Accident Analysis and Prevention 58 (2013) 309–317

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Chunking: A procedure to improve naturalistic data analysis

Marco Dozza[a,*], Jonas Bärgman[a], John D. Lee[b]

[a] Chalmers University of Technology, Applied Mechanics Dept., Sweden
[b] University of Wisconsin-Madison, Industrial and Systems Engineering, USA

ARTICLE INFO

Article history:
Received 27 September 2011
Received in revised form 7 March 2012
Accepted 15 March 2012

Keywords:
Accident causation
Impact assessment
Active safety
Intelligent transportation systems
Traffic and vehicle safety
Naturalistic data analysis
Field operational test

ABSTRACT

Every year, traffic accidents are responsible for more than 1,000,000 fatalities worldwide. Understanding the causes of traffic accidents and increasing safety on the road are priority issues for both legislators and the automotive industry. Recently, in Europe, the US and Japan, significant public funding has been allocated for performing large-scale naturalistic driving studies to better understand accident causation and the impact of safety systems on traffic safety. The data provided by these naturalistic driving studies has never been available before in this quantity and comprehensiveness and it promises to support a wide variety of data analyses. The volume and variety of the data also pose substantial challenges that demand new data reduction and analysis techniques. This paper presents a general procedure for the analysis of naturalistic driving data called chunking that can support many of these analyses by increasing their robustness and sensitivity. Chunking divides data into equivalent, elementary chunks of data to facilitate a robust and consistent calculation of parameters. This procedure was applied, as an example, to naturalistic driving data from the SeMiFOT study in Sweden and compared with alternative procedures from past studies in order to show its advantages and rationale in a specific example. Our results show how to apply the chunking procedure and how chunking can help avoid bias from data segments with heterogeneous durations (typically obtained from SQL queries). Finally, this paper shows how chunking can increase the robustness of parameter calculation, statistical sensitivity, and create a solid basis for further data analyses.

1. Introduction

In the US, more than 34,000 fatal motor-vehicle crashes occurred in 2008, corresponding to almost 18 fatalities per 100,000 licensed drivers (NHTSA, 2009). In Europe, the number of fatalities amounted to almost 39,000 in the same year (CARE, 2010). There are many ways to increase safety on our roads. For instance, performing research to understand the underlying causes of accidents can guide the development and legislation of appropriate countermeasures such as intelligent vehicle active safety systems (IVSs). The focus on the development and evaluation of the effect of IVSs is intensifying all over the world. Naturalistic driving studies are increasingly being used both to evaluate IVSs and to better understand what causes accidents. Dingus et al. (2006) defines naturalistic in this context as “Unobtrusive observation; observation of behavior taking place in its natural setting.” Typically, naturalistic driving studies rely on the collection of data from instrumented vehicles used by their drivers in their daily lives. The data collected often comes from many different types of data sources. Data sources can range from relatively simple accelerometers and GPS to such dissimilar sources as lane tracking cameras, vehicle tracking radar, as well as driver-state sensing such as eye-tracking systems. In a naturalistic driving study, data collection duration per driver ranges from a few weeks (Fancher et al., 1998; Leblanc et al., 2006; Najm et al., 2006; Reagan et al., 2006; Sayer et al., 2008) to several months or years (Hjälmdahl, 2004; Neale et al., 2005; Reagan et al., 2006; Carsten et al., 2008; euroFOT-Consortium, 2010). Such naturalistic data combine to form peta-scale databases that provide a unique window into the factors influencing driver behavior, but these databases also pose substantial challenges for analysis.

Analyzing data from naturalistic driving is complicated by the diversity of driving situations and trip types (Boyle et al., 2009; Victor et al., 2010). Compared to data collected in a simulator or field study, there is no experimental protocol that defines and regulates the driving situations. Consequently, the data collected is heterogeneous with respect to a number of variables such as weather, lighting, driving situations (e.g., traffic density), driver state (e.g., drowsiness), and vehicle dynamics (e.g., velocity). To analyze data from such diverse driving situations, these variables must be separated into their different states so that specific driving situations can be extracted from the data. For example, a velocity threshold can isolate a condition in which an IVS could potentially be active; for a lane departure warning system (LDW) (USDOT, 2005); this threshold would be 60 km/h for most of the systems now on the

[*] Corresponding author.
E-mail address: marco.dozza@chalmers.se (M. Dozza).

0001-4575 © 2012 Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.aap.2012.03.020

market. Selection conditions fragment the data into long and short segments of continuous time, also within a trip. Converting data fragmented into segments of heterogeneous size and variable states creates several challenges for data analysis such as the calculation of robust parameters to describe IVSs performance and assess crash causation.

Data fragmentation is a prevalent challenge. The European FESTA project (Festa-Consortium, 2008b) created guidelines for field operational tests that are currently followed by the major European field operational tests, such as euroFOT (euroFOT-Consortium, 2010) and teleFOT (teleFOT-Consortium, 2010). These guidelines list parameters (called performance indicators) for field operational test use (Festa-Consortium, 2008a). Approximately one third of the objective safety-related parameters proposed by FESTA are vulnerable to problems posed by data fragmentation.

The current analyses of naturalistic data have avoided substantial errors that fragmentation can cause. One reason for this success is that only a few experts have analyzed data from naturalistic studies because data access has been very restricted. In most analyses, data fragmentation and resultant issues were handled appropriately, or parameters that were robust with respect to fragmentation were used. Naturalistic data analyses and the number of analysts will grow significantly over the next few years, chiefly through the US Strategic Highway Research Program 2 (SHRP2, 2010), which will make naturalistic data available to many researchers. For this reason, it is important that methods that enhance comparability and robustness of naturalistic data analysis are developed and adopted soon. This paper presents a procedure that facilitates the calculation of robust parameters extracted from continuous naturalistic data. Such data procedure facilitates the analysis of naturalistic data, which is intrinsically diverse (e.g., in terms of trip durations, driving situations, and driver behavior).

This paper discusses fragmentation, which is intrinsic to naturalistic data analysis; the paper also presents a procedure for the analysis of fragmented data from quasi-experimental studies such as naturalistic driving studies. The procedure is called chunking and, in this paper, it was applied to one specific step in hypothesis testing (i.e., calculation of parameters in treatment conditions) to show how to apply the method when testing hypotheses. We believe that this procedure can support the development of robust and comparable methods for parameter calculations on naturalistic data. Also, this procedure can help new naturalistic data users to avoid biases due to fragmentation from basic SQL queries that may lead to improperly calculated parameters.

2. Methods

2.1. Data

A total of approximately 1142 h of naturalistic driving were collected from the Swedish national field operational test methodology project SeMiFOT (Victor et al., 2010). The data were collected from 14 drivers aged 45.5 ± 9.2 years (mean and SD), who had held driver licenses for 27.4 ± 9.2 years (mean and SD). Fifty percent of the drivers were women and fifty percent men. The data were collected over a period of approximately six months, primarily in the region of Västra Götaland, Sweden. In 49% of the 1142 h, velocity was below 50 km/h, in 16% between 50 and 70 km/h, in 18% between 70 and 90 km/h, and in 17% above 90 km/h. The thresholds 50, 70, and 90 km/h are standard speed limits on Swedish roads.

Seven Volvo Car Corporation leased vehicles were used by study participants as private cars in their everyday driving. A total of approximately 270 signals from different data sources were collected continuously from each vehicle. The data sources included GPS, vehicle controller area network (CAN) bus (ISO, 2003) video, and accelerometers, as well as eye tracker and extra lane tracker. Data was collected on a per trip basis, i.e. from engine start to engine stop. After data was collected on hard drives inside the vehicles, it was transferred to SAFER (SAFER, 2010) for processing and database upload. Prior to uploading the data onto an Oracle^{TM} SQL database, it was synchronized and re-sampled from its original frequency to 10 Hz. Analysis was performed using SQL queries combined with Matlab^{TM}.

2.2. Analysis procedures

In this paper, the two signals *velocity* and *lane offset* from the vehicle CAN bus were extracted from the database and analyzed following the procedures in Fig. 1. These signals were chosen because they relate to longitudinal control (speed selection and maintenance) and lateral control (lane keeping and curve negotiation) and are thus central for safety analysis. Specifically, lane offset is a signal used by Lane Departure Warnings (LDW). Velocity is a typical selection factor for naturalistic driving data because it can be used as a surrogate measure for road type, traffic density (in relation to posted speed limits), and safety (Nilsson, 2004; Cameron and Elvik, 2008; Turner-Fairbank Highway Research Center, 2010). Typical parameters, calculated using velocity and lane offset for naturalistic driving studies, are mean velocity (MV; (Hjälmdahl, 2004; Leblanc et al., 2006; Najm et al., 2006; Reagan et al., 2006; Carsten et al., 2008; Sayer et al., 2008) and standard deviation of lane position (SDLP; Orban et al., 2006; Alkim et al., 2007; Festa-Consortium, 2008a), respectively. These parameters have been used to evaluate longitudinal and lateral control and often serve as safety indicators. In this paper the term *parameter* refers to an indicator, calculated from naturalistic driving signals that is used for data analysis and, more specifically, for hypothesis testing. However, the results presented in this paper extend, as it will be clarified in the discussion, to other parameters such as those presented in FESTA (Festa-Consortium, 2008a).

This paper also refers to *segments* as intervals of continuous data that fulfill a specific criterion for data extraction such as an SQL query. More specifically, each trip in this study was divided into segments of time-continuous data (10-Hz sample rate) in which velocity was above 70 km/h (velocity threshold criterion from our SQL query). Each segment starts from the first occurrence of a sample above 70 km/h and ends when it falls below 70 km/h again (as shown in Fig. 2). Our selection criterion can be expressed with the following pseudo SQL query: SELECT velocity AND LaneOffset FROM All Trips WHERE velocity >70 km/h. After running this query, segments are individuated by finding sections of continuous (10 Hz) data. The threshold of 70 km/h was chosen because it was compatible with IVSs activation thresholds, and would thus be a valid condition in the evaluation of such a system in a field operational test. Furthermore, in Sweden 70 km/h is also the posted speed limit that divides urban and rural roads. The segments individuated in the process described above may have a length from a single sample (0.1 s) up to an entire trip that might last several hours (minus the time for accelerating to above 70 km/h and decelerating to 0 km/h). The number of segments in each trip may vary (zero to hundreds) depending on how long a trip was and how many times the driver crossed the threshold. Segments are key components in the analysis presented in this paper (step 1 in Fig. 1a). Instead of segmentation in time, other variables such as velocity or distance can be used depending on the analysis focus. The sequence of typical steps for data processing for hypothesis testing for naturalistic data is then according to Fig. 1: step 1, data fulfilling specific conditions are extracted using an SQL query, producing segments; step 2, the parameters are calculated for each of the segments individually, and step 3, all segments for a condition are merged into combined parameters.

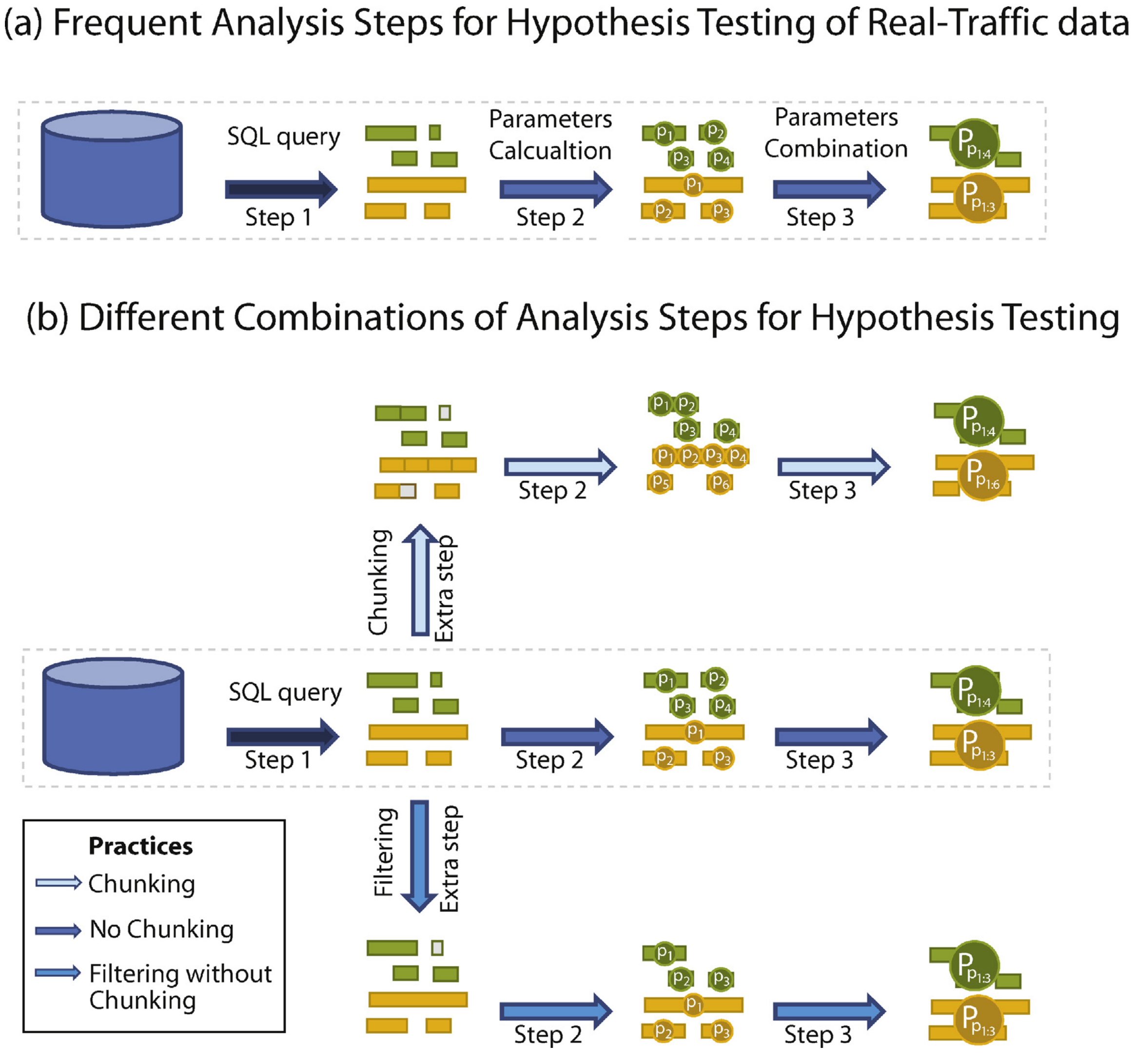

Fig. 1. (a) Three conventional steps followed in testing hypotheses on naturalistic driving data. Step 1: an SQL query retrieves the data; data is normally sorted according to two (or more) conditions (indicated with different colors in the figure) and consist of segments of different size. Step 2: parameters (such as mean velocity; indicated with pn in the figure) are calculated for individual segments of data. Step 3: parameters are combined and presented for the two (or more) conditions. (b) Three different practices for data analysis: the conventional practice from Panel A and the two alternatives discussed in this paper which add to one extra step for chunking and filtering to the conventional practice.

This paper introduces a new procedure to facilitate the transition between steps 1 and 2 in naturalistic driving data analysis (Fig. 1) called *chunking*. Chunking alone does not guarantee that baseline and treatment data are appropriately chosen. However, chunking may create the basis for a matching baseline data (Guo, 2009) since, later on in step 3 chunks from baseline and treatment can be matched accordingly to several vehicle, driver, or environment-related variables. This paper does not consider the baseline or treatment in steps 2 and 3 (Fig. 1), it only presents the chunking method to demonstrate its basic rationale and usefulness.

Chunking consists of the division of data within a segment into a number of sections of continuous data of equal length (Figs. 1 and 2). The nature and size of the chunks is decided by the analyst, and this paper presents an example of how these important decisions should be made. Fig. 2 shows, using a representative trip, (1) how sections of data are returned from an SQL query and segments are identified; (2) how segments are divided into chunks of equivalent duration, and (3) why some data is discarded when chunking. When segment duration does not add up to an integer multiple of the chunk duration, the *residuals* are discarded (Fig. 2). Also, segments of duration shorter than the chunking size are discarded (Fig. 2). To avoid bias in the analysis due to the removal of chunking residuals in the last part of the trip alone, the first chunk is started at a random point from the start of the segment (Fig. 2). The random start time is in the range between zero and the length of the residual for a segment.

This paper also discusses the effect of discarding segments that are too short after the SQL query without using chunking. This procedure is referred to as *filtering* (Fig. 1b). The data discarded by filtering is referred to as discarded data for *too-short segments*. Because both filtering and chunking discard segments shorter than a threshold, the amount of data discarded can be compared.

2.3. Parameter calculations

Mean velocity (MV) and standard deviation of lane position (SDLP) parameters were computed after obtaining data from different analysis steps (Fig. 1a and b). These parameters were calculated for the three analysis practices proposed in Fig. 1b, namely no-chunking, chunking, and filtering without chunking. Chunking was implemented to obtain chunks of data of equivalent duration corresponding to 5, 20, 60, 180, and 600 s. No-chunking with filtering

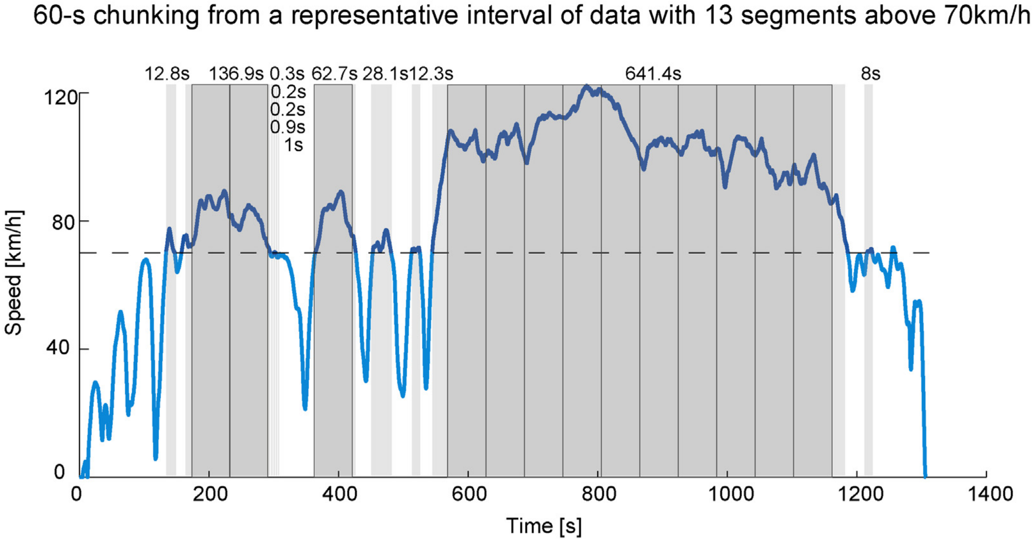

Fig. 2. 60-s chunking for one representative trip containing 13 segments with velocity above 70 km/h. The 13 segments are indicated with a darker shade and their durations are reported at the top of the figure (durations of 5 of the 13 segments were stacked on top of each other because of their short duration). 60-s chunking is shown for the 3 segments with duration above 60 s with darker bars. Light shades show the discarded data, both due to too-short segments and to chunking residuals.

was implemented for segments of 5, 20, 60, 180, and 600 s duration. After calculation of MV and SDLP for each segment or chunk (depending on the condition), two combined parameters were computed, namely mean MV and median SDLP. We combined MV and SDLP using different functions (mean vs median) to be able to discuss the interaction between the chunking procedure and the statistical properties of the parameter distribution. Further, the mean of MV is often used as a safety-risk indicator, and the median of the SDLP constitutes a more robust indicator of central tendency than the mean SDLP, because SDLP is more prone to extreme values. In the no-chunking condition, an additional calculation was performed to obtain a weighted average of MV and median of SDLP, with respect to segment duration (time). That is, mean MV and median SDLP for each segment was multiplied by the segment duration and divided by the sum of the duration of all segments before applying average and median functions respectively. In addition, the *average segment/chunk duration* (defined as the average length in time of the segments/chunks used for each specific parameter calculation) was calculated. In addition, the *number of data segments* (defined as the number of continuous pieces of data that the parameters were calculated on before being combined) was calculated. *Computation time*, defined as the percentage of time taken to complete analysis steps, in relation to the without-chunking condition, was also calculated. Finally, *total used time* (defined as the sum of all chunks durations used in the parameter calculation, after data discarded due to chunking and too-short segments were removed) was calculated.

2.4. Test for dependent observations

Autocorrelation analysis was used to determine the extent to which chunking created dependent observations. More specifically, segments containing more than 20 chunks were used to create sets of different chunk sizes. Analysis of autocorrelation of each such segment was pooled across all such segments for the different chunk sizes. This analysis was used to estimate the dependency across chunks and determine the extent to which traditional statistical analysis (which requires independence of observations) was still possible to apply after chunking. Autocorrelation analysis was performed for MV, SD of velocity, SDLP, and mean lane position. The SD of velocity and mean of lane position was chosen specifically in this analysis as a complement to MV and SDLP to show the effect of mean and median on autocorrelation results.

3. Results

Our sample query, selecting data from the SeMiFOT database for velocities above 70 km/h, returned 15,729 data segments with a mean duration of 91 s corresponding to 399 driving hours. The duration of these segments ranged from fractions of seconds to hours. Fig. 3 shows the distribution of segment durations from this query using a logarithmic scale. Time-wise, shorter segments occurred more often than longer segments. However, most of the data was in the longer segments; Fig. 3 shows how the cumulative distribution of time precedes the cumulative distribution of duration. In fact, 75% of the data was comprised of segments longer than 100 s even though 75% of the segments were shorter than 100 s (Fig. 3).

Average MV and median SDLP were strongly affected by chunking as well as by weighting and filtering. The mean MV without chunking or weighting was lowest: 77.9 km/h compared to a weighted average MV of 93.2 km/h, clearly indicating the bias from shorter segments in the simple average (Fig. 4a; Table 1). Chunking provided MV estimations proportionally higher the longer the

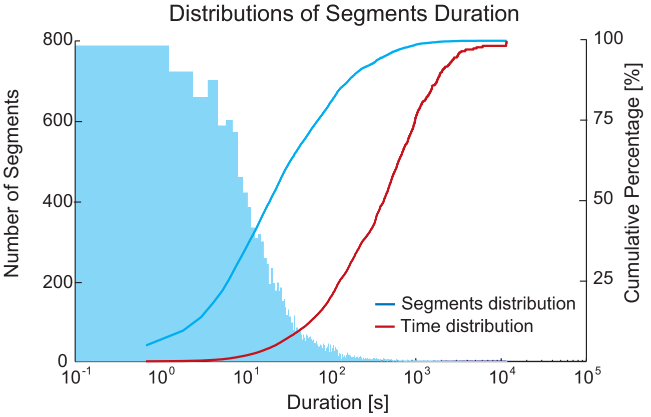

Fig. 3. Distribution of data segments from a simple query selecting data at 70 km/h. Cumulative percentage distributions are also reported for duration (time) and number of segments.

Table 1
Effect of chunking and weighting on parameters calculation, computation time, and data loss.

	Without chunking		With chunking				
	No weighting	Weighting	5-s chunks	20-s chunks	60-s chunks	180-s chunks	600-s chunks
Average segments length (s)	91.3	91.3	5	20	60	180	600
Average mean velocity (km/h)	77.9	93.2	93.8	95.0	96.8	99.6	102.9
Median SDLP (m)	0.271	0.332	0.102	0.211	0.299	0.35	0.353
Number of data segments	15,729	15,729	279,518	65,074	18,846	4681	745
Computation time rel. to no chunking (%)	100.0	100.0	1777.1	413.7	119.8	29.8	4.7
Discarded data for too-short segments (%)	0.0	0.0	0.5	4.3	12.8	31.7	61.1
Discarded data for chunking residuals (%)	0.0	0.0	2.1	5.0	8.4	9.6	7.8
Total used data (h)	399	399	388	362	314	234	124

Table 2
Effect of filtering without chunking on parameters calculation, computation time, and data loss.

	Filtering without chunking				
	5-s filter	20-s filter	60-s filter	180-s filter	600-s filter
Average segments length (s)	112.1	175.6	289.2	558.5	1188
Average mean velocity (km/h)	79.6	83.6	88.6	95.3	101.6
Median SDLP (m)	0.317	0.366	0.387	0.395	0.387
Number of data segments	12,698	7823	4328	1757	470
Computation time rel. to no chunking (%)	80.7	49.7	27.5	11.2	2.3
Discarded data for too-short segments (%)	0.5	4.3	12.8	31.7	61.1
Discarded data for chunking residuals (%)	0.0	0.0	0.0	0.0	0.0
Total used data (h)	397	382	348	272	155

chunk (the range was 93.8 km/h and 102.9 km/h corresponding to 5 s and 600 s chunks, respectively). This relation between chunk size and MV was anticipated considering that higher velocities require longer time due to vehicle dynamics. Filtering without chunking provided lower MV estimations compared to chunking, which suggests short segment bias cannot be eliminated with filtering (Table 2).

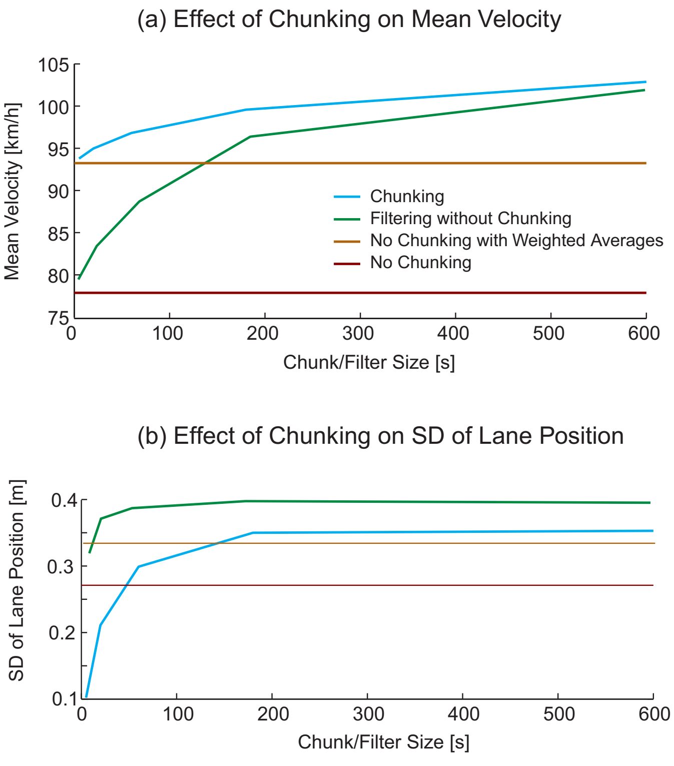

Fig. 4. (a) Average mean velocity as a function of chunking and filtering size. Average mean velocity is also reported for no-chunking and no-chunking with weighted averages. It is worth noticing that chunking with elementary (i.e. one-single data point) chunks is equivalent to no-chunking with weighted averages. (b) Median of standard deviation of lane position as a function of chunking and filtering size. Median of standard deviation of lane position is also reported for without-chunking and weighted medians without-chunking.

Chunking and chunk size also affected median SDLP. Larger chunks resulted in higher SDLP values, suggesting parameters such as road curvature or driver maneuvers such as corner-cutting, influence SDLP. Fig. 4b shows the effect of chunking on median SDLP. Median SDLP without chunking had values between the SDLP of 20 s and 60 s chunking (Table 1). Median SDLP was also calculated with weighting and provided an estimation similar to 180 s chunking. Filtering without chunking increased median SDLP compared to chunking, confirming that segment size affects SDLP. Further estimations of SDLP from filtering appear to be significantly higher than those obtained from all other methods followed in this paper.

Chunk size affects the number of available parameter values for statistical analysis. As shown in Table 1, the number of parameter/segment values available for statistical analysis with 5 s chunking was 279,518; this number decreased as chunks size increased finally reaching 745 for 600 s chunks (Table 1). Without chunking, the number of available parameter values for statistics was only 15,729 (lower than with any chunking below or equal to 60 s; see Table 1). Thus, 60 s or smaller chunking may result in higher power for the subsequent statistical analyses compared to no-chunking when controlling for multiple observations. However, in general, the smaller the chunks, the higher the power, but also the greater probability of dependent observations.

Chunk size affects computation time. As shown in Tables 1 and 2, computation time was proportional to the number of computed parameters. In other words, smaller chunks required more computation time. Computational time is a legitimate concern considering that size of naturalistic driving data is increasing and that project such as SHRP2 will produce data sets several times larger than previous studies. In this context, it is important to understand that if on the one hand, chunking in smaller chunks requires more processing power, chunking may also be used to create subsets of aggregated data for further analyses. In such case, chunking could actually be advantageous for computational time.

Chunk size affects the amount of data that is discarded and not used for analysis. Tables 1 and 2 show the amount of data lost due to the discarding of (1) chunking residuals and (2) too-short segments. Data loss ranged from a total of 2.6% to 68.9% (in the

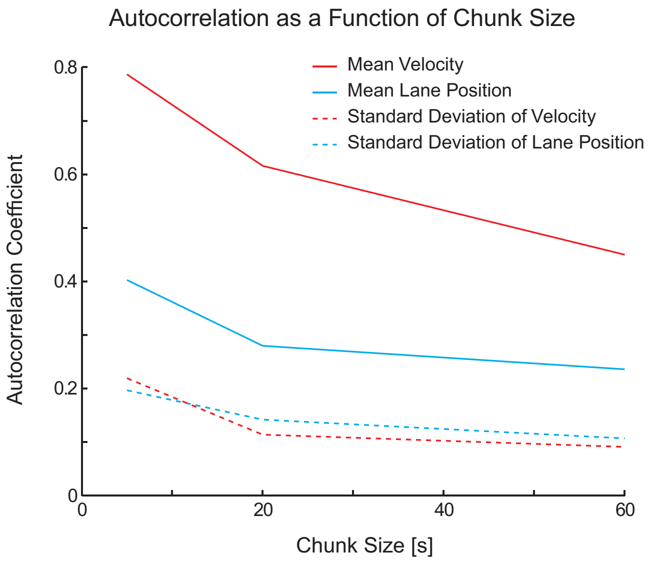

Fig. 5. Autocorrelation analysis for different chunk sizes.

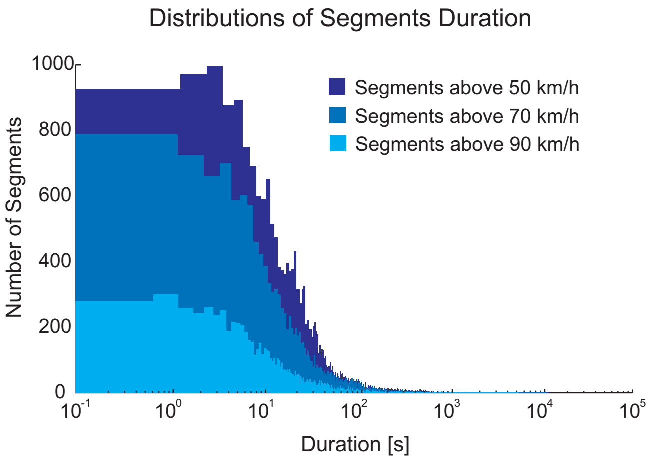

Fig. 6. Distributions of data segments from a simple query selecting data at different velocity thresholds (50, 70, and 90 km/h, respectively).

chunking range 5–600 s) with variable contributions from discarding too-short segments and chunking residuals. For small chunks, the large amount of discarded data came from chunking residuals whereas for large chunks it came from too-short segments. According to a simple linear regression from Table 1, an equal amount of data would have been discarded for 25-s chunks as for too-short segments and chunking residuals. Filtering without chunking also caused data to be discarded (Table 2). Discarded data from too-short segments is identical for (1) filtering without chunking and (2) chunking.

Chunk size also affects the autocorrelation, but this effect depends on the variable. SDLP, mean lane position, and SD of velocity processed for 5-, 20-, and 60-s chunks showed a low autocorrelation (Fig. 5). MV processed from 5-, 20-, and 60-s chunks showed a high autocorrelation (Fig. 5). Further, autocorrelation coefficients were inversely related to chunk size (i.e. short chunks presented higher values of autocorrelation). This pattern reflects the relationship between the autocorrelation and the time constant of the system. Small chunks of a signal with a long time constant, such as speed, will tend to be highly correlated.

4. Discussion

This paper presents a procedure – called *chunking* – for the analysis of naturalistic driving data. Chunking was compared with a few alternative procedures to demonstrate its rationale and advantages. Chunking is a procedure for data analysis aimed at assuring a more consistent and robust calculation of parameters from quasi-experimental data such as naturalistic driving data. More specifically, chunking divides data sets into equivalent subsets of data (chunks) before other data analysis steps, such as parameter calculation.

The data analysis presented in this paper is based on a simple and typical query extracting data for velocities above 70 km/h from a naturalistic driving database. This query is representative for many naturalistic driving analyses. In fact, a velocity threshold can be used for a variety of purposes such as identifying different types of roads and driving situations. Furthermore, 70 km/h is compatible with the activation threshold of several active safety system applications, thus making the query relevant for field operational tests. In addition, the velocity threshold does not shape the distribution of segment durations. Fig. 6 shows the similarity among distributions of segment durations obtained using different velocity thresholds. Finally, more complex queries including additional selection criteria (e.g. road type, weather, baseline and treatment for specific IVSs) would result in a yet more fragmented dataset even more suitable for chunking.

4.1. Sensitivity of parameters to segment duration

One of the biggest advantages of using chunking is that it guarantees parameters (such as SDLP) are computed on segments of equivalent duration. This is actually a basic requirement that drives the experimental design in controlled studies such as those performed in a simulator, but is not the case with naturalistic data. Naturalistic driving data collection cannot be controlled (as opposed to simulator studies), but analysis requirements for equivalent conditions remain the same. In fact, parameters calculated from naturalistic driving data are often affected by segment durations – as in studies in the simulator (Östlund et al., 2005) – but alternative solutions to experimental protocol design are needed. Fig. 7 shows how the SDLP depends on data segment durations. Fig. 7 shows the average SDLP from 33 sections of 620 s data with no lane change, from the SeMiFOT database. Each point in Fig. 7 was generated by calculating the SDLP mean across the 33 sections while increasing data length in 1-s steps. Chunking ensures that parameters are consistently calculated across equivalent data segments. Thus, chunking enables a more robust and repeatable calculation of the parameters, helping to compensate for the lack of control in the experimental protocol in naturalistic driving data collection.

This study shows how the filtering out of too-short segments – without chunking – also ensures that parameters are only computed on sufficiently long data segments, but (1) does not guarantee consistent calculation of parameters and (2) reduces the power of statistical analysis compared to chunking. In fact, as discussed above, parameters often depend on the duration of the data segments (Fig. 7) and, without chunking, these parameters are inconsistently calculated on segments of different duration. Further filtering of too-short segments before computing parameters significantly reduces the number of parameter values for use in statistics (7823 for filtering without chunking at 20 s and 65,074 for 20 s chunking; see Tables 1 and 2). This reduction results in a shorter processing time but – as a considerable drawback – produces less rich data sets for statistical analysis and less robust parameters. In addition, estimation of median SDLP

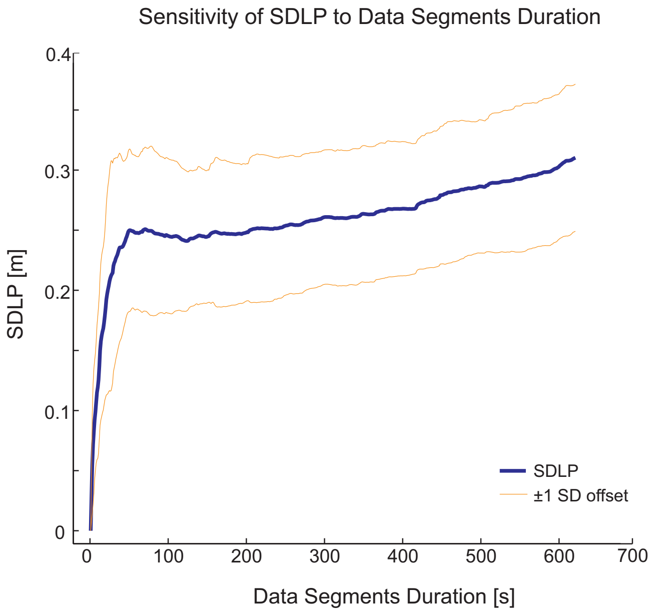

Fig. 7. Dependency of the parameter standard deviation of lane position on data chunk duration from 33 trips with velocity above 50 km/h, and without lane changes.

from filtering without chunking (Fig. 4b) proved to be significantly higher than without chunking and with weighting, suggesting a bias from longer segments and thus clearly demonstrating the limitations of this naive procedure for discarding short segments.

Benefits from chunking depend on how sensitive the analysis under consideration is to segment durations. Thus, not all analyses would necessarily need to use chunking. For instance, analyses comparing already equally long or equivalent data sets may not need chunking (e.g. when equally long events from triggered data collection are considered). Further, when parameters calculated in the analyses are intrinsically robust to segment lengths (Fancher et al., 1998; Hjälmdahl, 2004; Leblanc et al., 2006; Orban et al., 2006; Reagan et al., 2006; Battelle, 2007; Carsten et al., 2008), chunking is not necessary.

Autocorrelation analyses showed that chunks derived from a segment exhibit strong autocorrelation in some cases but not all. Autocorrelation indicates a violation of the assumption of independence that is relied upon by many analyses. For instance, parameters computed from lane position were not autocorrelated; however, MV from chunks of all sizes was found to be highly autocorrelated. Nevertheless, SD of velocity did not show the same autocorrelation. These results suggest that, in general the extent to which chunking introduces dependent observations depends on the specific measure (e.g. lane position vs velocity) and different statistical properties (e.g. mean vs standard deviation) under analysis. Further, it should be considered that the normal segmentation occurring when querying an SQL database may also introduce dependent observations, especially when the parameter of interest has a long time constant as MV. Analysis of autocorrelation is needed to verify that chunking does not introduce dependent observations. However, using the data presented in Fig. 5 to decide, for instance, to augment the size of MV chunks may have some unwanted side effects. In fact, such a decision may indirectly result in selecting only specific road types (such as freeway) if the criteria determining segment length is a speed threshold as in our example. In other words, by selecting only large-size chunks we select only long segments from the SQL query. If the SQL query was based on speed, then we are selecting only stretches of road where it was possible to maintain a high-speed for a long time.

It is worth noting that in case of dependent observations (e.g. MV) chunking may still be useful if paired with statistical procedures such as bootstrap (Abdelhak and Iskander, 2007) which can generate a distribution not affected by the dependent observation to use as a reference to test statistical significance.

When segment durations play an important role in the calculation and/or combination of the parameters under consideration (Festa-Consortium, 2008a), chunking would be of great help. Examples of functions in which segment durations play a particularly important role are functions related to (1) descriptive statistics (e.g. min, max, and median) and (2) frequency spectrum analysis (e.g. mean frequency, median frequency, and frequency content ratio).

4.2. *Chunk size*

When using chunking, the most critical decision is chunk size. This paper shows how such decision is a trade-off between: (1) parameter calculation robustness, (2) statistical analysis power, and (3) computation time. General guidelines for decisions on chunk sizes are hard to prescribe because they depend on a number of analysis-specific aspects such as experimental set-up and context. Nevertheless, as an example let us assume that MV and SDLP are chosen as safety indicators for the evaluation of an LDW active only above 70 km/h. Note that the intention of this paper is not to prove or disprove the validity of MV and SDLP as safety indicators. Let us also assume that an analyst were to calculate how average MV and median SDLP change between a baseline and a treatment condition for LDW in the SeMiFOT database using the same drivers and vehicles which were considered in this paper. According to the results presented in the paper, we would suggest the use of chunking and, more specifically, 5-s chunks for MV and 60 s chunks for SDLP. In fact, for MV, weighted average is equivalent to calculating MV over the whole data set (appending all segments together) and can be considered ground truth for MV. Often the primary interest concerns driver response over a shorter time period, such as when segments shorter than a few seconds would not capture vehicle dynamics and driver control in response to LDW (Gordon et al., 2009). A sensitivity analysis can be done to find the shortest interval size that would still capture driving responses to LDW. For this example, we assume that such an analysis would return 5 s as the shortest interval. A 5 s filtering period is significantly biased toward too low MV in relation to weighted average (ground truth). Weighting is therefore desirable. However, 5-s chunking is still the best choice since it also increases the number of available parameter values by 22 times compared to 5 s filtering with weighting with limited additional data loss (2.1%; Tables 1 and 2).

There is no easy way to calculate a ground truth value for SDLP as there is for MV because appending segments together would introduce artifacts into the estimation of the standard deviation. The most stringent requirement for chunking size comes from the nature of SDLP from Fig. 7. The curve shown in Fig. 7 stabilizes at approximately 60 s, implying the computation of SDLP on data sections longer than, or equal to, 60 s. Filtering segments shorter than 60 s without chunking would result in an overestimation of SDLP (Fig. 4b) probably due to the bias induced from long segments affected by road curvature and driving maneuvers. Weighting medians after filtering without chunking would not eliminate this bias, thus the best choice is again chunking. Furthermore, since shorter chunking results in higher power, the shortest possible chunk size should be used in order to increase the potential for statistical analysis to show statistical significance and power. Finally, computation time was not considered when determining chunk-size due to the relatively small data set, making this aspect negligible in our case.

4.3. Chunking in dimensions different than time

This paper refers to segment *duration*s and chunk *sizes* because chunking is not necessarily based on time. Chunks are elementary intervals of data designed to be equivalent to one another. The examples shown in this paper were – for the sake of simplicity and clarity – limited to chunks equivalent in time (i.e. with equal durations or sample numbers). However, chunks can be equivalent in terms of traveled km (e.g. 2-km chunks) or number of off-road glances (e.g. 3-glance chunks). In general, depending on the analysis, different types of chunking strategies should be implemented; for example chunking by traveled km would not be appropriate for eye tracking analysis (e.g. multiple glances would occur in the same spatial point in intersections, when the vehicle is standing still). Future studies should investigate how chunking may be combined with entropy estimation to define where chunks start and stop using e.g. *voting experts* algorithms (Hewlett and Cohen, 2010).

Independently of the algorithm used to generate chunks, any time chunking would result in dependent observation (as it may without chunking), this aspect should be taken into account and controlled for in the statistical analysis. The issue of dependent observations when chunking in other parameter space (distance, entropy, etc.) has also to be verified case by case. It is worth noting that if the probability of accidents is proportional to exposure to higher speed for example, then dependent measures from chunking would still have their ecological validity.

4.4. Combination of chunking with clustering

The benefit of chunking data is not limited to parameter computation and calculation. Chunking can divide data into equivalent data sets that can be used to find matches between scenarios, for instance in treatment and baseline conditions when analyzing data from field operational tests. In such case, chunks should be clustered according to several attributes (such as road type, current maneuver, weather, etc.). Future studies should address the issue of clustering data to individuate different scenarios and its potential combination with chunking to individuate matching baseline and treatment conditions. Baseline definition for field operational tests data analysis is not always straightforward. Treatment and baseline conditions need to be equivalent under a number of factors (e.g. road type, velocity distribution, weather, and traffic density) for their comparison not to suffer from biases. Chunking can help match baseline and treatment when combined with clustering of the possible confounding factors mentioned above. It is worth noting that matching chunks can lead to *paired* statistical analyses which may be a more robust than comparisons between non-paired baseline and treatment conditions.

5. Conclusions

Naturalistic data offers great potential to reveal important insights into IVSs efficacy and crash causation, but also presents many new challenges for data analysis. Chunking is a general procedure that addresses a central challenge that applies to many naturalistic driving data analyses. Chunking divides datasets under analysis into equivalent, elementary chunks of data to facilitate the robust and consistent calculation of parameters. Chunking can also increase the number of parameter values for statistical analyses, thus increasing their power.

One can regard chunking as a combination of (1) filtering out data segments that are too-small and dividing long segments into equal sized chunks, (2) weighting parameters across segments according to any quantity (e.g. time; as shown in this paper), and (3) controlling for inconsistent estimation of parameters from equivalent data sets. Chunking can also serve as a basis for further analyses and can be used to simplify other procedures such as matching data between different conditions (e.g. treatment and baseline for IVSs evaluation).

Acknowledgments

The authors would like to acknowledge Rasmus Nisslert and Jordanka Kovaceva for their participation in Matlab$^{\text{TM}}$ coding. This study was supported by the Swedish National Strategic Transport grant, the SeMiFOT2 project, and the SAFER BASFOT2 project.

References

Abdelhak, Z.M., Iskander, D.R., 2007. Bootstrap methods and applications: a tutorial for the signal processing practitioner. IEEE Signal Processing Magazine 24 (4), 10–19.

Alkim, T., Bootsma, G., Looman, P., 2007. The Assisted Driver—Systems that Support Driving Ministry of Transport. Public Works and Water Management, Rijkswaterstaat, Delft, Netherlands.

Battelle, 2007. Final Report: Evaluation of the Volvo Intelligent Vehicle Initiative Field Operational Test, Version 1.3.

Boyle, L.N., Lee, J.D., Heyens, K.M., McGehee, D.V., Hallmark, S., Ward, N.J., 2009. S02 Integration of Analysis Methods and Development of Analysis Plan: Phase I Report. SHRP2 SO2 Phase I Report. University of Iowa.

Cameron, M.H.E., Elvik, R., 2008. Nilsson's power model connecting speed and road trauma: does it apply on urban roads? In: Australian Road Safety Research Policing Education Conference, 2008 ed., Adelaide, Australia, p. 20.

CARE, 2010. Road Safety Evaluation in the EU. European Commission, Brussels.

Carsten, O., Fowkes, M., Lai, F., Chorlton, K., Jamson, S., Tate, F., Simpkin, B., 2008. ISA – UK Intelligent Speed Adaptation. University of Leeds, Leeds.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermerland, J., Knipling, R.R., 2006. The 100-car naturalistic driving study - phase ii - results of the 100-car field experiment. Technical Report DOT HS 810 593.

EuroFOT-Consortium, 2010. Eurofot – Bringing Intelligent Vehicles on the Road. EuroFOT Consortium, http://www.eurofot-ip.eu/.

Fancher, P., Ervin, R., Sayer, J., Hagan, M., Bogard, S., Bareket, Z., Mefford, M., Haugen, J., 1998. Intelligent Cruise Control Field Operational Test – Final Report. The University of Michigan Transportation Research Institute, Ann Arbor.

Festa-Consortium, 2008a. Festa D2.1 PI Matrix – Final, Final ed. FESTA-Consortium.

Festa-Consortium, 2008b. Festa Handbook, Version 2.

Gordon, T., Blankespoor, A., Barnes, B., Blower, D., Green, P., Kostyniuk, L., 2009. Yaw rate error – a dynamic measure of lane keeping control performance for the retrospective analysis of naturalistic driving data. In: 21st International Technical Conference on the Enhanced Safety of Vehicles, Stuttgart, Germany, pp. 09-0326.

Guo, F., 2009. Modeling 100-car Safety Events: A Case-based Approach for Analyzing Naturalistic Driving Data – Final Report. Blacksburg, Virginia.

Hewlett, D., Cohen, P., 2010. Artificial general segmentation. In: Advances in Intelligent Systems Research.

Hjälmdahl, M., 2004. In-vehicle speed adaptation – on the effectiveness of a voluntary system. Doctoral thesis. Lund University.

ISO 11898-1:2003. International Organization for Standardization.

Leblanc, D., Sayer, J., Winkler, C., Ervin, R., Bogard, S., Devonshire, J., Mefford, M., Hagan, M., Bareket, Z., Goodsell, R., Gordon, T., 2006. Road Departure Crash Warning System Field Operational Test: Methodology and Results. The University of Michigan Transportation Research Institute, Ann Arbor.

Najm, W.G., Stearns, M.D., Howarth, H., Koopmann, J., Hitz, J., 2006. Evaluation of an Automotive Rear-end Collision Avoidance System.

Neale, V.L., Dingus, T.A., Klauer, S.G., Sudweeks, J., Goodman, M., 2005. An overview of the 100-car naturalistic study and findings. In: Proceedings – 19th International Technical Conference on the Enhanced Safety of Vehicles, Washington, DC, June 6–9. NHTSA, Washington, DC.

NHTSA, 2009. FARS 2008. Annual Assessment ed. NHTSA. The FARS Traffic Fatality Database.

Nilsson, G., 2004. Traffic Safety Dimensions and the Power Model to Describe the Effect of Speed on Safety. Lund Institute of Technology.

Orban, J., Hadden, J., Stark, G., Brown, V., 2006. Evaluation of the Mack Intelligent Vehicle Initiative Field Operational Test – Final Report. Battelle Memorial Institute, Columbus, OH, USA.

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., Horst, D., Juch, S., Mattes, S., Fohel, U., 2005. Driving Performance Assessment Methods and Metrics. AIDE D2.2.5.

Reagan, M.A., Triggs, T.J., Young, K.L., Tomasevic, N., Mitsopoulos, E., Stephan, K., Tingvall, C., 2006. On-road Evaluation of Intelligent Speed Adaptation, Following Distance Warning and seatbelt Reminding System: Final Results of the Tac Safecar Project. Monash University Accident Research Centre, Victoria, Australia.

SAFER, 2010. SAFER – Vehicle and Traffic Safety Centre. Chalmers University of Technology, SAFER, http://www.chalmers.se/safer/SV/.

Sayer, J., Leblanc, D., Bogard, S., Hagan, M., Sardar, H., Buonarosa, M.L., Barnes, M., 2008. Integrated Vehicle-based Safety Systems Preliminary Field Operational Test Plan. The University of Michigan, Transportation Research Institute, Ann Arbor.

SHRP2, 2010. Strategic Highway Research Program – SHRP2. Transportation Research Board, Washington, DC, http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Blank2.aspx.

teleFOT-Consortium, 2010. http://www.telefot.eu/.

Turner-Fairbank Highway Research Center, U.D.O.T., 2010. Synthesis of Safety Research Related to Speed and Speed Limits. McLean, Virginia.

USDOT, 2005. Lane Departure Warning Systems. U.S. Department of Transportation – Federal Motor Carrier Safety Administration.

Victor, T., Bärgman, J., Hjälmdahl, M., Kircher, K., Svanberg, E., Hurtig, S., Gellerman, H., Moeschlin, F., 2010. Sweden-Michigan Naturalistic Field Operational Test (SeMiFOT) – Phase 1: Final Report. SAFER, Göteborg, Sweden.