Transportation Research Part A 61 (2014) 27–40

Contents lists available at ScienceDirect

Transportation Research Part A

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

Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data

Johannes Paefgen [a,*], Thorsten Staake [b], Elgar Fleisch [a,c]

[a] Operations Management, University of St. Gallen, Dufourstrasse 40a, 9000 St. Gallen, Switzerland
[b] MIS and Energy-Efficient Systems, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany
[c] Information Management, ETH Zurich, Weinbergstrasse 58, 8092 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 8 February 2013
Received in revised form 21 October 2013
Accepted 24 November 2013

Keywords:
GPS trajectories
In-vehicle data recorder
Risk exposure
Accident research
Pay-as-you-drive insurance
Multivariate logistic regression

ABSTRACT

The increasing adoption of in-vehicle data recorders (IVDR) for commercial purposes such as Pay-as-you-drive (PAYD) insurance is generating new opportunities for transportation researchers. An important yet currently underrepresented theme of IVDR-based studies is the relationship between the risk of accident involvement and exposure variables that differentiate various driving conditions. Using an extensive commercial data set, we develop a methodology for the extraction of exposure metrics from location trajectories and estimate a range of multivariate logistic regression models in a case-control study design. We achieve high model fit (Nagelkerke's R^2 0.646, Hosmer–Lemeshow significance 0.848) and gain insights into the non-linear relationship between mileage and accident risk. We validate our results with official accident statistics and outline further research opportunities. We hope this work provides a blueprint supporting a standardized conceptualization of exposure to accident risk in the transportation research community that improves the comparability of future studies on the subject.

1. Introduction

The advent of low-cost sensing and data transmission technology in a substantial number of road vehicles is providing transportation researchers with new opportunities for empirical research. While in 1976 the recording of driving data for research purposes in a sensorized vehicle required extensive hardware modifications and bulky equipment such as digital tape recorders (Helander and Hagvall, 1976), today's technology exhibits higher performance at miniature scale, and can be retrofitted to vehicles in the field. Data collection devices installed in vehicles – commonly referred to as in-vehicle data recorders (IVDR) – are providing driving and travel information from hundreds to thousands of vehicles over years of operation (Huang et al., 2010; Toledo et al., 2008; Musicant et al., 2010; Jun et al., 2010). Various studies have shown that IVDR data is more reliable and of higher resolution than conventional, self-reported driving data, thus increasing the validity and quality of insights inferred (Wolf et al., 2003, 2001; Blanchard et al., 2010; Forrest and Pearson, 2005; Stopher et al., 2007). Furthermore, IVDR offer a way to study the behavior of motorists in a minimally intrusive manner, thus improving the representativeness of results from a sample for the overall population.

Hence, transportation researchers have recently developed *naturalistic driving* studies, which employ IVDR for the collection of positioning, acceleration, and video data over extended periods of naturalistic driving (Jovanis et al., 2011; Shankar et al., 2008; Gordon et al., 2011; Wu and Jovanis, 2012). These studies yield an unprecedented level of detail with regard to driving behavior, but methodologically enter unchartered territory, and many researchers have to design their own data

* Corresponding author. Address: Dossenstrasse 13, 9470 Buchs SG, Switzerland. Tel.: +41 71 224 7240.
E-mail address: johannes.paefgen@outlook.com (J. Paefgen).

0965-8564/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tra.2013.11.010

collection systems (DAS) to suit the objectives of their specific research. While the field may see the adoption of comprehensive standards for devices and data formats in the future, as of today research efforts still require extensive financial resources and personnel for the equipment of vehicles, provision of information processing systems, and the administration of study participants. For instance, dividing the budget of a recent naturalistic driving study by the number of IVDR-equipped vehicles, one obtains per-vehicle costs of data acquisition of more than USD 20,000.

A transitional alternative to dedicated research solutions represent vehicle fleets equipped with IVDR for commercial purposes. Various business sectors such as insurance, car rental, car sharing, logistics, as well as companies operating field services in general, carry an economic interest in the type and extent of vehicle use. For transportation researchers, their fleets may offer a rich data source, featuring large sample sizes without saddling researchers with high technology costs. In the paper at hand, we utilize IVDR data made available by a provider of Pay-as-you-drive (PAYD) insurance. Under a PAYD policy, insurers calculate premiums based on the actual vehicle usage of the policyholder instead of conventional lump-sum payments, thereby improving actuarial risk differentiation and incentivizing risk mitigation by policyholders (Bolderdijk et al., 2011; Desyllas and Sako, 2012).

Within the transportation research domain, this paper focuses on the analysis of antecedents of accident involvement. From such analysis, researchers intend to inform policymakers and make suggestions towards improved driver training, traffic regulations, the design of road features and intersections, as well as vehicle safety features. We present a case-control study based on an IVDR dataset obtained from a PAYD insurance service provider. We conduct a comparative analysis of different types of vehicle *exposure* with respect to their effect on accident involvement. A plethora of previous work has discussed the relationship between mileage exposure and accident involvement (Foldvary, 1975; Janke, 1991; Progressive Insurance, 2005; White, 1976; Lourens et al., 1999; Jovanis and Chang, 1986; Langford et al., 2008; Staplin et al., 2008). Yet, studies that pursue a differentiated modeling approach to exposure and discriminate environmental conditions under which mileage was accumulated are sparse. In particular, the available sample size and resolution in such studies typically constrains the number of independent variables and prohibits the use of multivariate models. Our study utilizes location trajectories collected from two samples of 1000 (control group) and 600 (case group) vehicles, collected over 24 and 6 months, respectively, resulting in a total of 27,600 vehicle months. We devise a methodology for the extraction of exposure metrics from location data and discuss a range of multivariate logistic regression models estimated with exposure variables.

After careful modification in several steps, the final model exhibits high predictive performance. We validate the model through comparison with official accident statistics. Furthermore, we obtain insights regarding the characteristic of the nonlinear relationship between mileage and accident risk. In the concluding section of the paper, we discuss limitations to our work and outline continuing research opportunities. We propose that an exposure aggregation procedure developed in the paper may serve as a blueprint for the standardization of exposure variables that can facilitate comparative and meta-level analyses in subsequent work. Finally, we comment on critical issues in the collaboration between researchers and fleet operators, encouraging the use of commercially obtained IVDR data in future studies by the transportation research community.

2. Related work and research objective

2.1. IVDR data in accident research

Early studies in transportation research utilizing IVDR data reach back several decades, when the technological complexity of data acquisition and processing as well as the associated costs prohibited the simultaneous collection of driving data from large numbers of vehicles. When we prepared a review of such studies for this paper, we found that the sample sizes of IVDR-based studies have reached representative scales only recently (Table 1). One may consider the Texas Mileage Study published by Progressive Insurance, a PAYD provider in the US, outstanding with respect to both the number of observed vehicles and the observation period. The corresponding report (Progressive Insurance, 2005) presents a regression analysis of the relationship between annual mileages and incurred insurance losses for different coverage types, comprising more than 200,000 vehicles. For a basic linear regression model, the study obtains goodness-of-fit indicators of $R^2 > 0.82$. The study does not address any other variables besides annual mileage, nor does it give a detailed description of sample selection. While the study arguably does not offer any genuinely novel insights, it does give an indication of the potential that comes with PAYD insurance data from a research perspective. The sample size remains unmatched in other scientific undertakings in this domain, and in theory, it allows for the development of sophisticated models that combine a plethora of driving variables in the estimation of accident risk.

A more rigorous approach to IVDR-based accident research pursues a range of studies collectively referred to as *naturalistic driving*. The label naturalistic driving signifies non-intrusive data collection that informs researchers of detailed driving style and travel behavior during everyday vehicle use. A major part of recent publications in this domain builds upon data collected in the course of the Virginia Tech Transportation Institute (VTTI) 100-Car Naturalistic Driving Study conducted between 2001 and 2006 by the US National Highway Traffic and Safety Administration (Jovanis et al., 2011; Wu and Jovanis, 2012; Shankar et al., 2008). In the trial, researchers used kinematic measurements to identify critical driving events, or crash surrogates, which they then screened by means of video and survey data, delivering an unprecedented level of detail in measurement. The trial comprised 100 vehicles. A second, significantly larger naturalistic driving study is currently underway as part of the US Strategic Highway Research Program (2nd Strategic Highway Research Program, 2012). The study will equip

Table 1
IVDR-based studies with relevant research, in loose chronological order.

Dataset	Sample description	Observation period	Region	Reference(s)	Research scope
SAMOVAR	270 Commercial vehicles	24 Months	Netherlands	Wouters and Bos (2000)	Effect of driver feedback
n.a.	61 Private drivers, 1 IVDR-equipped vehicle	60 Min	Virginia, US	Boyce and Geller (2002)	Analysis of crash surrogate events
ISA Trial	4840	18 Months	Sweden	Hjälmdahl and Varhelyi (2004) Biding and Lind (2002)	Evaluation of different intelligent speed adaptation devices
n.a.	125 Bus drivers in 1 IVDR-equipped vehicle	Approx. 30 min	Sweden	Wahlberg (2004)	Estimation of past crash involvement
Texas Mileage Study	203,941 Vehicles insured by Progressive Casualty Insurance Company	36 Months	Texas, US	Progressive Insurance (2005)	Relationship between mileage and insurance claims
Commute Atlanta Program	167 Private vehicles	6 Months	Atlanta GA, US	Jun et al. (2007) Jun et al. (2010) Ogle et al. (2005)	Relationship between mileage, velocity, acceleration and accident involvement
Drive Diagnostics System	191 Commercial vehicles	Between 6 and 35 months	Israel	Toledo et al. (2008)	Relationship between critical driving events and accident involvement; effect of driver feedback
Green-Box	109 Commercial vehicles	6 Months	Israel	Oren Musicant et al. (2010)	Analysis of crash surrogate events
Virginia Tech (VTTI) 100-car Naturalistic Driving Study	100 Private and commercial vehicles	12 Months	Washington DC and Virginia, US	Shankar et al. (2008) Jovanis et al. (2011) Wu and Jovanis (2012)	Analysis of crash surrogate events
SHRP 2 Field Operational Test	78 Private vehicles	10 Months	Michigan, US	Gordon et al. (2011)	Analysis of crash surrogate events
SHRP 2 Naturalistic Driving Study	3000 Private and commercial vehicles	24 Months	6 states in the US	n.a.	n.a.
Presented Study	1600 Private and commercial vehicles	6 Months (case group) 24 months (control group)	Northern Italy	n.a.	Comparison of exposure across various driving conditions

3000 vehicles in several US states with IVDRs and will thus cover a more diverse and representative sample, with a planned budget of approximately USD 67 million. Various preparative studies and results from field operational tests (FOT) of equipment for this research program have already been published, for instance the work by Gordon et al. (2011) that was concerned with the validation of three measures of lane departure as crash surrogates.

In Europe, proposals for studies similar to the SHRP 2 are currently under discussion, such as the PROLOGUE, DaCoTa and 2BESAFE initiatives under the 7th European Union research framework program, yet to our knowledge there are no published results available at present. Various other studies with sample sizes in the lower three-digit range have used IVDR data in accident research. We omit a detailed discussion of these due to limited relevance and for the sake of brevity.

2.2. Accident risk exposure and qualitative differentiation of driven mileage

The term *exposure* appears frequently in transportation research publications and can generally be defined as the quantified potential for loss that might occur as the result of some activity. With respect to accident risk, prior work introduced and investigated various measures of driving-related activity. A common interpretation of exposure in this context is the *accumulated mileage* of a vehicle, implying that with mileage, the ceteris paribus probability of accident involvement increases (Wolfe, 1982). Previous work has investigated the role of mileage as a single predictor of accident risk (Chipman, 1982; Foldvary, 1975; Jovanis and Chang, 1986; Lourens et al., 1999). Analogously, exposure can also refer to the driving duration of a vehicle (Wolfe, 1982; Chipman et al., 1992; Chipman et al., 1993). Besides accidence occurrence, researchers have also established correlations between mileage and crash fatalities (Litman, 2011) as well as mileage and insurance claims (Bordoff and Noel, 2008; Ferreira and Minike, 2010). While the Texas Mileage Study referenced in the previous section found a linear relationship between mileage and insurance claims, the literature has also discussed non-linear relationships particularly for low-mileage drivers (Janke, 1991; Staplin et al., 2008; Langford et al., 2008). Risk and Shaoul (1982), by pointing out that the measure of vehicle mileage "seems to reflect mainly the extent rather than the degree of accident risk exposure", introduced a more differentiated view of exposure. While extent signifies a quantitative representation of exposure, e.g., accumulated mileage, degree refers to qualitative characteristics of exposure. For instance, the authors discuss road properties as a criterion for the discrimination of exposure degrees.

As a variable, mileage is comparatively straightforward to obtain even for large vehicle samples, whereas the qualitative characteristics under which it was accumulated were difficult to collect prior to the advent of IVDR-based field studies. Jun et al. (2010) have recently demonstrated the various exposure types accessible through IVDR data in the context of the Commute Atlanta Program. Their case-control study design, comprising 167 IVDR-equipped vehicles, examines the differences between vehicles that were either accident-involved or accident-free over the duration of a limited observation period. They compare several velocity-derived exposure metrics across the two groups controlling for road type (freeways, arterials, or local roads) and day times.

It appears, however, that prior work has not yet addressed the integrative modeling of accident risk based on the accumulated exposure of vehicles using differentiated driving characteristics. In the opinion of the authors, this constitutes an important research objective for which IVDR data are quite suitable. The Texas Mileage Study – as the only study with a similar scope of substantial sample size – does not take a differentiated perspective on exposure modeling and when this paper went to press has forgone publication in peer-reviewed outlets. Specifically, a shortcoming of previous IVDR-based studies that investigate the relationship between exposure variables and accident involvement is the restriction to univariate modeling, i.e., testing is typically restricted to individual factor effects. Multivariate exposure models would allow for a comparative assessment of independent variables and reveal correlations between them. At the same time, they require sample sizes that exceed most published IVDR studies by an order of magnitude. Furthermore, there appears to be no common understanding as to how different "degrees" of exposure, or driving situations, can be integrated with the "extent" of exposure, e.g., mileage, in a single, holistic model in the sense of Risk and Shaoul. The objective of the paper at hand is to develop a methodological and empirical approach to address these issues.

3. Data sampling and processing

3.1. IVDR data from PAYD insurance

We obtained the IVDR data used in our study from a major European PAYD insurance service provider. In the handling of the data, we adhered to strict privacy-protecting measures. In particular, we did not access information regarding the driver profiles or their insurance company. Each vehicle was equipped with an on-board unit that included a GPS sensor and wireless transmission capabilities. During vehicle operation, position updates were carried out every couple of seconds and aggregated on the device level to reduce costs of data transmission and storage. For aggregation, the system calculated traveled distance from incremental position updates and generated new data entries every segment of 2000 m. Next to a vehicle's latitude and longitude, data points consisted of a time stamp, ignition status of the vehicle, and driven distance since the previously generated data point. Segment distance lay below 2000 m when vehicle ignition was turned off and the current trip ended. Segment distance could in some cases exceed the 2000 m interval if no position update was available for some time owing to signal obstruction, for example, causing the segment to end only with the next valid position measurement.

Through straightforward computations, we extended raw data points to include the elapsed time since the last update, which in turn allowed us to compute the average velocity for the previously driven distance. In addition, the system inferred a road type indicator from data point GPS coordinates using map matching techniques. For each officially mapped road, Italian authorities provide an administrative classification, which distinguishes the road types "Autostrada" (highway), "Strada statale/regionale/provincial" (extra-urban roads), and "Strada comunale" (urban roads). This classification was adopted for labeling collected data points. Lastly, start and end locations of vehicle trips were available from data points generated upon changes of the vehicle ignition status (i.e., engine start and switch off).

3.2. Sample selection

In its entirety, the IVDR database is computationally intractable by means available to us; thus, we resorted to a randomized sampling procedure. Sampling followed a case-control research design (Schulz and Grimes, 2002) and was carried out as follows: We randomly drew a sample of 600 vehicles that had an accident in 2008. This sample contained 6 months of location data prior to the accident event. *Accident* comprises the categories "incident", "incident with injuries", and "incident with death", according to the Italian traffic authorities. We used stratified sampling to achieve an even distribution of accident events over the year, so that one-twelfth, i.e., 50 vehicles, shared the same month in which the accident occurred. By sampling an equal number of accident events for each month, we hoped to eliminate the effect of seasonal variations on accident frequencies in our analysis. While seasonally varying conditions may significantly influence accident risk, they pose a challenge in a case-control study setting which requires the determination of average exposure for the respective groups: To represent variations in individual mileage exposure over seasons adequately, a significantly larger time horizon would have to be taken into account. We rejected this approach for the study at hand due to both limited availability of historical data and the questionable stability of mileage accumulation by drivers across several years.

No location data beyond an accident event were included in the sample, as previous work has reported strong variations in driving patterns in the aftermath of an accident (Mayou et al., 1993). As a control group, we furthermore randomly drew a second sample of 1000 vehicles from the data pool with 24 months of location data without accident involvement throughout this period, spanning from July 2007 to June 2009. Thus, a total of 27,600 vehicle months was included in the study.

A number of vehicles were eliminated from both samples for the following reasons:

- further accident events in the 6-month observation periods of accident-involved vehicles that would affect vehicle usage,
- errors in data recording or storage that made it impossible to process the resulting log files,
- failure of GPS sensors over prolonged periods, so that no location data was available for certain vehicles even though ignition status indicated vehicle use, and
- non-continuous GPS readings that resulted in excessively long traveled distances.

These instances were unambiguously identifiable and resulted in a reduction of the accident-involved sample by 17 vehicles (2.8%) to 583 and of the accident-free sample by 16 vehicles (1.6%) to 984. No further elimination of outliers was undertaken, since we argue that their effect on the results of logistic regression analysis is negligible due to the large sample size. Both samples combined cover approximately 45.7×10^6 km driven distance in 1.0×10^6 hours of vehicle operation.

3.3. Aggregation to exposure matrix

The combined datasets comprise 2,679,425 data points, i.e., trip segments, with an average of 958.5 data points per vehicle. In order to prepare the statistical analysis of vehicle exposure, we further process these segments to aggregated exposure metrics on an individual vehicle level. We define an N-by-M accounting matrix $\mathbf{E}$, where each row index n corresponds to one of the 1567 vehicles in the combined datasets, and each column index m represents a distinct condition under which a given vehicle accumulated a fraction of its mileage. When data points are processed, the aggregation algorithm identifies the driving condition for a given segment and increments the corresponding entry in $\mathbf{E}$ by the mileage associated with that segment. An overview of the aggregation process is depicted in Fig. 1.

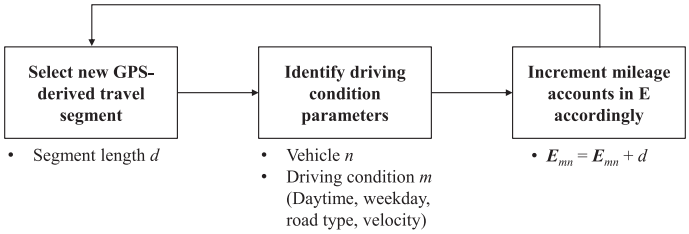

Fig. 1. Exposure aggregation process.

Table 2
Column structure of exposure aggregation matrix **E**.

Time of day	Day of week	Road type	Velocity interval	Σ
24 Columns	7 Columns	3 Columns	5 Columns	39 Columns

Choosing reasonable criteria for the exposure conditions that determine the columns of matrix **E** is non-trivial. For the dataset at hand, we intend to take into account the following information that we can directly infer for a given segment:

- at which time of the day and
- on which day of the week a vehicle was operated,
- which road type was predominantly used (as indicated by segment start location),
- and what the average velocity was.

Aggregating driven mileage separately for different driving situations requires a discretization of the continuous situational variables daytime and velocity. We initially choose a high resolution for discretization and discuss a fusion of adjacent categories according to a similarity criterion in Section 4. We discretize the time of day variable in hourly intervals. With respect to velocity, we consider five intervals of 30 km/h width, where the last interval is open-ended and thus captures all vehicle operation above 120 km/h. For the weekday variable, we consider each of the seven weekdays as a separate category. Lastly, we maintain the already established differentiation between highways, urban, and extra-urban roads for the road type variable.

If the above conditions were taken into account simultaneously, each column in **E** would correspond to a specific combination of a daytime interval, a day of the week, a road type, and a velocity interval. This would result in 2520 different mileage exposure counts and yield an excessive number of variables unsuitable for modeling purposes. We therefore aggregate situational exposure in parallel, i.e., for each of the named criteria separately. Thus, the number of columns in **E** is reduced to 39 as displayed in Table 2. We acknowledge that the described approach precludes accounting for exposure as specified by an overlay of intervals (e.g., a certain time interval on a certain day) and thus precludes the analysis of interaction effects. Note that exposure generated by a given segment is registered in four columns simultaneously, each one corresponding to the active interval of one of the four factors.

We compute the overall mileage exposure (i.e., total number of driven km) for each driver in the sample by summing up a subset of column entries in **E** for a specific category. The resulting sum is the same across categories (time of day, day of week, etc.) due to the separate accounting of these conditions. Next, we divide all entries in a row of **E** by the resulting value, i.e., the mileage exposure for each vehicle. In consequence, the modified entries in **E** represent the *fraction* of exposure accumulated under specific conditions. For example, a value of 0.2 in the column corresponding to road type 'highway' signifies that the vehicle has accumulated 20% of its mileage exposure under these conditions. As the overall accumulated mileage is also of relevance, it is stored in an additional column, normalized by the observation period of the respective group (i.e., 6 and 24 months, respectively), so that we obtain a per-month mileage exposure value. At this stage, the raw data processing yields 40 variables per vehicle, represented as columns of matrix **E**. Histograms of average monthly mileage and the exemplary velocity exposure interval between 60 and 90 km/h across the rows of **E** are shown in Fig. 2.

4. Exposure-based logistic regression models

In order to systematically analyze differences between case and control group, we employ logistic regression modeling (Christensen, 1997). In logistic regression, a linear combination $g = \beta_0 + \sum_i \beta_i x_i$ of exposure variables x_i is the argument of a logistic function, the output of which is interpreted as the probability of event occurrence,

$$\widehat{P}(\text{case}|x_i) = 1 - \widehat{P}(\text{control}|x_i) = \frac{e^g}{1+e^g}.$$

Using maximum likelihood methods, the intercept β_0 and the coefficients β_i are estimated such that the errors between predicted probabilities and actual event observation – one in the case group, zero in the control group – are minimized.

We provide three goodness-of-fit measures for logistic regression models. These are pseudo-R^2 indicators according to Cox and Snell (1968) and Nagelkerke (1991), as well as the Hosmer–Lemeshow test statistic (Lemeshow and Hosmer, 1982). We evaluate different logistic regression model variants based on modified sets of exposure variables. In order to avoid under-fitting or over-fitting of our model, we employ a backwards stepwise estimation approach, where variables are iteratively removed from the model if their significance exceeds a threshold of $p = 0.1$ according to the Wald test.

The computation of minimally required sample size or statistical power for multivariate logistic regression models is a complex problem, see, e.g., (Schoenfeld and Borenstein, 2005). No exact formulas are available for a larger number of non-normal distributed predictors, and in particular for the mixed variable-type model in Section 4.4. We therefore resort to approximate simulation studies that give a lower bound on sample cases per predictor (Peduzzi et al., 1996). A conservative minimum for this ratio given by the authors is 36. Considering our sample size of 1567 vehicles, this renders feasible the estimation of models with up to 43 predictors.

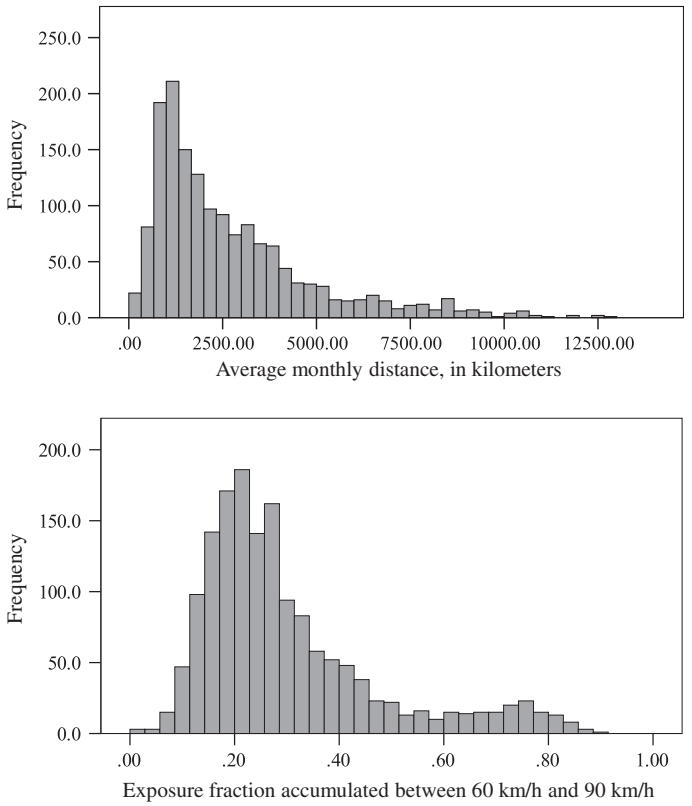

Fig. 2. Histograms of average monthly mileage (top) and the fraction of mileage accumulated within the 60–90 km/h velocity interval (bottom).

4.1. Full variable set

For an initial model, we consider the entire variable set contained in the exposure matrix **E** as introduced in Section 3. The stepwise estimation yields 35 variables out of the available 40 included in the model. These are all daytime intervals, six of the seven weekday intervals, urban and highway road-type exposure, and the velocity intervals 0–30 km/h and 60–90 km/h. Average monthly mileage is also included. Pseudo-R^2 values show a good fit with *0.387* (Cox and Snell) and *0.528* (Nagelkerke), while the Hosmer–Lemeshow test is not significant ($\chi^2 = 8.539$, $p = 0.383$) and thus confirms good model fit. However, the model exhibits high error terms, in particular for the daytime variables (*>2800*), which we attribute to high collinearity, i.e., non-independent error terms. As a remedy for this, we proceed with a selective merging of adjacent exposure intervals to reduce collinearity in the subsequent section. We omit a detailed account of coefficients and *p*-values for this model for the sake of brevity.

4.2. Merged-intervals variable set

As a preliminary indicator for the merger of exposure intervals, we conduct an exploratory factor analysis across intervals of the daytime and weekday variable groups. We employ the well-established method of Principal Components Analysis (Jolliffe, 2005) in order to obtain a target estimate of the number of merged intervals. The corresponding Scree plots are given in Fig. 3. For the 24 intervals in the daytime variable group, factor analysis yielded seven factors with Eigenvalues larger than one, however with a distinct bend in the Scree plot after the fourth factor. For the weekday variable group, only one factor with an Eigenvalue larger than one is obtained, while the corresponding Scree plot exhibits a bend after the second factor. As an alternative indicator for the similarity between intervals, we furthermore compare the mean differences between case and control samples across different intervals. Combining both analyses, we conclude (i) to merge the daytime variables into four new intervals (00–05 h, 05–18 h, 18–21 h, and 21–24 h) and (ii) to merge the day of week variables into two new intervals (Monday through Thursday, and Friday through Sunday).

After three iterations, the logistic regression model contains 9 variables, which are given in Table 3 together with β-coefficients, error terms and *p*-values. Based on standard errors, average monthly mileage is observed to be the strongest indicator. Its small coefficient is due to the comparatively large values (in km) of the corresponding variable. With pseudo-R^2 values of *0.353* (Cox and Snell) and *0.482* (Nagelkerke), goodness-of-fit is slightly lower than in the previously discussed

Table 3
Coefficients, error terms and significance for model based on merged-interval variable set.

Variable	Coefficient	Standard error	Significance
Time of day			
05–18 h	4.346	1.047	<.001
18–21 h	7.277	1.543	<.001
21–24 h	13.121	2.241	<.001
Day of week			
Monday through Thursday	4.748	.839	<.001
Road type			
Urban	2.616	.706	<.001
Highway	−2.804	.490	<.001
Velocity interval			
0–30 km/h	−6.625	1.543	<.001
60–90 km/h	−8.613	.695	<.001
Average monthly mileage	.001	.000	<.001
Constant	−7.261	1.249	<.001

model, and the Hosmer–Lemeshow χ^2 of *12.169* is not significant ($p = 0.144$). Error terms now take acceptable values (below *0.595*) except for the constant term (*1.903*), confirming the successful reduction of collinearity in comparison to the non-merged-interval variables.

4.3. LN-transformed variable set

For further improvement of model fit, we examine the distributional characteristics of exposure variables. We observe most predictor distributions, and particularly average monthly mileage, to be left-skewed; see Zhang et al. (2011). Under the assumption that they approximately follow a log-normal distribution, we transform variables by taking their natural logarithm in order to improve their resemblance to a normal distribution. For the two exemplary variables of Fig. 2, we plot the observed cumulative probability of untransformed and LN-transformed variables against the cumulative probability of a normal distribution in Fig. 4. The benefit of LN-transformation was not equal for all exposure variables, although it was evident for a majority.

While normality of predictor variables is not a requirement of logistic regression, it typically improves the model fit (Christensen, 1997). We estimate a logistic regression model based on a LN-transformed, merged-interval variable set. After seven removal iterations, the model again contains nine variables, which are given in Table 4 together with β-coefficients, error terms and p-values. The model achieves improved pseudo-R^2 values of *0.454* (Cox and Snell) and *0.614* (Nagelkerke). However, the Hosmer–Lemeshow test becomes significant ($\chi^2 = 14.372$, $p = 0.073$), indicating insufficient predictive performance. Subsequently, we remedy this observed degradation of the Hosmer–Lemeshow statistic by extending our model to include non-linearities.

4.4. Categorical mileage exposure

With a value of the Wald statistic of *255.733*, average monthly mileage exceeds all other predictors by a multiple in terms of the significance of its effect on the log-odds in the model discussed above. Based on a detailed analysis of descriptive

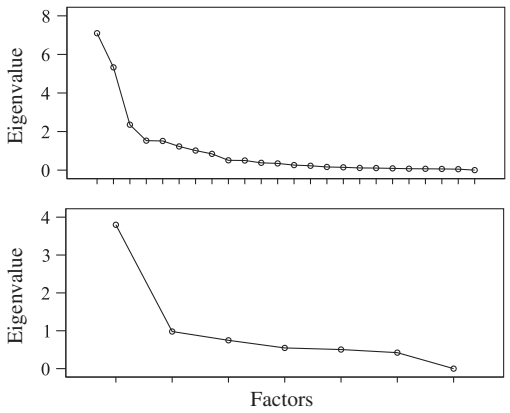

Fig. 3. Scree plots for daytime (top) and weekday (bottom) exposure factors from PCA.

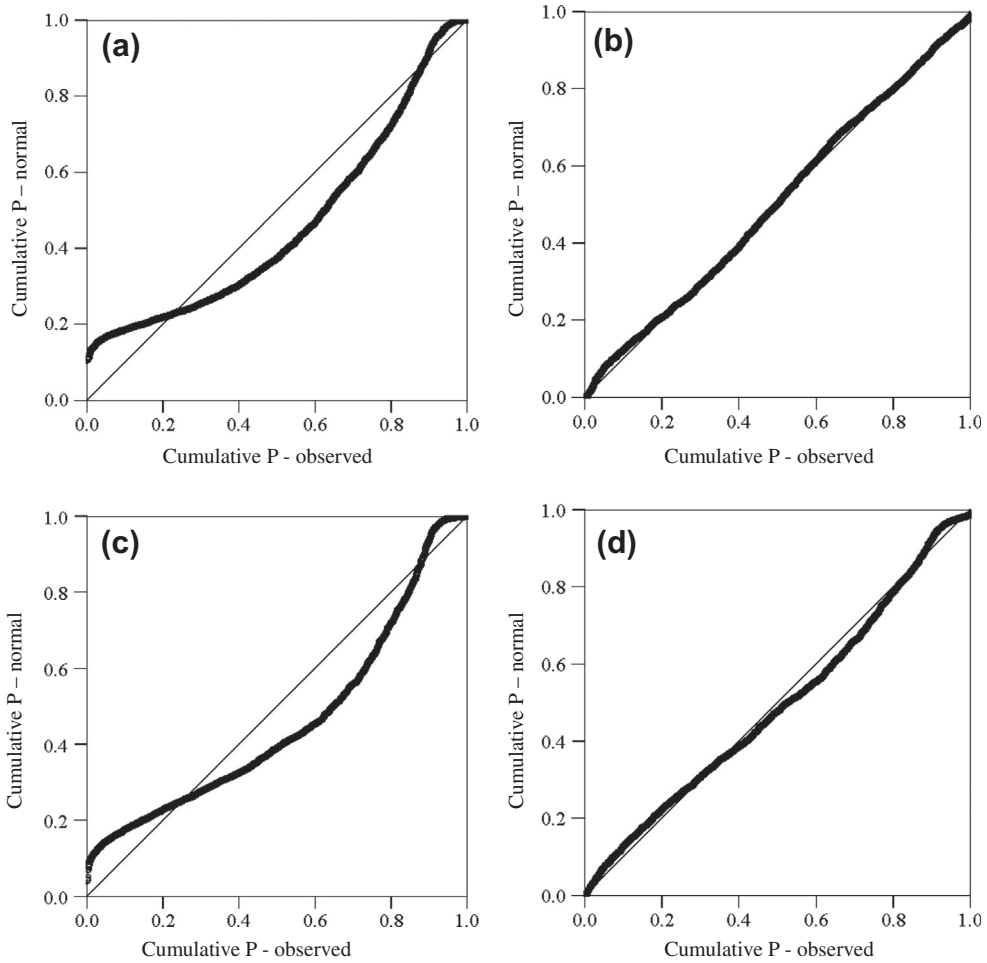

Fig. 4. P–P plots for (a) average monthly mileage, (b) the same after LN-transformation, (c) relative exposure at 60–90 km/h, and (d) the same after LN-transformation.

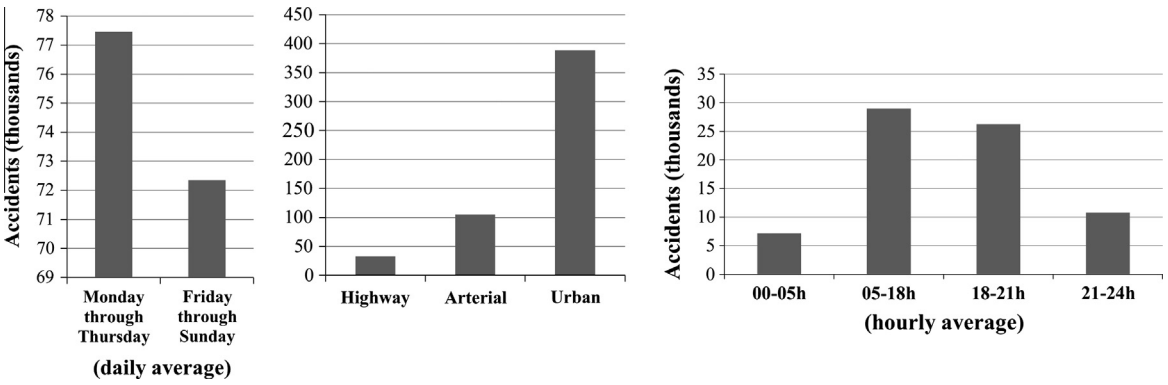

Fig. 5. Distribution of the 526,900 road traffic accidents that occurred in Italy in 2008 according to day of week, daytime, and road type (ISTAT, 2009).

statistics, we hypothesize a non-linear relationship between mileage exposure and the log-odds. Christensen (1997) has suggested remedying the issue of non-linearity in logistic regression by transforming metric variables to ordinal levels and including them as a categorical predictor. We therefore generate binary dummy variables that associate each case in the sample with a certain average monthly mileage interval, using the lowest interval as the reference level. We discretize in 5 and 10 intervals, with an equal percentage of cases in each respective interval, i.e., 20% and 10%. The resulting model parameters are given in Tables 5 and 6. For the 5-bin mileage variable, pseudo-R^2 values are approximately the same as

Table 4
Coefficients, error terms and significance for model based on merged-interval and LN-transformed variable set.

Variable	Coefficient	Standard error	Significance
Time of day (LN-transformed)			
05–18 h	−1.671	.595	.005
18–21 h	.567	.131	<.001
Day of week (LN-transformed)			
Friday through Sunday	−1.216	.403	.003
Road type (LN-transformed)			
Urban	1.065	.241	<.001
Highway	−.305	.118	.010
Velocity interval (LN-transformed)			
0–30 km/h	1.021	.345	.003
60–90 km/h	−1.573	.248	<.001
90–120 km/h	.454	.150	.002
Average monthly mileage (LN-transformed)	3.829	.239	<.001
Constant	−28.973	1.903	<.001

Table 5
Coefficients, error terms and significance for model based on merged-interval and LN-transformed variable set, categorical mileage exposure (5 bins).

Variable	Coefficient	Standard error	Significance
Time of day (LN-transformed)			
05–18 h	−.424	.118	<.001
18–21 h	.720	.164	<.001
Day of week (LN-transformed)			
Friday through Sunday	−.352	.125	.005
Road type (LN-transformed)			
Urban	.679	.152	<.001
Velocity interval (LN-transformed)			
0–30 km/h	.455	.141	.001
60–90 km/h	−.655	.128	<.001
90–120 km/h	.292	.165	.077
Average monthly mileage (categorical)			
<1021 km	0 (reference)	–	–
1021–1585 km	.933	.357	.009
1585–2522 km	2.989	.339	<.001
2522–3966 km	4.979	.371	<.001
>3966 km	6.478	.483	<.001
Constant	−3.997	.322	<.001

for the model described in Section 4.3, albeit with a considerably improved Hosmer–Lemeshow test result ($\chi^2 = 10.026$, $p = 0.263$). For the 10-bin mileage variable, pseudo-R^2 values rise to *0.478* (Cox and Snell) and *0.646* (Nagelkerke) and the Hosmer–Lemeshow χ^2 of *4.097* tested highly insignificant ($p = 0.848$).

4.5. Discussion

We subsume goodness-of-fit measures for all discussed models in Table 7. Clearly, the last model outperforms the previous ones in terms of both pseudo-*R*-squares and Hosmer–Lemeshow test results. An interesting observation is the apparent independence of these two measures, as evident from the introduction of the categorical monthly mileage variable. Furthermore, we report the AIC (Akaike, 1974) and BIC (Schwarz, 1978) information criteria computed from

$$\mathrm{AIC} = 2k - 2\ln(L), \text{ and } \mathrm{BIC} = k\ln(n) - 2\ln(L),$$

where k is the number of variables in the model, $n = 1567$ the sample size, and L is the likelihood. While the AIC is consistent in that the 10-bin categorical model receives the lowest, i.e., best score, the non-categorical model is slightly favorable according to the BIC, though very close to the 10-bin model.

From the β-coefficients in Table 6, we are able to infer a ranking of different driving exposure types according to their influence on the accident involvement log-odds. We find that

- The risk of accident involvement is lower for the daytime interval between 5:00 and 18:00 h, while higher for the interval between 18:00 and 21:00 h.
- Driving exposure accumulated on weekends, including Fridays, is associated with lower risk.

Table 6
Coefficients, error terms and significance values for model based on merged-interval and LN-transformed variable set, categorical mileage exposure (10 bins).

Variable	Coefficient	Standard error	Significance
Time of day (LN-transformed)			
05–18 h	−.397	.123	.001
18–21 h	.698	.168	<.001
Day of week (LN-transformed)			
Friday through Sunday	−.266	.133	.046
Road type (LN-transformed)			
Urban	.737	.163	<.001
Highway	−.289	.142	.042
Velocity interval (LN-transformed)			
0–30 km/h	.380	.152	.012
60–90 km/h	−.733	.137	<.001
90–120 km/h	.306	.177	.084
Average monthly mileage (categorical)			
<767 km	0 (reference)	–	–
676–1021 km	.361	.576	.531
1021–1274 km	1.023	.544	.060
1274–1585 km	1.394	.529	.008
1585–2000 km	2.623	.508	<.001
2000–2522 km	4.075	.514	<.001
2522–3188 km	4.672	.523	<.001
3188–3966 km	6.539	.579	<.001
3966–5705 km	7.193	.637	<.001
>5705 km	7.032	.721	<.001
Constant	−4.418	.475	<.001

Table 7
Goodness-of-fit measures for logistic regression models.

Model	−2 Log likelihood	k	Information Criteria		Pseudo-R^2		Hosmer–Lemeshow test	
			AIC	BIC	Cox and Snell	Nagelkerke	χ^2	Sig.
Full variable set (35 out of 40)	1301.044	35	1441.044	1558.536	.387	.528	8.539	.383
Merged-interval	1385.644	10	1425.644	1459.213	.353	.482	12.169	.144
Merged-interval, LN transformed	925.959	10	945.959	999.528	.454	.614	14.372	.073
Merged-interval, LN transformed, 5-cat. mileage	925.928	12	949.928	1014.211	.453	.612	10.026	.263
Merged-interval, LN transformed, 10-cat. mileage	870.506	18	906.506	1002.93	.478	.646	4.097	.848

- Urban driving is associated with high risk, while driving on highways has the lowest risk per fraction of mileage.
- The mid-range of velocities (60–90 km/h) has the lowest risk of accident involvement, while both the lowest velocity interval (0–30 km/h) and the second-highest interval (90–120 km/h) are associated with higher risk.

The coefficients for velocity intervals seem counterintuitive, as the literature generally associates higher velocities with higher risk of accident involvement. However, one has to consider that the duration of vehicle operation for a given amount of mileage is inversely proportional to velocity. This effect may also contribute to the elevated coefficient of urban driving.

We compare the inferred rankings of exposure situations with averaged accident statistics for the year in which the accidents observed in our sample occurred (Fig. 5). For the weekday and the road type statistics, both show the same trend. For daytime intervals, the result of normalization becomes evident. While the hourly average of accidents is higher between 5:00 and 18:00 h, vehicles in our sample drove an hourly average of 158.3 km per month for this interval, compared to 81.5 km between 18:00 and 20:00 h. This confirms the validity of our model and demonstrates the additional insights accessible through an exposure-based analysis of road accident involvement.

With respect to average monthly mileage, a conversion from a metric to a categorical variable revealed a non-linear relationship to the log-odds. To demonstrate this finding, we plot the lower mileage-interval bounds against the β-coefficients of the corresponding dummy variables in Fig. 6. Coefficient values monotonously increase up to 4000 km but slightly decrease for the top interval. Up to this point, two segments are distinguishable. Below approximately 1600 km, a linear function overestimates the per-mile risk contribution, while above this limit it is underestimated.

Our results appear to contradict the linear mileage–risk relationship as suggested by, e.g., the Texas Mileage Study. Furthermore, we do not observe the 'low-mileage bias' in terms of an elevated risk at lower mileage exposure as discussed in Section 2.2. However, both discrepancies may be explained by the simultaneous consideration of mileage and situational

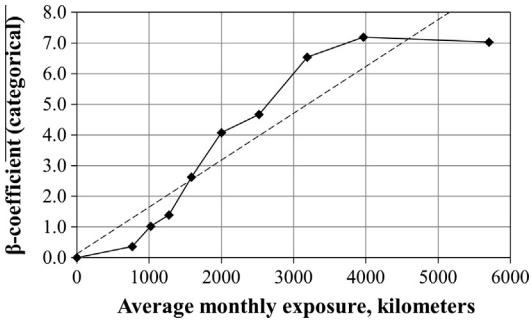

Fig. 6. Non-linear progression of coefficients of categorical dummy-variables with average monthly mileage.

exposure variables in our analysis. Controlling for the influence of road type, daytime, etc. – factors that have been suggested as causal for the 'low-mileage bias' (Janke, 1991) – it is not surprising that our assessment of the mileage–risk relationship diverges from univariate models.

5. Conclusion

In this study, we have proposed a methodology for multivariate modeling of the exposure–accident relationship with IVDR data. Based on location trajectories from 1600 vehicles obtained from a PAYD insurance provider, we have developed and validated several models that explain differences between accident-involved and accident-free vehicles in a case-control study. The discussed models combine mileage as a measure of the "extent" of exposure with several groups of situational variables that represent the "degree" of exposure, such as daytime, weekday, road type, and velocity.

From model coefficients, we were able to infer a ranking of situational variables with respect to their contribution to the risk of accident involvement. A comparison with official accident statistics demonstrated the value of a differentiated concept of exposure, which supports the interpretation of observed accident frequencies. Furthermore, we presented evidence that when the driving situation is controlled for, the relationship between mileage and accident involvement deviates from a linear function.

5.1. Limitations and research opportunities

We acknowledge principal limitations of case-control studies. In particular, we point out that odds ratio estimates are different from relative risk and do not allow for any inference of probability of accident involvement for an individual vehicle, or accident frequencies. Such information could be inferred in cohort studies that observe a given vehicle population over time, which would arguably require significantly larger sample sizes than in the presented study. The presented case-control study provides insights into the relative contribution of a multitude of influence factors for accident risk and outlines a holistic modeling approach to driving exposure. Given the availability of even larger sample sizes, and over longer observation periods, a cohort study with count-regression models (Lord and Mannering, 2010) is likely to yield valuable additional information.

Another possible extension of the presented study is a separation of accident types. Information regarding driver profile and vehicle, which was not available to us for privacy reasons, would enable additional differentiation. Furthermore, the influence of additional variables extracted from raw data would be worthwhile to consider. These could for instance include trip lengths, the 'familiarity' of routes and locations, or speed limit violations. In addition, the presented framework for exposure modeling is based solely on mileage. An alternate approach is to use driving duration as a primary measure of exposure.

A further limitation of our research is the limited representativeness of the dataset. The sample is regionally restricted to Italy, and only contains vehicles operated under a PAYD insurance contract. As a majority of previous IVDR-based studies uses data from vehicles in the US, a European sample is, in our opinion, desirable. However, we call for researchers in other regions to reiterate the proposed modeling methodology and verify if our results can be reproduced with similar datasets.

5.2. Use of commercially collected IVDR by the transportation research community

This concluding section extends the technical part of this paper with some more general remarks that may facilitate the use of commercially collected IVDR data in research. In our opinion, the increasing number of IVDR-equipped vehicles that come with commercial services such as PAYD insurance represents a momentous opportunity for transportation researchers. Besides the focus of our study, the used sample may deliver additional insights in fields such as the analysis of route choices, commuting patterns, or travel times. In order to unlock the potential of such data, we deem two objectives as particularly important from a research policy perspective.

First, within the domain of accident analysis, it is worthwhile to consider a standardized conceptualization of exposure in terms of variables that can be derived from IVDR data. Compared to other means of empirical data acquisitions, IVDR are more objective and reliable, as measurement parameters can be precisely defined. This makes standardization feasible, which would facilitate the exchange of such data among researchers and support the use of meta-analyses to aggregate evidence across individual studies. We propose the aggregation of situational exposure developed in Section 3.3 as a blueprint for future work in this regard.

Secondly, consideration should be given to technical, legal, and economical frameworks that enable the collaboration between researchers and commercial entities for the exchange of IVDR data. A primary issue in this regard is certainly the privacy of vehicle owners. A further challenge is the willingness of data providers to collaborate, although compared to the budget requirement of large-scale, dedicated research studies, financial compensation is a viable option. Ultimately, given the enormous efforts undertaken by agencies such as the NHTSA and its European and international counterparts to improve road safety, legislative action regarding the access to IVDR data from large vehicle fleets may also be worth considering.

References

2nd Strategic Highway Research Program, 2012. Naturalistic Driving Study. <http://www.shrp2nds.us/> (accessed 23.11.12).

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. AC-19 (6), 716–723.

Biding, T., Lind, G., 2002. Intelligent Speed Adaptation, Results of Large-scale Trials in Borlänge, Lidköping, Lund and Umea during the Period 1999 to 2002. Technical Report. Swedish National Road Administration.

Blanchard, R., Myers, A.M., Porter, M.M., 2010. Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers. Accid. Anal. Prev. 42 (2), 523–529.

Bolderdijk, J.W., Knockaert, J., Steg, E.M., Verhoef, E.T., 2011. Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment. Accid. Anal. Prev. 43 (3), 1181–1186.

Bordoff, J.E., Noel, P.J., 2008. Pay-as-you-drive Auto Insurance: A Simple Way to Reduce Driving-related Harms and Increase Equity. The Brookings Institution. <www.brookings.edu/papers/2008/07_payd_bordoffnoel.aspx>.

Boyce, T.E., Geller, E.S., 2002. An instrumented vehicle assessment of problem behavior and driving style: do younger males really take more risks. Accid. Anal. Prev. 34 (1), 51–64.

Chipman, M., 1982. The role of exposure, experience and demerit point levels in the risk of collision. Accid. Anal. Prev. 14 (6), 475–483.

Chipman, M.L., MacGregor, C.G., Smiley, A.M., Lee-Gosselin, M., 1992. Time vs. distance as measures of exposure in driving surveys. Accid. Anal. Prev. 24 (6), 679–684.

Chipman, M.L., MacGregor, C.G., Smiley, A.M., Lee-Gosselin, M., 1993. The role of exposure in comparisons of crash risk among different drivers and driving environments. Accid. Anal. Prev. 25 (2), 207–211.

Christensen, R., 1997. Log-linear Models and Logistic Regression, second ed. Springer, New York.

Cox, D.R., Snell, E.J., 1968. A general definition of residuals. J. Roy. Stat. Soc. Ser. B Stat. Methodol. 30 (2), 248–275.

Desyllas, P., Sako, M., 2012. Profiting from business model innovation: evidence from Pay-As-You-Drive auto insurance. Res. Policy 42 (1), 101–116.

Ferreira, J., Minike, E., 2010. Pay-as-you-drive Auto Insurance in Massachusetts: A Risk Assessment and Report on Consumer, Industry and Environmental Benefits. Department of Urban Studies and Planning, Massachusetts Institute of Technology. <www.clf.org/our-work/healthy-communities/modernizing-transportation/pay-as-you-drive-auto-insurance-payd>.

Foldvary, L., 1975. Road accident involvement per miles travelled, Part II. Accid. Anal. Prev. 11 (2), 191–205.

Forrest, T.L., Pearson, D.F., 2005. Comparison of trip determination methods in household travel surveys enhanced by a global positioning system. Transp. Res. Rec. 1917, 63–71.

Gordon, T.J., Kostyniuk, L.P., Green, P.E., Barnes, M.A., Blower, D., Blankespoor, A.D., Bogard, S.E., 2011. Analysis of crash rates and surrogate events. Transp. Res. Rec. 2237, 1–9.

Helander, M., Hagvall, B., 1976. An instrumented vehicle for studies of driver behavior. Accid. Anal. Prev. 8, 271–277.

Hjälmdahl, M., Varhelyi, A., 2004. Validation of in-car observations, a method for driver assessment. Transp. Res. Part A Policy Pract. 38 (2), 127–142.

Huang, H., Zhu, Y., Li, X., Li, M., Wu, M.-Y., 2010. META: a mobility model of MEtropolitan TAxis extracted from GPS traces. In: Proceedings of the 2010 IEEE Wireless Communication and Networking Conference.

ISTAT, 2009. *Incidenti Stradali*. Report of the Italian Institute of Statistics.

Janke, M.K., 1991. Accidents, mileage, and the exaggeration of risk. Accid. Anal. Prev. 23 (2–3), 183–188.

Jolliffe, I., 2005. Principal component analysis. In: Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd., Hoboken, NJ.

Jovanis, P.P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. Transp. Res. Rec. 1068, 42–51.

Jovanis, P.P., Aguero-Valverde, J., Wu, K.-F., Shankar, V., 2011. Analysis of naturalistic driving event data. Transp. Res. Rec. 2236, 49–57.

Jun, J., Ogle, J., Guensler, R., 2007. Relationships between crash involvement and temporal–spatial driving behavior activity patterns: use of data for vehicles with global positioning systems. Transp. Res. Rec. 2019, 246–255.

Jun, J., Guensler, R., Ogle, J., 2010. Differences in observed speed patterns between crash-involved and crash-not-involved drivers: application of in-vehicle monitoring technology. Transp. Res. Part C Emerg. Technol. 19 (4), 569–578.

Langford, J., Koppel, S., McCarthy, D., Srinivasan, S., 2008. In defence of the "low-mileage bias". Accid. Anal. Prev. 40 (6), 1996–1999.

Lemeshow, S., Hosmer, D.W., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. Am. J. Epidemiol. 115 (1), 92–106.

Litman, T., 2011. Distance-based Vehicle Insurance Feasibility, Benefits and Costs: Comprehensive Technical Report. VTPI (www.vtpi.org). <www.vtpi.org/dbvi_com.pdf>.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44 (5), 291–305.

Lourens, P.F., Vissers, J.A., Jessurun, M., 1999. Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. Accid. Anal. Prev. 31, 593–597.

Mayou, R., Bryant, B., Duthie, R., 1993. Psychiatric consequences of road traffic accidents. Br. Med. J. 307 (6905), 647.

Musicant, Oren, Bar-gera, H., Schechtman, E., 2010. Electronic records of undesirable driving events. Transp. Res. Part F Traffic Psychol. Behav. 13 (2), 71–79.

Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika 78 (3), 691–692.

Ogle, J., Guensler, R., Elango, V., 2005. Georgia's Commute Atlanta value pricing program: recruitment methods and travel diary response rates. Transp. Res. Rec. 1931, 28–37.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R., 1996. A simulation study of the number of events per variable in logistic regression analysis. Clin. Epidemiol. 49 (12), 1373–1379.

Progressive Insurance, 2005. Texas Mileage Study: Relationship between Annual Mileage and Insurance Losses. Report.

Risk, A., Shaoul, J., 1982. Exposure to risk and the risk of exposure. Accid. Anal. Prev. 14 (5), 353–357.

Schoenfeld, D.A., Borenstein, M., 2005. Calculating the power or sample size for the logistic and proportional hazards models. J. Stat. Comput. Simul. 75 (10), 771–785.
Schulz, K.F., Grimes, D.A., 2002. Case-control studies: research in reverse. Lancet Epidemiol. Ser. 359, 431–434.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Shankar, V., Jovanis, P.P., Aguero-Valverde, J., Gross, F., 2008. Analysis of naturalistic driving data: prospective view on methodological paradigms. Transp. Res. Rec. 2061, 1–8.
Staplin, L., Gish, K.W., Joyce, J., 2008. "Low mileage bias" and related policy implications – a cautionary note. Accid. Anal. Prev. 40 (3), 1249–1252.
Stopher, P., Fitzgerald, C., Xu, M., 2007. Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation 34, 723–741.
Toledo, T., Musicant, O., Lotan, T., 2008. In-vehicle data recorders for monitoring and feedback on drivers' behavior. Transp. Res. Part C Emerg. Technol. 16 (3), 320–331.
Wahlberg, A.af., 2004. The stability of driver acceleration behavior, and a replication of its relation to bus accidents. Accid. Anal. Prev. 36 (1), 83–92.
White, S.B., 1976. On the use of annual vehicle miles of travel estimates from vehicle owners. Accid. Anal. Prev. 8 (4), 257–261.
Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary experiment to derive trip purpose from Global Positioning System travel data. Transp. Res. Rec. 1768, 125–134.
Wolf, J., Oliveira, M., Thompson, M., 2003. Impact of underreporting on mileage and travel time estimates: results from Global Positioning System-enhanced household travel survey. Transp. Res. Rec. 1854, 189–198.
Wolfe, A.C., 1982. The concept of exposure to the risk of a road traffic accident and an overview of exposure data collection methods. Accid. Anal. Prev. 14 (5), 337–340.
Wouters, I.J., Bos, J.M., 2000. Traffic accident reduction by monitoring driver behaviour with in-car data recorders. Accid. Anal. Prev. 32 (5), 643–650.
Wu, K.-F., Jovanis, P.P., 2012. Crashes and crash-surrogate events: exploratory modeling with naturalistic driving data. Accid. Anal. Prev. 45, 507–516.
Zhang, D., Li, N., Zhou, Z.-H., Chen, C., Sun, L., Li, S., 2011. IBAT: detecting anomalous taxi trajectories from GPS traces. In: Ubiquitous Computing. ACM, Beijing, China, pp. 99–108.