script” in some software package (for example in Matlab, Excel, or SPSS) to calculate a PI from measures. Scripts are typically comprised of software code not only implementing the basic algorithm for calculation, but also implementing specialized code for exception management. Below, a number of recommendations are identified for successful implementation of calculation algorithms:

Development of automatic and efficient PI calculations. FOTs are not field experiments, they typically involve much larger amounts of data. The scale of the dataset and the large variation in driving situations because of less constrained driving become a seriously limiting factor unless well-structured automatic or semi-automatic calculations are implemented. Particular

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

efforts should be planned for implementing batch processing, data-subset selection and handling, and easy procedures for manual review.

PIs will be calculated on imperfect data. Exception management code is important to achieve a robust algorithm because often much of the data that is collected is not of perfect quality. Thus there is a strong need to create special solutions for "exceptions to perfect data". The prudent researcher should count on some data being poor because of the real-world conditions they were collected in (sensors will fail, human error is an issue, etc), but also because of the need to identify and exclude exceptional driving situations from calculation. In fact, four main categories of imperfect data have to be dealt with in the PI calculation: (1) good data, (2) poor data that can be used if reconstructed, filtered, or otherwise fixed, (3) poor data that should be excluded, and (4) missing data. Of the four, it is particularly important to identify poor data that should be excluded from analysis. Iterative interaction between PI calculation and data quality analysis and processing phases should be expected because some quality requirements for calculating values may only be encountered once calculations have commenced. Even if requirements on quality exist, the unfortunate reality is that the data may only partially meet these requirements.

PI calculation requires situation or context identification. As previously noted (e.g. D2.1), PI calculation requires a "denominator" to make a measure comparable. Risk estimation requires a "denominator" or exposure measures as well, in order to determine how often a certain event (e.g. lane departure) occurs *per something* (e.g. near-crash per left turn, lane departure warnings per drowsy event, or crash per rear-end crash type). These denominators may also involve a considerable amount of effort to correctly classify. Recall that a denominator makes a measure comparable (per time interval/per distance/in a certain location, etc). For example a denominator, such as a baseline time interval created to match the events found in the treatment condition, is needed to make a measure comparable (e.g. number of situations requiring a blind spot warning with and without a system). The fact that test exposure is largely uncontrolled (not tightly controlled as in experiments) means that analysis is largely conducted by first identifying the important contextual influences, and then performing the analyses to create a "controlled" subset of data to compare with. The identification of situation or context is sometimes an easy task (e.g. rain detected by a rain sensor), and sometimes quite difficult (e.g. defining an aggressive overtaking manoeuvre).

Critical events as a special case of PIs in FOTs. The ability to find and classify crash-relevant events (crashes, near-crashes, incidents) is a unique possibility enabled by FOTs to study direct safety measures. This possibility should be exploited by using a process of identification of critical events from review of triggered events. Experimental field or simulator studies or do not produce any useful amount of naturally-occurring critical event, whereas FOTs have been used for this purpose (e.g. in the US Volvo Trucks FOT, see Lehmer, 2007). Kinematic trigger conditions (e.g. Lateral acceleration >0.20 g, Following interval <0.5 s) , ABS activation) are used to find a list of *potentially* relevant critical events. Not all triggered events represent a true conflict. Irrelevant events are filtered out by manual review of video and/or by adding conditions (e.g. The host vehicle was in a curve [yaw rate >2 deg/s for 3 s] and the lead vehicle was stopped or on-coming, or The lead vehicle was in a different lane [lateral distance to target >2 ft]). Some events are straightforward and simple to identify, for example hard braking defined as peak deceleration >0.7g, and may not need to be saved as a discrete or transition variable. However, many events involve a considerable

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

amount of effort to find. The definition of these trigger values and the associated processes to filter out irrelevant events are of particular importance for enabling efficient analyses.

Statistical estimation of PIs

There are five kinds of data on which are calculated the performance indicators: direct measures logged from a sensor, indirect measures, events, self reported measures and situational variables. Care should be taken to use appropriate statistical methods to analyse the PIs. The methods used must consider the type of data and the probability distribution governing the process. The direct and indirect measures are considered as continuous stochastic processes sampled in time and space such as an instantaneous speed measured every half-second. The events are typically represented either by point processes such as a number of counts of accidents or incidents, or by stochastic processes taken values in a discrete "state" space, such as a series of on/off of a system. The self reported measures are qualitative variables, most often discrete coded values.

The statistical analysis of quantitative random variables such as speed follows the classical methods of statistical estimation of a probability distribution, which could depends on a set of parameters such as a mean and variance in the case of a gaussian distribution (Basawa,1980). By means of an histogram and correlogram, one has to identify the family of probability distribution governing the process. The dynamics of the process could be modelled by an autoregressive process if necessary.

The statistical analysis of the majority of performance indicators based on events draws on a Markovian or semi-Markovian formalism of changes in states over time (Taylor, 1994). For example in the LAVIA evaluation, the states relate to system modes, the activation of systems (in/out), the driving situation in the LAVIA zone, the state of operation/breakdown of the system, the display of a recommended speed and the use of the kick-down. The system and its environment develop over time. Changes of state occur randomly in accordance with the laws of probability. A homogeneous semi-Markovian process is described by a matrix of probabilities for transitions between states and by conditional distributions for the duration of one state knowing the next one. The first indicator derived from it is the rate of occurrence per unit of time (hour) or per unit of distance (kilometre), for instance the average number of kick-downs per kilometre or per hour. This measures the frequency of occurrence starting from any state of the vehicle system. The second indicator is either the duration of the state, or the distance travelled in that state. The distribution of probabilities for that length of time or distance is calculated with the help of a histogram for the density or the empirical average for that expectancy (Lassarre, Romon, 2006).

The statistical analysis of self reported measures relies on the estimation of frequencies issued from cross-tabulation of data. A statistical qualitative model such as a logistic model based on proportions of the modalities of a variable, for example the degree of acceptance of a system (in favour, hesitating, reject) is required to control situational variables. Multivariate analysis could be carried out by means of correspondence analysis or structural models in order to work with optimal scales as new performance indicators.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6.4 Step 4 : Hypothesis testing

Hypothesis testing in FOT takes the form of a null hypothesis : No effect of the system on a performance indicator, like the 85th percentile speed, against an alternative such as a decrease of x% of the performance indicator. To carry out the test, one relies on two samples of data with/without the system from which the performance indicator is estimated with its variance. Comparing the performance indicator of the two samples with/without intervention, whether independent (two groups of drivers S and A of size n_S and n_A) or paired (same driver with/without), is done using standard techniques such as a t-test on normally distributed data. A Student equality test of the theoretical averages in the case of Gaussian variables with identical variance is done through the statistic F

$$F = \frac{\bar{PI}_S - \bar{PI}_A}{\sqrt{\sigma^2\left(\frac{1}{n_S} + \frac{1}{n_A}\right)}}$$

An estimate of the variance can be calculated as a weighted average of the empirical variances of each sample. In the paired case, one works directly on $\Delta = PI_S - PI_A$ with $F = \frac{\bar{\Delta}}{\sqrt{\sigma^2/n - 1}}$. Even if the statistical units : driver; situation and measurement are embedded, we can consider, as an approximation, the variance only at the driver's stage, because the observations are clustered and consequently correlated.

Here the assumption is that there is an immediate and constant difference between the use and non-use of the system, i. e. there is no learning function, no drifting process, no erosion of the effect.

However, the assumption of a constant effect is often inappropriate. To get a complete view of the sources of variability and to handle the problem of serially correlated data, multi-level models are recommended (Goldstein, 2003). The performance indicator for each driver i, situation j and measurement k depends on a constant, a fixed effect of the system with a dummy variable T (=1 with the system, = 0 without), two random effects related to the driver and to the situation, some effects of explanatory variables Z and a residual:

$$PI_{ijt} = \mu + \lambda T_{ijt} + u_i + u_j + \beta Z_{ijt} + \varepsilon_{ijt}$$

The test of the impact (λ=0 in case of null hypothesis) is carried out by comparing the estimated parameter λ to its estimated variance. This model can be adapted to distributions different from a Gaussian distribution by means of generalized linear mixed model. An auto-correlation structure of the residual could be introduced if necessary. Other form of impact than immediate and constant can be tested by means of non linear function.

With such models, drivers or situations with missing data have to be included except in the case of nonignorable drop-out in case of Missing not at random. Elimination of drivers or situations because of missing data in order to keep complete data set may cause bias in the estimation of the impact.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

It is assumed that data will have been cleaned up in the data quality control phase. Nevertheless, to be sure that the estimation will be influenced minimally by outliers, one can use either robust estimates such as trimmed mean and variance or non parametric test such a Wilcoxon rank test or a robust MM regression (Gibbons, 2003) (Wasserman, 2007) (Lecoutre, 1987). Such tests provide protection against violation of the assumption of normal distribution of the performance indicator.

When a parametric approach is too fastidious in case of combination of PIs, a non parametric approach such as a bootstrap can be used to estimate the variance of the estimates.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6.4.1 Additional Step 4: Data mining

Data mining techniques allow the uncovering of patterns in the data that may not be revealed with the more traditional hypothesis testing approach. Such technique scan therefore be extremely useful as a means of exploratory data analysis and for revealing relationships that have not been anticipated. The data collected in a FOT is a huge resource for subsequent analysis, which may continue long after the formal conclusion of the FOT.

One relatively simple technique for pattern recognition is to categorise a dataset into groups. At the data analysis state, categorisation is normally made based on participants (e.g. gender, age, attitude, and personality traits etc); for example, the behavioural difference between male and female drivers in the presence of an ADAS system. Using dependent variables to categorise participants is more useful at the study design stage.

While dichotomous categories are valid for some variables, such as male vs female, or control vs treatment groups, or system A vs system B, there is generally little consensus about how best to split data according to continuous variables, such as age and driving experience etc. Mean or median split is prevailing when dichotomy is desired; for example, aged over 40 and under 40 years of age when 40 being the mean or median from the participants. However, when the number of required groups is more than two, cluster analysis is a commonly employed technique.

Cluster analysis tries to identify homogeneous groups of observations in a set of data according to a selected variable (e.g. demographic variables or performance indicators), where homogeneity refers to the within-group variation is minimised but the between-group variation is maximised. It is worth noting that when multiple variables are available for categorising the data into sub-groups, not all of the variables would necessarily lead to identical sub-groups. There are also different algorithms that could be used for cluster analysis, which may also lead to different sub-groups. Most commonly used methods for cluster analysis are k-means, two-step, and hierarchical clusters:

- *Hierarchical cluster analysis* allows the researcher to select a definition of distance as well as a linking method for establishing clusters, and determine how many clusters best suit the data. Hierarchical cluster analysis suits categorical variables.
- *K-means cluster analysis* allows the researcher to specify the number of clusters in advance and assign cases to the K clusters. K-means cluster analysis suits continuous variables.
- *Two-step cluster analysis* creates pre-clusters then clusters the pre-clusters. Two-step cluster analysis suits both categorical or continuous variables. Due to the nature of its algorithm, two-step clustering is capable of handling very large dataset.

Cluster analysis is a function available through many popular statistical packages, such as SPSS, SAS, and Minitab etc.

It is worth noting that data categorisation is an analysis technique to help achieving meaningful analysis but it does not guarantee transferability; i.e. the clustering method adopted for one dataset might not be applicable to another dataset due to the potential

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

difference in variances and distribution among different datasets. For example, it may work well by categorising a set of data collected from the UK part of a cross-country FOT into 3 groups for further analysis but 3 clusters might not always fit into the data collected from other counties.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

6.5 Step 5 : Global assessment

This section deals with the issue of identification of models and methodologies for generalise results from a certain FOT to a global level on traffic safety, environmental effects and traffic flow. One problem when generalizing results from a FOT is that it is often not known how close the participants in the FOT represent the target population. This leads to the situation that the estimates are biased and that it is very difficult to obtain valid variance estimators (Särndahl et al., 1992). If it is not possible to obtain a proper variance estimate it is recommended to perform a sensitivity analysis to estimate uncertainties. Furthermore, the problem of bias also need to be carefully considered by for example performing different comparisons such as compare speed and headway distributions of test drivers in the reference group to speed and headway of the traffic in general, e.g. measured by induction loops.
When generalizing results from a FOT, i.e. scaling-up FOT results to a more global level, it is often necessary to control for: usage, market penetration and compliance (the system might be switched off by the driver) and reliability of the system. Also total vehicle km driven by road type (motorway, rural, urban) and time of day is needed. The process of how to go from the FOT data to safety effects, traffic flow and environmental effects is illustrated in Figure 5. In this process two steps need to be taken. One is scaling up the FOT results, for example to higher penetration levels or larger regions. The other is to translate the results from the level of performance indicators (for example, time headway distribution) to the level of effects (for example, effect on the number of fatalities). For each type of effect there are (at least) two different ways of generalize the results: through micro-simulation or directly.

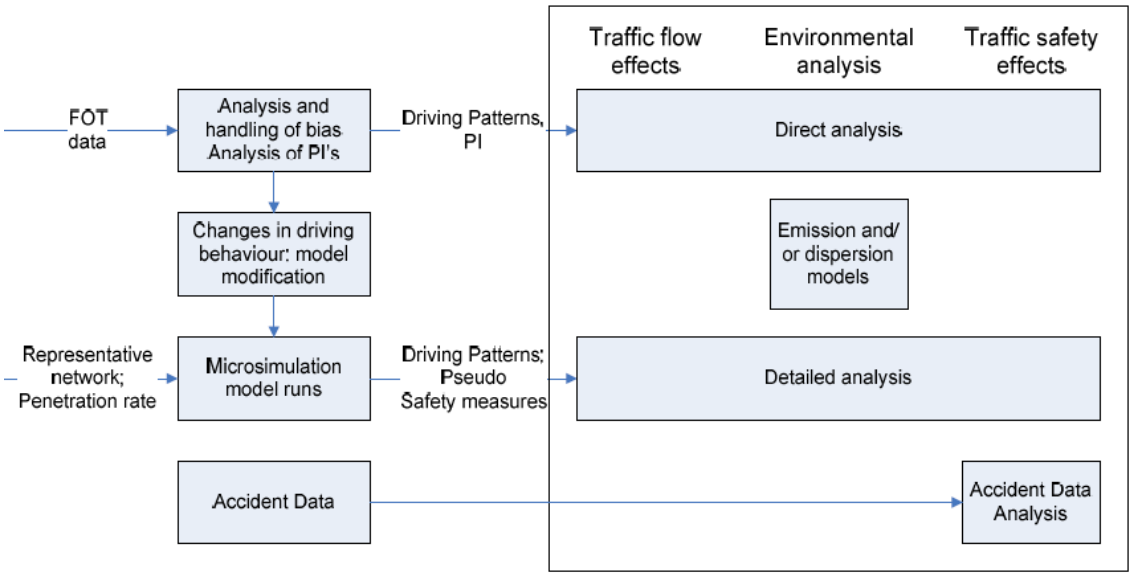


Figure 5 : Block diagram of scaling-up process.

The direct route includes both estimation directly from the sample itself and estimation through individual or aggregated models. Some advantages of the direct route are that it is rather cheap and quick. The alternative is micro-simulation, or in full: microscopic traffic simulation (i.e., a simulation model of traffic that has such a level of detail that individual

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

driver/vehicle units are being simulated). Advantages of micro-simulation are that they can be more reliable and precise and can incorporate indirect effects (like congestion).

Tools

There are now three tools with the following order from simple/cheap to complicated/ expensive. Direct sampling estimates and indirect model based estimates (described in Conceptual scaling models) and Micro simulation.

Conceptual scaling models

Computational models (not simulation) that directly relate FOT data to effects can be used for scaling up. Often this type of analysis is used for safety analysis and environmental analysis but can also be performed for traffic flow as illustrated in the figures above.

Traffic flow: Traffic flow impacts of FOT results can be estimated directly. The headway between vehicles is for example closely related to road capacity. FOT data that suggest changes in headway can therefore be scaled up to a road capacity effect. Similar considerations can be made for other possible FOT results. However, traffic flow effects of FOT results can be difficult to estimate with this direct method. There can for example be impacts on several parameters of importance for the traffic flow condition. Use of traffic simulation to aggregate effects on individual vehicles to the traffic system level is more appropriate in such cases.

Environmental: Estimation of environmental effects is an indirect model based process. By logged data from the FOT, for example driving patterns, exhaust emissions for the actual test vehicles can be estimated. Road conditions like gradients can also be included. There is a limit for such models describing effects for petrol engines with catalytic converters since only fuel consumption and CO_2 can be estimated with acceptable accuracy, one example of a model is VETO, Hammarström et al. (1989). For diesel engines the accuracy should be acceptable also for at least HC and NO_x. By use of micro-simulation models exhaust emissions can be estimated for other vehicles then the vehicles in the FOT.

Traffic safety: The desired measure in safety analysis is usually the reduction of the number of target accidents in the situation when the system is active compared to the baseline situation. Both direct and indirect model based methods are used for traffic safety. The effects can be estimated in several ways, Kulmala et al. (2007), Erke et al (2007):

1. Empirical evidence on safety impacts (verified results e.g. experimental design)
    a. Direct study of the reduction of the number of accidents for a certain accident type in the FOT. This direct path is often impossible to use since there might be no or rather few accidents to study.
    b. Study of the reduction of incidents/conflicts or near accidents; see for example Batelle (2007).
2. Expert evaluations of safety impacts (predicted results)
3. Indirect evidence of safety impacts such as changes in mean speed.

In Kulmala et al. (2007) a three step procedure is suggested to estimate the safety impact:

1. Describe the system
2. Assess how quickly the system will penetrate the market
3. Estimate the actual effectiveness of the system.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

Micro-simulation

There are two different types of micro-simulation:

1. Network level simulation. Due to their limited scale, FOTs will not show effects on a network level. Micro-simulation is the (probably only) tool to scale up FOT effects in which the effects of *penetration level* can be investigated. Micro-simulation results can be used to obtain traffic flow effects, environmental effects, and also safety effects (via so-called surrogate safety measures, see e.g. Gettman and Head, 2003). Typically, micro-simulation yields statistical data on the traffic flow that requires further analysis and interpretation, as well as further data sources, in order to find these effects. This is true in particular for safety and environmental effects. For example, for safety analysis, accident data are required to estimate effects on the number of fatalities in a certain region and period.
    - The micro-simulation package has to allow far-reaching adaptation of vehicle and driver models, so that the changes in behaviour that was found in the FOT can be implemented on the level of individual driver-vehicle units. These could be changes in speed control behaviour, car-following behaviour, lane change behaviour, etc. This model adaptation typically requires (1) interpretation of the FOT results, (2) translating the effects in terms of the driver/vehicle model components (=model specification), (3) implementation (including verification), (4) conducting the actual model runs.
    - For network level it is typically impossible to analyze the whole network (e.g. because of lack of data or computing power). Hence we need a network that is representative for the population, regarding road layout, normal driver behaviour, traffic load, etc. A problem is that real traffic data are not easily obtained! Some possibilities: (1) Induction loops, which give speed and time headway, are usually only available on motorways; (2) Incidental measurements from helicopters, cameras, OBUs (On Board Units) etc.

2. Focussing on safety effects, an alternative simulation approach is to go to a more detailed level (accident sites) and analyse these use cases. In this setting, goal of micro-simulation is first to accurately reproduce (reconstruct) actual accidents in the simulation model, and second to test if the FOTs ITS systems would change the outcome. For this approach, highly detailed accident scenarios/use cases are needed, as well as high-fidelity simulation models

Models

Traffic micro-simulation models consider individual vehicles in the traffic stream. There is consequently a potential to incorporate FOT results in the driver/vehicle models of the simulation. Impacts on the traffic system level can then be estimated through traffic simulations including varying percentages of system-equipped vehicles.
The basic result of a traffic micro-simulation model run is a set of trajectories for the vehicles that have traversed the studied road network during the simulated time. Traffic flow, road safety and environmental impacts can in the following analysis be studied based on these resulting vehicle trajectories.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

Here we describe some various aggregated and individual models that can be used, i.e. effect models to convert for instance speed to safety effects (power model) or changes in speed variation or headway to traffic flow.

Models for traffic flow

FOT results are with advantage scaled up to traffic flow impacts using micro-simulation models. The modelling detail of traffic micro-simulation does however place restrictions on the practical size of the simulated road network. Macroscopic or mesoscopic traffic models combine the possibility to study larger networks with reasonable calibration efforts. These models are commonly based on speed-flow or speed-density relationships. Large area impacts of FOT results can therefore be estimated by applying speed-flow relationships obtained from micro-simulation for macro- or mesoscopic traffic modelling.

Examples of traffic micro-simulation models include AIMSUN [1] and VISSIM [2], for urban or motorway road networks and TWOSIM, Kim et al (2007) and RuTSim, Tapani et al. (2007) for rural highways. PARAMICS [3] is another example. DYNAMEQ [4] and CONTRAM [5] are mesoscopic traffic models. EMME [4] is a commonly applied macroscopic traffic model.

Environmental models

Exhaust emissions from road traffic is a most complex process to describe. Models for exhaust emissions in general include three parts: Cold start emissions, hot engine emissions and evaporative emissions.

An exhaust emission model can roughly be described as: ***Σ(Traffic activity) x (Emission factor)=Total emissions***

Of course traffic activity data then has a high correlation to total emissions. Traffic activity data includes: mileage, engine starts and parking.

Mileage is of importance in two ways, both the total value and the distribution on traffic situations. Parking is of importance both for cold start and for evaporative emissions.

In most cases total exhaust emissions per substance will increase when mileage and engine starts increase but not for sure. If the emission factors would decrease in parallel total emissions could decrease. An ICT can influence both traffic activity and emission factors per traffic situation. In order to estimate ICT effects on this level there are two possibilities: exhaust emission measurements or use of micro simulation models.

The most accepted and used exhaust emission models on an EU level should be ARTEMIS, Keller et al. (2007) and COPERT, Kourdis et al (2000). These models are used from a local up to a national level. They should include most substances of interest. In order to use this kind of model one needs input data for the level of evaluation. In addition to traffic activity data one needs data for: the vehicle fleet; road network; meteorological conditions; fuel quality etc. If the driving pattern is influenced per traffic situation such data for the FOT vehicles are directly available. In order to estimate driving pattern changes for all vehicles per traffic situation traffic micro simulation models could be used. In order to estimate emission factors for these alternative driving patterns there is need for exhaust emission measurements or exhaust emission models on an individual level. Examples of individual models are VETO (Hammarström et al. (1989)) and PHEM (Rexeis (2007).

In order to estimate socio economic costs (see deliverable2.6) for exhaust emissions there are values for different substances available, these values may vary between countries. One example from Sweden is: SIKA (2005). These values represent all types of costs caused of exhaust emissions. If such values are available there is no need for dispersion models.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

In the TAC SafeCar project, Reagan et al.(2006) a Positive Kinetic Energy (PKE) model was used to estimate fuel consumption and vehicle emissions.

Safety models

Speed has a close relation to safety. The speed of a vehicle will influence not only the likelihood of a crash occurring, but will also be a critical factor in determining the severity of a crash outcome. This double risk factor is unique for speed. The relationship between speed and safety can be estimated by various models such as the Power model, Nilsson (2004), that estimate effects of changes in mean speed on traffic crashes. It suggests that a 5% increase in mean speed leads to approximately a 10% increase in crashes involving injury and a 20% increase in those involving fatalities. More examples of models for speed-safety relationship are reviewed in Aarts and van Schagen (2006). In general it is important to consider under which assumptions the models are valid, the Power model for example is valid under the assumption that mean speed is the only factor that have changed in the system. Therefore these models are more suitable for FOTs with systems mainly dealing with speed.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

References

Aarts L. & van Schagen, I. (2006) Driving speed and the risk of road crashes: A review. Accident, Analysis &Prevention, 38, Page 215-224.

Batelle (2007) Final report Evaluation of the Volvo Intelligent Vehicle Initiative Field Operational Test Version 1.3

Cochran W. G. (1977). Sampling techniques. 3rd edition, Wiley & Sons, New-York.

Erke, A., Veisten, K., Elvik, R. (eds) (2007?) Cost-benefit analysis and cost effectiveness analysis. In-Safety Deliverable D5.2

Everitt B., Dunn F. (2000), Applied multivariate data analysis, Arnold Publication.

Gibbons J. D. and Chakraborti S. (2003). Nonparametric Statistical Inference, 4th Ed. CRC.

Goldstein H. (2003). Multilevel statistical models. 3rd edition, Arnold, London.

Hammarström, U. and Karlsson, B. (1989) VETO - a computer program for calculation of transport costs as a function of road standard. VTI meddelande 501. Swedish Road and Traffic Research Institute. Linköping.

Keller, M. and Kljun, N. ARTEMIS Road Emission MOdel. Model description. (EU Commission – DG Tren – Contract 1999-RD.10429). Workpackage 1100 – Model version 04c. Deliverable 13. INFRAS. Bern.

Kim, J. and L. Elefteriadou (2007), Capacity Estimation for Two-Lane Two-way Highways Using Simulation, In proceedings of the 86th TRB meeting.

Kouridis, C & Ntziachristos, L & Samaras, Z: (2002) COPERT III: Computer programme to calculate emissions from road transport. Users manual (Version 2.1). European Environment Agency. Technical report No 50. Copenhagen.

Kulmala, R., Rämä, P., Schirokoff, A., Sihvola, N. (2007) Safety effects of co-operative intelligent vehicle safety system. eIMPACT

Lassarre, S. et Saad, F. (2006). Présentation générale du dispositif expérimental: justification des choix. In. *Carnet de route du LAVIA. Limiteur s'adaptant à la vitesse autorisée*. Paris : Actes du colloque LAVIA, pp. 11-17.

Lassarre, S., Romon, S. (2006). Utilisation du LAVIA et influence sur les vitesses pratiquées en vue de l'évaluation de l'utilité. In. *Carnet de route du LAVIA. Limiteur s'adaptant à la vitesse autorisée*. Paris : Actes du colloque LAVIA, pp 53-60.

Lebart L., Morineau A., Piron M. (1997), Statistique exploratoire multidimensionnelle, Dunod, Paris.

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

Lecoutre J. P., Tassi P. (1987). Statistique non paramétrique et robustesse. Econometrica, Paris.

Michon, J.A. (1985) A critical view of driver behaviour models: what do we know, what should we do? In L. Evans and R.C. Schwing (Eds.) *Human Behaviour and Traffic Safety* (pp. 485-524). New York: Plenum Press.

Nilsson, G. (2004) Traffic safety dimensions and the power model to describe the effect of speed on safety. Bulletin 221, Lund Institute of Technology, Lund Univiversity.

Reagan, M., Triggs, T., Young, K., Tomasevic, N., Mitsopoulus, E., Stephan, K., Tingvall, C. (2006) On-road Evaluation of Intelligent Speed Adaptation, Following Distance Warning and Seatbelt Reminder System: Final Results of the TAC SafeCar Project. Final Report. Monash University.

Rexeis, M. (2007) In-Use Fahrzeugtests an einem schweren Nutzfahrzeug und Erstellung der Eingabedaten zur Berechnung der Emissionsfaktoren mit dem Modell Phem. TUG, SECTION: Thermodynamics and Emissions Research - Emission. Graz. 2007.

Saad, F. (1997). Contribution of observation and verbal report techniques to an analysis of road situations and drivers' activity. In T. Rothengatter & E. Carbonell vaya (Eds), *Traffic and Transport Psychology, Theory and application* (pp 183-192). Pergamon.

Saad, F (2006). Some critical issues when studying Behavioural Adaptations to new driver support systems. *Cognition, Technology & Work*, 8, 175-181.

Saad, F. and Dionisio, C. (2007). Pre-evaluation of the "Mandatory Active" LAVIA: assessment of usability, utility and acceptance. In proceeding of the 14th world congress & exhibition on Intelligent Transport Systems and Services. 8-12 October 2007, Pekin, Paper 2257.

SIKA PM 2005:16. Kalkylvärden och kalkylmetoder (Arbetsgruppen för samhällsekonomiska kalkyler ASEK). En sammanfattning av verksgruppens rekommendationer 2005

Särndahl, C-E., Swensson, B., Wretman, J. (1992) Model Assisted Survey Sampling. Springer Verlag.

Tapani, A. (2005), Versatile Model for Simulation of Rural Road Traffic, Transportation Research Record 1934.

Taylor H., Karlin S. (1994), An introduction to stochastic modeling, Academic Press.

Wannacott T. and Ronald Wannacott R. (1990) , *Introductory Statistics for Business and Economics*, 4th edition, John Wiley & Sons.

Wasserman L. (2007). All of Nonparametric Statistics, Springer.

[1] www.aimsun.com
[2] www.ptv.de

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.

[3] www.paramics-online.com
[4] www.inro.ca
[5] www.contram.com

The FESTA Support Action has been funded by the European Commission DG-Information Society and Media in the 7th Framework Programme. The content of this sole responsability of the project partners listed herein and does not necessarily represent the view of the European Commission or its services.