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/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
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™ SQL database, it was synchronized and re-sampled from its original frequency to 10 Hz. Analysis was performed using SQL queries combined with Matlab™.

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.