

Manuscript Number:

Title: Multivariate Exposure Modeling of Accident Risk: Insights from  
Pay-as-you-drive Insurance Data

Article Type: Research Paper

Keywords: GPS trajectories; In-vehicle data recorder; risk exposure;  
accident research; Pay-as-you-drive insurance; multivariate logistic  
regression; low-mileage bias.

Corresponding Author: Mr. Johannes Paefgen,

Corresponding Author's Institution: University of St. Gallen

First Author: Johannes Paefgen

Order of Authors: Johannes Paefgen; Thorsten Staake, Dr.; Elgar Fleisch,  
Prof. Dr.

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.

## Highlights

- We extract exposure metrics from location trajectories of 1,600 vehicles
- Metrics are used in multivariate logistic regression to predict accident involvement
- After various transformations, a Nagelkerke  $R^2$  goodness-of-fit of 0.646 is achieved
- Influence of driving conditions and the mileage-risk relationship are discussed
- Study shows that Pay-as-you-drive insurance data can yield novel insights

## **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 & Hagvall 1976), today's technology exhibits higher performance on a miniature scale and at a fraction of the cost. Data collection units 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; Oren Musicant et al. 2010; Jun et al. 2010). Various researchers have furthermore demonstrated 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; Wolf et al. 2001; Blanchard et al. 2010; Forrest & Pearson 2005; Stopher et al. 2007).

One important research application of IVDR data is the analysis of antecedents and consequences of accident involvement. This is a major objective of prominent naturalistic driving studies, which unobtrusively collect data from large vehicle samples over long periods with sophisticated sensing solutions that include positioning, acceleration, and video data (Jovanis et al. 2011; Shankar et al. 2008; Gordon et al. 2011; Tarko 2012; K.-F. Wu & Jovanis 2012). While these studies yield an unprecedented level of detail with regard to driving behavior, they also require extensive financial resources and personnel for the equipment of vehicles, provision of information processing systems, and the administration of study participants. Bringing these costs in relationship to the size of acquired samples, one arrives at per-vehicle investments in naturalistic driving studies that can exceed USD 20,000.

Vehicle fleets equipped with IVDR for commercial purposes constitute an alternative source of data, featuring large sample sizes without saddling researchers with high equipment costs. For this paper, we utilize IVDR data made available by PAYD insurance providers. Under PAYD insurance, premiums are calculated based on the actual vehicle usage instead of conventional lump-sum payments, thereby

improving actuarial risk differentiation and incentivizing risk mitigation by policyholders (Desyllas & Sako 2012). For premium calculation, data is collected (i) with auditable quality, (ii) in sample sizes up to hundreds of thousands of vehicles, and (iii) over prolonged observation periods. A general objective of the paper at hand is to demonstrate the potential use that IVDR data obtained from PAYD insurance providers has in accident research.

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 & 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 1,000 (control group) and 600 (case group) vehicles, collected over 24 and 6 months, respectively. 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. The model is validated through comparison with official accident statistics. Furthermore, we obtain insights regarding the characteristic of the non-linear 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.

## **2. RELATED WORK**

### **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 reached representative scales only recently (Table 1). The Texas Mileage study published by Progressive Insurance, a PAYD provider in the US, may be considered 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 mileage 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 a naturalistic driving trial conducted between 2001 and 2006 by the US National Highway Traffic and Safety Administration (Gordon et al. 2011; Jovanis et al. 2011; Tarko 2012; K.-F. Wu & Jovanis 2012; Shankar et al. 2008). In the trial, kinematic measurements were used to identify critical driving events, which were 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 3,000 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. In Europe, proposals for similar studies are currently under discussion, such as the PROLOGUE, DaCoTa and 2BESAFE initiatives under the 7<sup>th</sup> 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. A lack of standardization with regard to which data is collected and how it is reported makes it difficult to compare the results across these studies and with the previously mentioned studies. An additional limitation is that a majority of these studies come from US driver populations, while Asian and European samples appear underrepresented. While several IVDR studies in China used extensive samples of taxi fleets that exceeded more than 7,600 vehicles (Huang et al. 2010; Zhang et al. 2011), these were not evaluated from an accident analysis perspective.

In our opinion, current IVDR-based research activities exhibit a shortcoming when it comes to the modeling of accident risk based on the accumulated exposure of vehicles. This also does not appear to be a research objective among current and future naturalistic driving studies. The Texas Mileage Study – as the only study with this focus of substantial sample size – does not take a differentiated perspective on exposure modeling and as this paper went to press had not been published in peer-reviewed outlets. To make this point clearer, we discuss the conceptualization of exposure to accident risk for IVDR-based studies in the next section.

Project Name	Sample Description	Observation Period	Region	Reference(s)	Relevant Research
SAMOVAR	270 commercial vehicles	24 months	Netherlands	(Wouters & Bos 2000)	Effect of driver feedback
n.a.	61 private drivers, 1 IVDR-equipped vehicle	60 minutes	Virginia, US	(Boyce & Geller 2002)	Analysis of critical driving events
ISA Trial	4,840	18 months	Sweden	(Hjälmdahl & Varhelyi 2004) (Biding & Lind 2002)	Evaluation of different intelligent speed adaptation devices
n.a.	125 bus drivers in 1 IVDR-equipped vehicle	approx. 30 minutes	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
DriveDiagnostics 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 critical driving events
NHTSA / VDOT Naturalistic Driving Study	100 private and commercial vehicles	12 months	Washington DC and Virginia, US	(Shankar et al. 2008) (Gordon et al. 2011) (Jovanis et al. 2011) (Tarko 2012) (K.-F. Wu & Jovanis 2012)	Analysis of critical driving events
SHRP 2 Naturalistic Driving Study (Planned)	3,000 private and commercial vehicles	24 months	6 states in the US	n.a.	n.a.
Presented Study	1,600 private and commercial vehicles	6 months (case group) 24 months (control group)	Northern Italy	n.a.	Comparison of exposure across various driving conditions

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

## 2.2. Conceptualization of Accident Risk Exposure

The term *exposure* appears frequently in transportation research publications yet appears to lack an unambiguous, overarching definition. A common interpretation of exposure 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 (M. Chipman 1982; Foldvary 1975; Jovanis & Chang 1986; Lourens et al. 1999). Analogously, exposure can also refer to the driving duration of a vehicle (Wolfe 1982; M. L. Chipman et al. 1992; M. L. Chipman et al. 1993). While the Texas Mileage Study for example referenced in the previous section found a linear relationship between mileage and insurance claims, non-linear relationships particularly for low-mileage drivers have also been discussed in the literature (Janke 1991; Staplin et al. 2008; Langford et al. 2008). A more differentiated view of exposure was introduced by (Risk & Shaoul 1982), who pointed out that the measure of vehicle mileage “seems to reflect mainly the extent rather than the degree of accident risk 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.

While mileage is comparatively simple to obtain even for large vehicle samples, differentiated data on driving exposure was 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.

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; However, they also 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. The objective of the paper at hand is to devise a remedy for these issues, both from a methodological and an empirical perspective.

### **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 travelled distance from incremental position updates and generated new data entries every segment of 2,000 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 2,000 m when vehicle ignition was turned off and the current trip ended. Segment distance could in some cases exceed the 2,000 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. Figure 1 gives an example of segment end locations plotted for a single vehicle in the sample accumulated over 24 months. 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 locations, which distinguished urban roads, extra-urban roads, and highways. 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).

< Figure 1 about here >

### **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 & 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 six 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. 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 1,000 vehicles from the data pool with twenty-four months of location data without accident involvement throughout this period, spanning from July 2007 to June 2009.

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

- Further accident events in the six-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 travelled 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  $45.7 \times 10^6$  kilometers driven distance in  $1.0 \times 10^6$  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 1,567 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 Figure 2.

*< Figure 2 about here >*

Choosing reasonable criteria for the exposure conditions that determine the columns of matrix  $\mathbf{E}$  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 2,520 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.

<b>Time of day</b>	<b>Day of week</b>	<b>Road type</b>	<b>Velocity interval</b>	<b><math>\Sigma</math></b>
24 columns	7 columns	3 columns	5 columns	39 columns

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

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 percent 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 Figure 3.

*< Figure 3 about here >*

#### **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,

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

Using maximum likelihood methods, the intercept  $\beta_0$  and the coefficients  $\beta_i$  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 & Snell 1968) and (Nagelkerke 1991), as well as the Hosmer-Lemeshow test statistic (Lemeshow & 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 & 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 1,567 vehicles, this renders feasible the

estimation of models with up to 43 predictors.

#### **4.1. Full Variable Set**

For an initial model, we consider the entire variable set contained in the exposure matrix  $\mathbf{E}$  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 & 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 ( $> 2,800$ ), 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 Figure 4. 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 (00-05h, 05-18h, 18-21h, and 21-24h) and (ii) to merge the day of week variables into two new intervals (Monday through Thursday, and Friday through Sunday).

< Figure 4 about here >

After three iterations, the logistic regression model contains 9 variables, which are given in Table 3 together with  $\beta$ -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 & Snell) and 0.482 (Nagelkerke), goodness-of-fit is slightly lower than in the previously discussed model, and the Hosmer-Lemeshow  $\chi^2$  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 **Error! Reference source not found.** 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 Figure 3, we plot the observed cumulative probability of untransformed and LN-transformed variables against the cumulative probability of a normal distribution in Figure 5. The benefit of LN-transformation was not equal for all exposure variables, although it was evident for a majority.

< Figure 5 about here >

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  $\beta$ -coefficients, error terms and  $p$ -values. The model achieves improved pseudo- $R^2$  values of 0.454 (Cox & 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 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 Table 5 and Table 6. For the 5-bin mileage variable, pseudo- $R^2$  values are approximately the same as 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 & Snell) and 0.646 (Nagelkerke) and the Hosmer-Lemeshow  $\chi^2$  of 4.097 tested highly insignificant ( $p = 0.848$ ).



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

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

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

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

<b>Variable</b>	<b>Coefficient</b>	<b>Standard Error</b>	<b>Significance</b>
<b>Time of day (LN-transformed)</b>			
05-18h	<i>-.424</i>	<i>.118</i>	<i>&lt;.001</i>
18-21h	<i>.720</i>	<i>.164</i>	<i>&lt;.001</i>
<b>Day of week (LN-transformed)</b>			
Friday through Sunday	<i>-.352</i>	<i>.125</i>	<i>.005</i>
<b>Road type (LN-transformed)</b>			
Urban	<i>.679</i>	<i>.152</i>	<i>&lt;.001</i>
<b>Velocity interval (LN-transformed)</b>			
0-30 km/h	<i>.455</i>	<i>.141</i>	<i>.001</i>
60-90 km/h	<i>-.655</i>	<i>.128</i>	<i>&lt;.001</i>
90-120 km/h	<i>.292</i>	<i>.165</i>	<i>.077</i>
<b>Avg. monthly mileage (categorical)</b>			
< 1,021 km	<i>0 (reference)</i>	<i>-</i>	<i>-</i>
1,021 - 1,585 km	<i>.933</i>	<i>.357</i>	<i>.009</i>
1,585 - 2,522 km	<i>2.989</i>	<i>.339</i>	<i>&lt;.001</i>
2,522 - 3,966 km	<i>4.979</i>	<i>.371</i>	<i>&lt;.001</i>
> 3,966 km	<i>6.478</i>	<i>.483</i>	<i>&lt;.001</i>
<b>Constant</b>	<i>-3.997</i>	<i>.322</i>	<i>&lt;.001</i>

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

<b>Variable</b>	<b>Coefficient</b>	<b>Standard Error</b>	<b>Significance</b>
<b>Time of day (LN-transformed)</b>			
05-18h	-.397	.123	.001
18-21h	.698	.168	<.001
<b>Day of week (LN-transformed)</b>			
Friday through Sunday	-.266	.133	.046
<b>Road type (LN-transformed)</b>			
Urban	.737	.163	<.001
Highway	-.289	.142	.042
<b>Velocity interval (LN-transformed)</b>			
0-30 km/h	.380	.152	.012
60-90 km/h	-.733	.137	<.001
90-120 km/h	.306	.177	.084
<b>Avg. monthly mileage (categorical)</b>			
< 767 km	0 (reference)	-	-
676 - 1,021 km	.361	.576	.531
1,021 – 1,274km	1.023	.544	.060
1,274 - 1,585 km	1.394	.529	.008
1,585 – 2,000 km	2.623	.508	<.001
2,000 – 2,522 km	4.075	.514	<.001
2,522 – 3,188 km	4.672	.523	<.001
3,188 - 3,966 km	6.539	.579	<.001
3,966 – 5,705 km	7.193	.637	<.001
> 5,705 km	7.032	.721	<.001
<b>Constant</b>	-4.418	.475	<.001

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

#### 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

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

where  $k$  is the number of variables in the model,  $n = 1,567$  the sample size, and  $L$  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  $\beta$ -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 hours, while higher for the interval between 18:00 and 21:00 hours.
- Driving exposure accumulated on weekends, including Fridays, is associated with lower risk.
- 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 to 90 km/h) has the lowest risk of accident involvement, while both the lowest velocity interval (0 to 30 km/h) and the second-highest interval (90 to 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 (Figure 6 and 7). 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 hours, 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 hours. This confirms the validity of our model and demonstrates the additional insights accessible through an exposure-based analysis of road accident involvement.

*< Figure 6 and 7 about here >*

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  $\beta$ -coefficients of the corresponding dummy variables in Figure 8. Coefficient values monotonously increase up to 4,000 km but slightly decrease for the top interval. Up to this point, two segments are distinguishable. Below approximately 1,600 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 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.

*< Figure 8 about here >*

Model	-2 Log likelihood	k	Information Criteria		Pseudo-R <sup>2</sup>		Hosmer-Lemeshow Test	
			AIC	BIC	Cox & Snell	Nagelkerke	$\chi^2$	Sig.
Full variable set (35 out of 40)	1,301.044	35	1,441.044	1,558.536	.387	.528	8.539	.383
Merged-interval	1,385.644	10	1,425.644	1,459.213	.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	1,014.211	.453	.612	10.026	.263
Merged-interval, LN transformed, 10-cat. mileage	870.506	18	906.506	1,002.93	.478	.646	4.097	.848

Table 7. Goodness-of-fit measures for logistic regression 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 1,600 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 & 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. Additional Remarks**

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.

## **6. REFERENCES**

- 2nd Strategic Highway Research Program, 2012. Naturalistic Driving Study. Available at: <http://www.shrp2nds.us/> [Accessed November 23, 2012].
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6), pp.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. a, Myers, A.M. & Porter, M.M., 2010. Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers. *Accident Analysis & Prevention*, 42(2), pp.523–529.
- Boyce, T.E. & Geller, E.S., 2002. An instrumented vehicle assessment of problem behavior and driving style: do younger males really take more risks. *Accident Analysis & Prevention*, 34(1), pp.51–64.
- Chipman, M., 1982. The role of exposure, experience and demerit point levels in the risk of collision. *Accident Analysis & Prevention*, 14(6), pp.475–483.
- Chipman, M.L. et al., 1993. The role of exposure in comparisons of crash risk among different drivers and driving environments. *Accident Analysis & Prevention*, 25(2), pp.207–211.
- Chipman, M.L. et al., 1992. Time vs. distance as measures of exposure in driving surveys. *Accident Analysis & Prevention*, 24(6), pp.679–684.
- Christensen, R., 1997. *Log-linear models and logistic regression* 2nd ed., New York: Springer.
- Cox, D.R. & Snell, E.J., 1968. A general definition of residuals. *Journal of the Royal Statistical Society, Series B (Methodological)*, 30(2), pp.248–275.
- Desyllas, P. & Sako, M., 2012. Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance. *Research Policy*, 42(1), pp.101–116.
- Foldvary, L., 1975. Road accident involvement per miles travelled, part II. *Accident Analysis & Prevention*, 11(2), pp.191–205.
- Forrest, T.L. & Pearson, D.F., 2005. Comparison of trip determination methods in household travel surveys enhanced by a global positioning system. *Transportation Research Record: Journal of the Transportation Research Board*, 1917, pp.63–71.
- Gordon, T.J. et al., 2011. Analysis of crash rates and surrogate events. *Transportation Research Record: Journal of the Transportation Research Board*, 2237, pp.1–9.
- Helander, M. & Hagvall, B., 1976. An instrumented vehicle for studies of driver behavior. *Accident Analysis & Prevention*, 8, pp.271–277.

- Hjälmdahl, M. & Varhelyi, A., 2004. Validation of in-car observations, a method for driver assessment. *Transportation Research Part A: Policy and Practice*, 38(2), pp.127–142.
- Huang, H. et al., 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. *Accident Analysis & Prevention*, 23(2-3), pp.183–188.
- Jolliffe, I., 2005. Principal component analysis. In *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ: John Wiley & Sons, Ltd.
- Jovanis, P.P. et al., 2011. Analysis of naturalistic driving event data. *Transportation Research Record: Journal of the Transportation Research Board*, 2236, pp.49–57.
- Jovanis, P.P. & Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. *Transportation Research Record: Journal of the Transportation Research Board*, 1068, pp.42–51.
- 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. *Transportation Research Part C: Emerging Technologies*, 19(4), pp.569–578.
- 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. *Transportation Research Record: Journal of the Transportation Research Board*, 2019, pp.246–255.
- Langford, J. et al., 2008. In defence of the “low-mileage bias”. *Accident Analysis & Prevention*, 40(6), pp.1996–1999.
- Lemeshow, S. & Hosmer, D.W., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115(1), pp.92–106.
- Lord, D. & Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), pp.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. *Accident Analysis & Prevention*, 31, pp.593–597.
- Mayou, R., Bryant, B. & Duthie, R., 1993. Psychiatric consequences of road traffic accidents. *British Medical Journal*, 307(6905), p.647.
- Musicant, Oren, Bar-gera, H. & Schechtman, E., 2010. Electronic records of undesirable driving events. *Transportation Research Part F: Psychology and Behaviour*, 13(2), pp.71–79.

- Nagelkerke, N.J.D., 1991. A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78(3), pp.691–692.
- Ogle, J., Guensler, R. & Elango, V., 2005. Georgia's Commute Atlanta value pricing program: Recruitment methods and travel diary response rates. *Transportation Research Record: Journal of the Transportation Research Board*, 1931, pp.28–37.
- Peduzzi, P. et al., 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), pp.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. *Accident Analysis & Prevention*, 14(5), pp.353–357.
- Schoenfeld, D.A. & Borenstein, M., 2005. Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation*, 75(10), pp.771–785.
- Schulz, K.F. & Grimes, D.A., 2002. Case-control studies: research in reverse. *The Lancet Epidemiology Series*, 359, pp.431–434.
- Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics*, 6(2), pp.461–464
- Shankar, V. et al., 2008. Analysis of naturalistic driving data: Prospective view on methodological paradigms. *Transportation Research Record: Journal of the Transportation Research Board*, 2061, pp.1–8.
- Staplin, L., Gish, K.W. & Joyce, J., 2008. “Low mileage bias” and related policy implications—a cautionary note. *Accident Analysis & Prevention*, 40(3), pp.1249–1252.
- Stopher, P., FitzGerald, C. & Xu, M., 2007. Assessing the accuracy of the Sydney Household Travel Survey with GPS. *Transportation*, 34, pp.723–741.
- Tarko, A.P., 2012. Use of crash surrogates and exceedance statistics to estimate road safety. *Accident Analysis & Prevention*, 45, pp.230–240.
- Toledo, T., Musicant, O & Lotan, T., 2008. In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transportation Research Part C: Emerging Technologies*, 16(3), pp.320–331.
- Wahlberg, A., 2004. The stability of driver acceleration behavior, and a replication of its relation to bus accidents. *Accident Analysis & Prevention*, 36(1), pp.83–92.
- White, S.B., 1976. On the use of annual vehicle miles of travel estimates from vehicle owners. *Accident Analysis & Prevention*, 8(4), pp.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. *Transportation Research Record: Journal of the Transportation Research Board*, 1768, pp.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. *Transportation Research Record: Journal of the Transportation Research Board*, 1854, pp.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. *Accident Analysis & Prevention*, 14(5), pp.337–340.
- Wouters, I.J. & Bos, J.M., 2000. Traffic accident reduction by monitoring driver behaviour with in-car data recorders. *Accident Analysis & Prevention*, 32(5), pp.643–650.
- Wu, K.-F. & Jovanis, P.P., 2012. Crashes and crash-surrogate events: exploratory modeling with naturalistic driving data. *Accident Analysis & Prevention*, 45, pp.507–516.
- Zhang, D. et al., 2011. iBAT: Detecting anomalous taxi trajectories from GPS traces. In *Ubiquitous Computing*. Beijing, China: ACM, pp. 99–108.

## **FIGURE CAPTIONS**

*Figure 1. Exemplary vehicle location trajectory, segment end-points (coordinate axes omitted for privacy reasons).*

*Figure 2. Exposure aggregation process.*

*Figure 3. Histograms of average monthly mileage (top) and the fraction of mileage accumulated within the 60-90 km/h velocity interval (bottom).*

*Figure 4. P-P Plots for (a) avg. monthly mileage, (b) the same after LN-transformation, (c) relative exposure at 60-90 km/h, and (d) the same after LN-transformation.*

*Figure 5. Scree plots for daytime (top) and weekday (bottom) exposure factors from PCA.*

*Figure 6. 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).*

*Figure 7. Non-linear progression of coefficients of categorical dummy-variables with average monthly mileage.*

Figure 1.tif  
[Click here to download high resolution image](#)



Figure 2.tif

[Click here to download high resolution image](#)

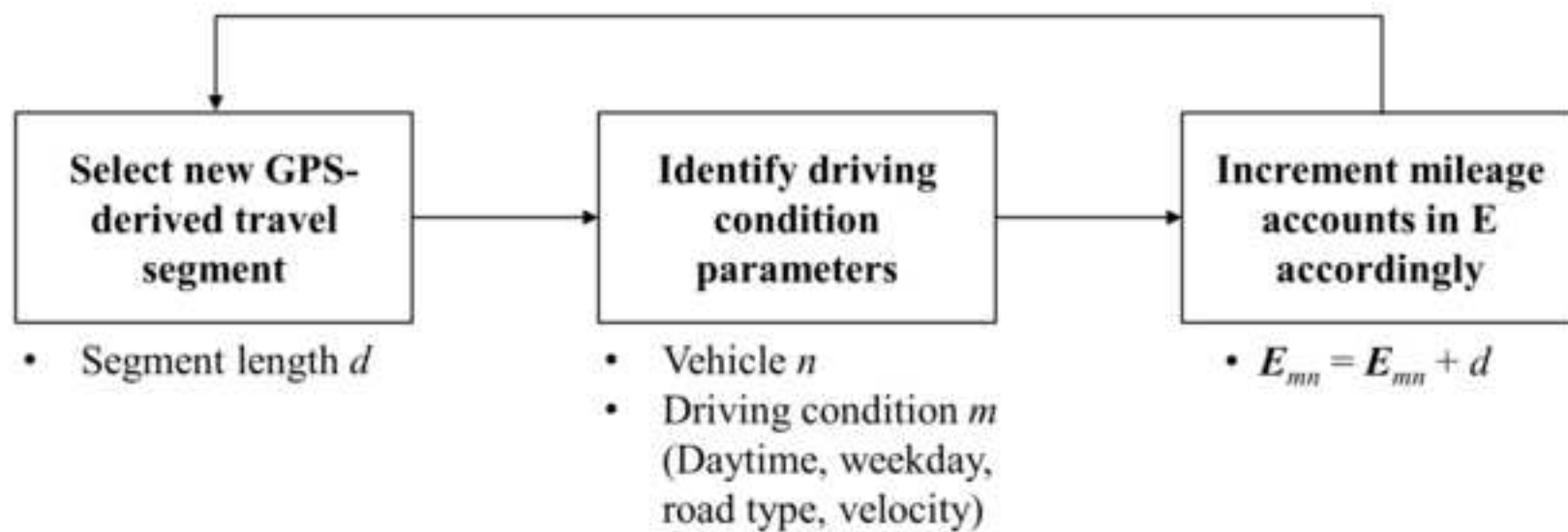




Figure 3.tif

[Click here to download high resolution image](#)

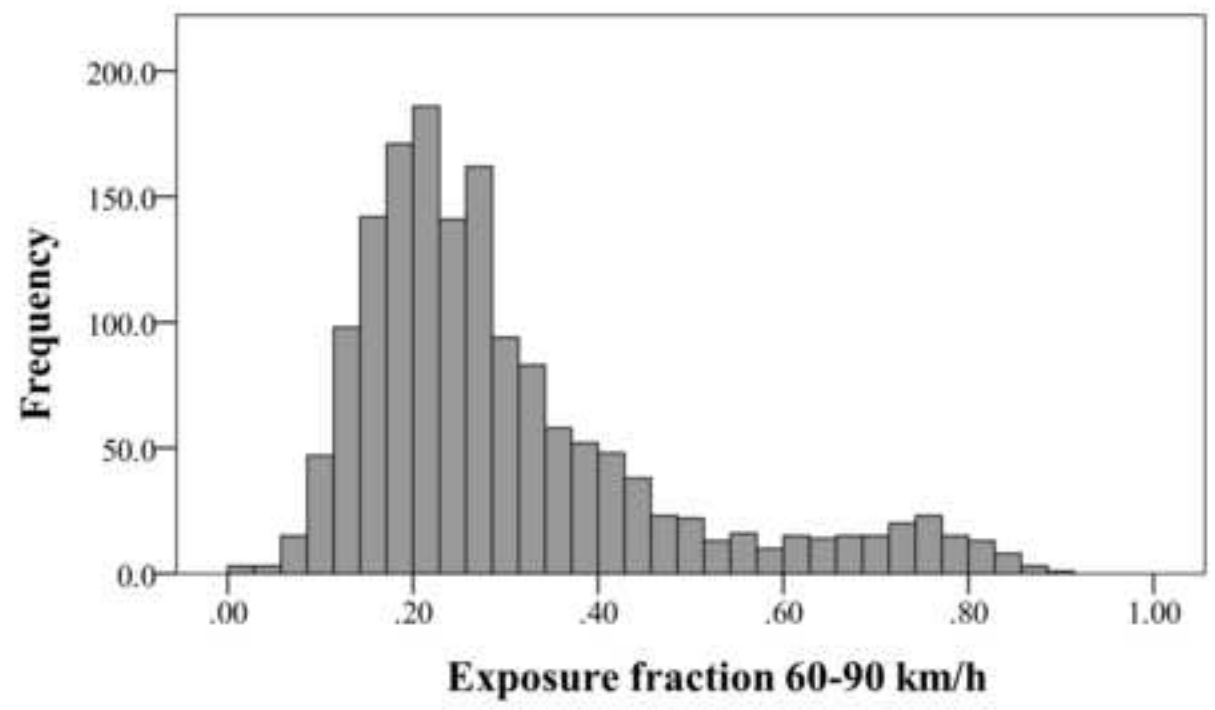
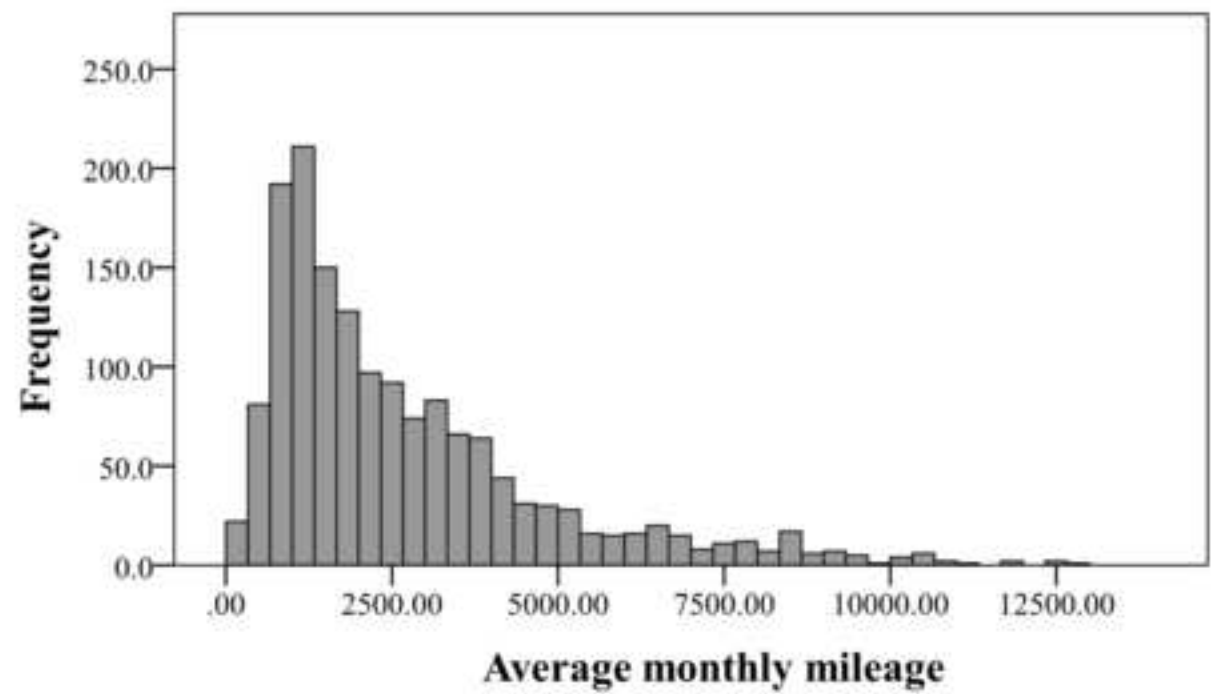


Figure 4.tif

[Click here to download high resolution image](#)

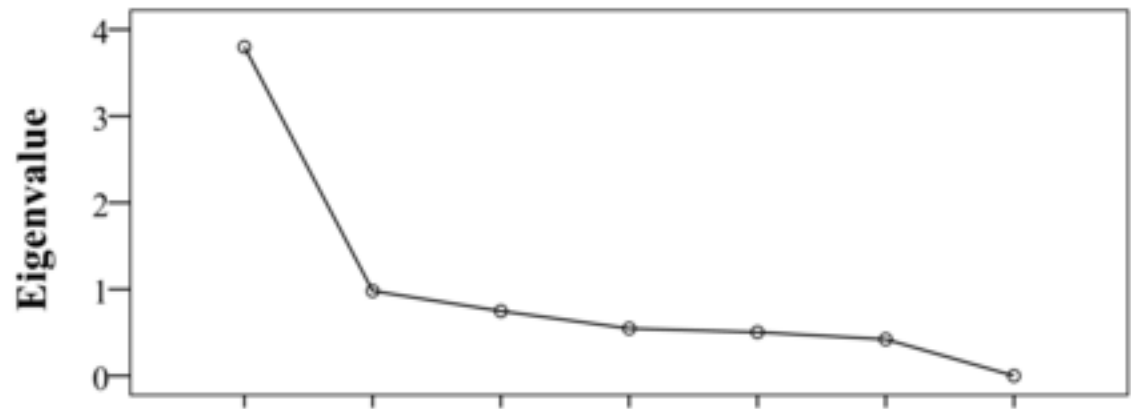
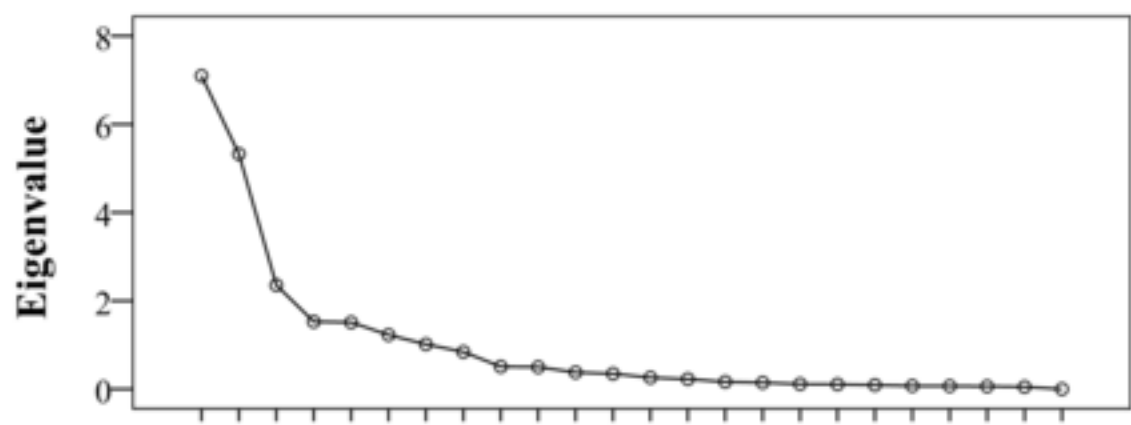


Figure 5.tif

[Click here to download high resolution image](#)

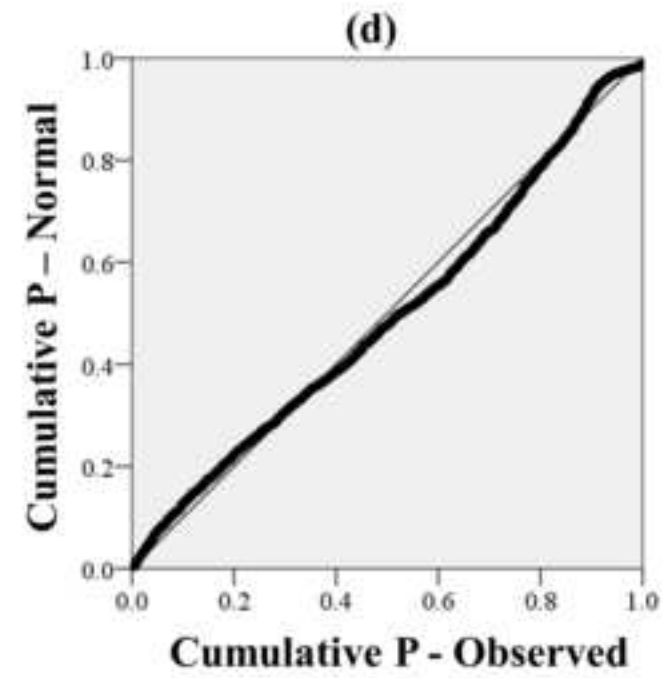
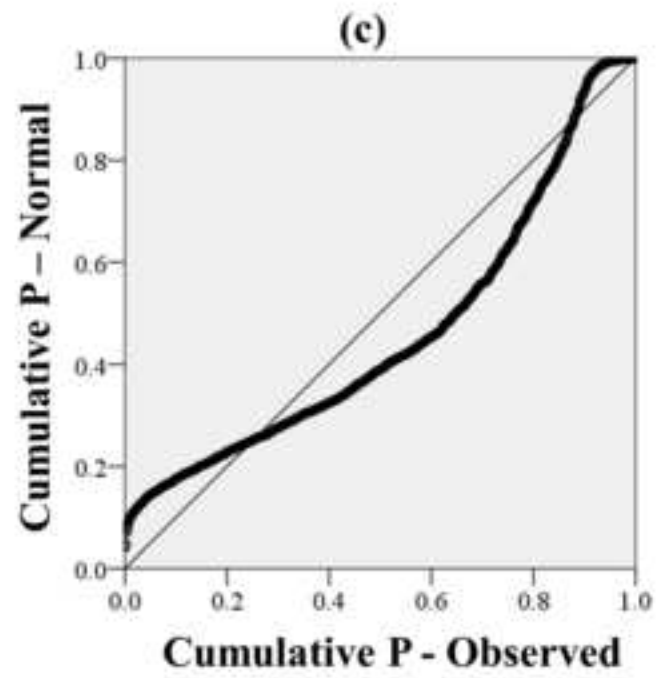
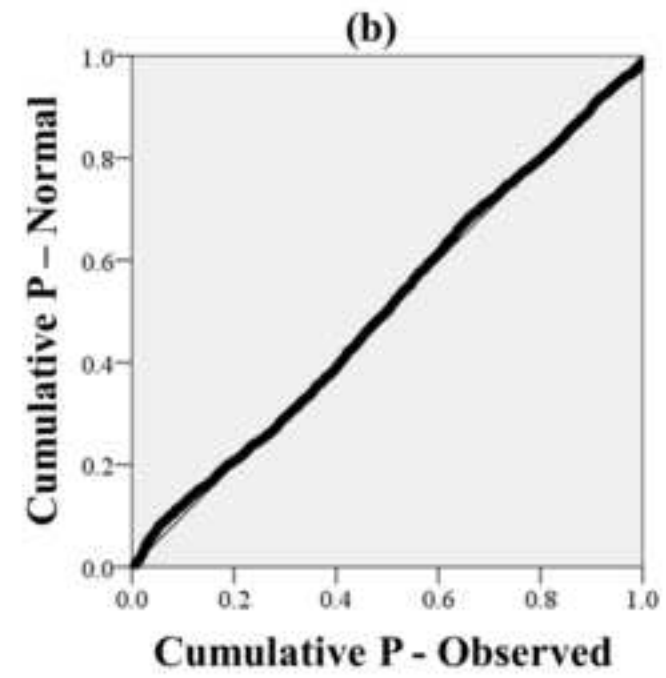
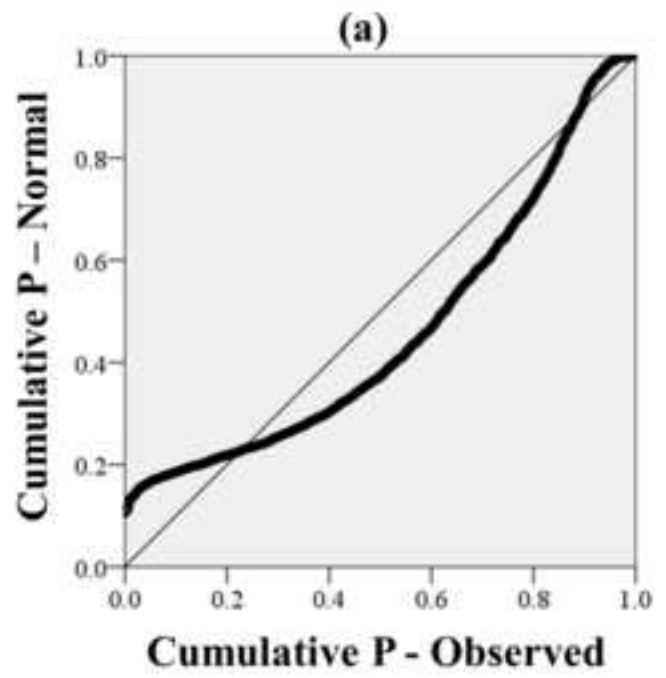


Figure 6.tif

[Click here to download high resolution image](#)

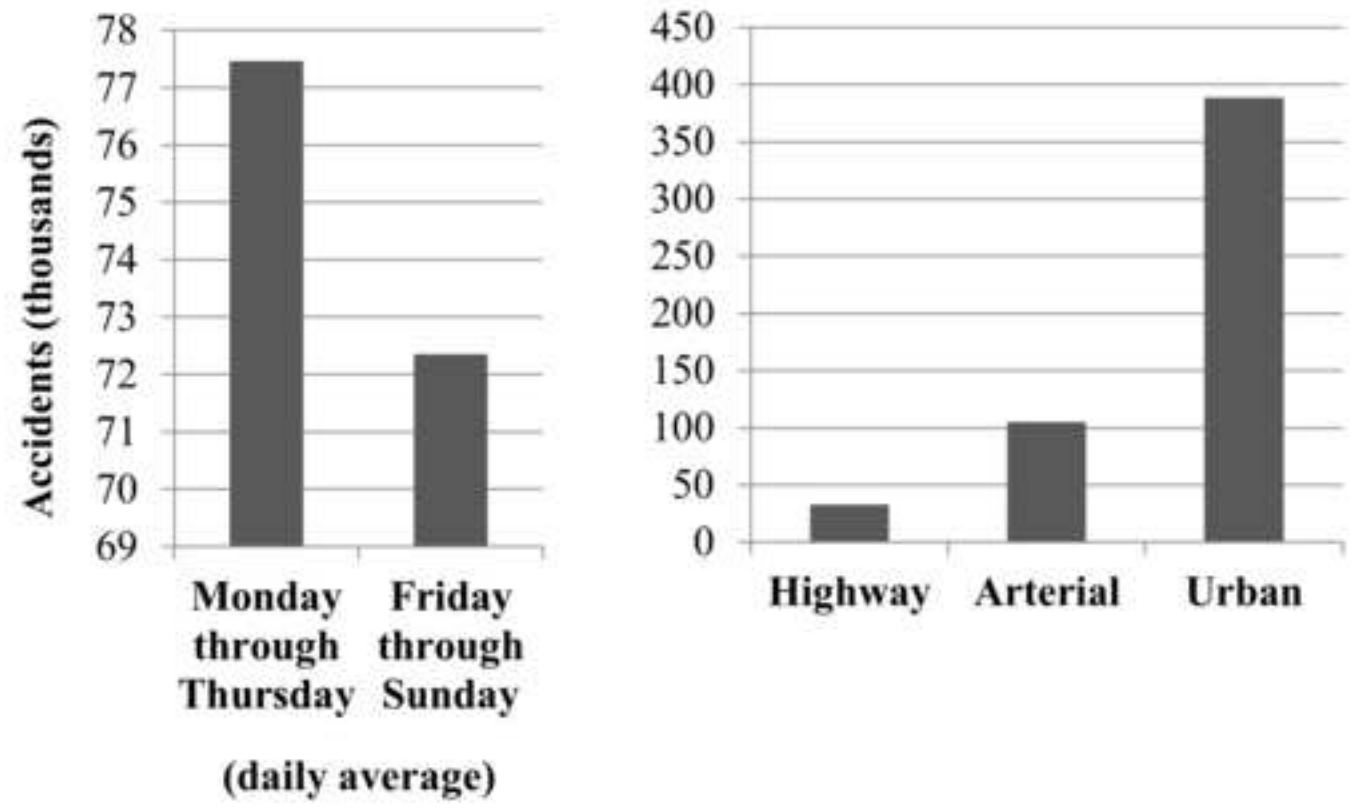


Figure 7.tif

[Click here to download high resolution image](#)

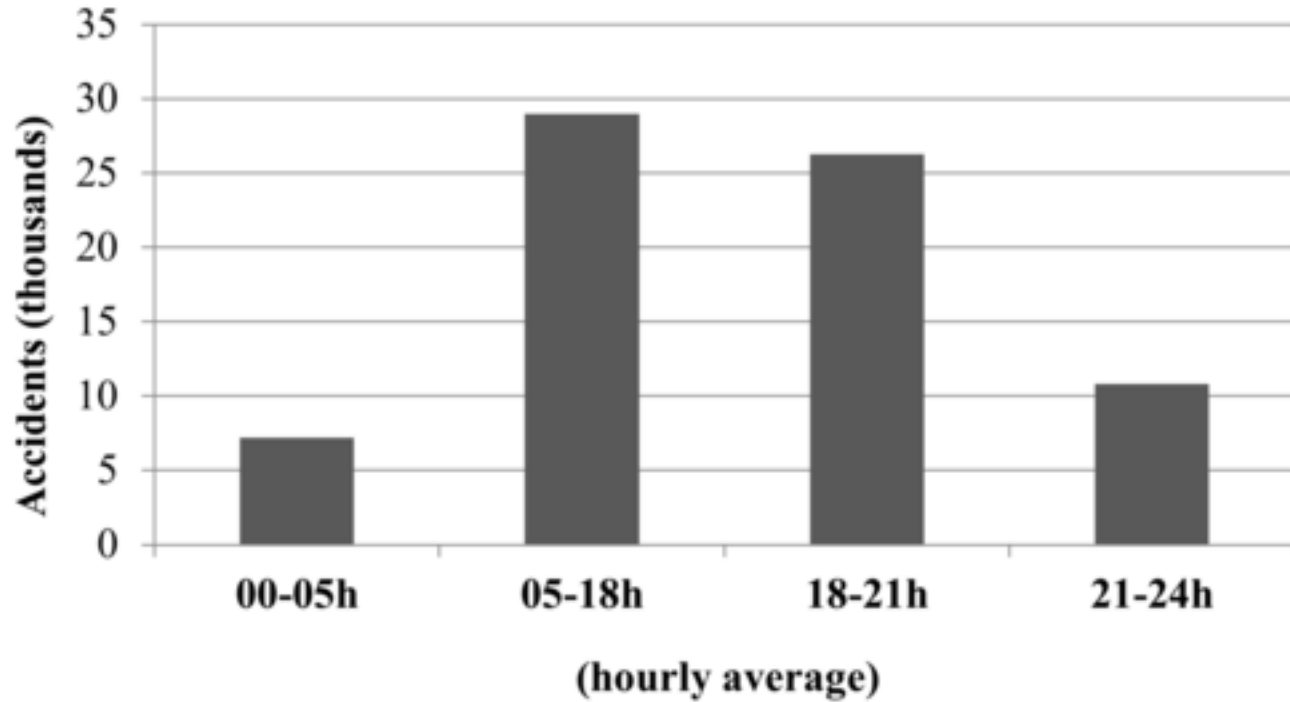


Figure 8.tif

[Click here to download high resolution image](#)

