Transportation Research Part A xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Transportation Research Part A

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

Spatial prediction of traffic accidents with critical driving events – Insights from a nationwide field study

Benjamin Ryder[a,*], Andre Dahlinger[b], Bernhard Gahr[b], Peter Zundritsch[a], Felix Wortmann[b], Elgar Fleisch[a,b]

[a] *Information Management, ETH Zurich, Weinbergstrasse 58, 8092 Zurich, Switzerland*
[b] *ITEM-HSG, University of St. Gallen, Dufourstrasse 40a, 9000 St. Gallen, Switzerland*

ARTICLE INFO

Keywords:
Naturalistic driving data
Critical driving events
Crash rate
Crash frequency
Traffic safety
Spatial regression

ABSTRACT

Despite the fact that semi-autonomous vehicles will become more and more prevalent in the coming decades, recent studies have highlighted that traffic accidents will persist as a core issue for road users, insurers, and policy makers alike. Researchers and industry players see potential in the technology embedded in semi-autonomous vehicles to combat this challenge by reliably predicting locations with a high likelihood of traffic accidents. This technology can be leveraged to detect accidents and 'near miss incidents', such as heavy braking and evasive manoeuvres, otherwise known as Critical Driving Events (CDEs). The locations of CDEs could identify areas of high accident exposure, offering automotive insurers a unique opportunity to reduce traffic accidents through the adoption of active loss prevention business models, such as providing safe-routing services and in-vehicle warnings. To date, there is limited empirical evidence on whether the Crash Frequency and Crash Rate of locations can be accurately identified through CDEs. To address this research gap, an 18-week naturalistic driving field study of 72 vehicles was conducted in Switzerland, covering over 690,000 km. Data collected from the CAN Bus of these vehicles indicate that there is a proportional relationship between the CDEs of the fleet, and the Crash Frequency and Crash Rate of a location. Furthermore, a nationwide spatial regression analysis was applied to determine Crash Frequency across the majority of the Swiss road network. We identify the relationship between Crash Frequency, and the CDEs and Trip Frequency of the fleet, along with additional explanatory variables for urban and highway locations. These insights provide first evidence that insurance companies and other industry players with access to a nationwide semi-autonomous fleet can determine existing and emerging locations of high accident probability, enabling more proactive business models and safety focused services.

1. Introduction

Within the insurance industry, the ever-increasing digitisation of the physical world has, until recently, primarily been viewed as an advanced concept with limited short- and mid-term impact (Berger, 2015). Hence, many insurers adopted a cautious attitude to this new technology paradigm. Meanwhile, early-adopters have been disrupting the way insurers traditionally conduct business by demonstrating how vehicle sensors, along with wearable technology, Global Position System (GPS), mobile and modern telematics can proactively engage policyholders and help improve risk assessment in loss prevention (Koenig et al., 2016). This trend is likely to

* Corresponding author.
E-mail addresses: bryder@ethz.ch (B. Ryder), andre.dahlinger@unisg.ch (A. Dahlinger), bernhard.gahr@unisg.ch (B. Gahr), zpeter@student.ethz.ch (P. Zundritsch), felix.wortmann@unisg.ch (F. Wortmann), elgar.fleisch@unisg.ch (E. Fleisch).

https://doi.org/10.1016/j.tra.2018.05.007

continue, with the rise of fully- and semi-autonomous vehicles triggering fundamental change and the adoption of inexpensive mobility on-demand services (Krueger et al., 2016). In the insurance industry, as advice-led customer interactions become both more frequent and in real-time, a shift is starting as companies pivot more toward active loss prevention (Reifel et al., 2014). As such, many insurers are moving away from purely 'reactive' business models and incorporating 'preventative' measures into their products in order to help their customers and cut long-term costs. Some of the first companies to adopt this approach were insurers in the health domain, which have introduced preventive actions into their portfolios, such as offering customers incentives to engage in healthy behaviour (CSS Insurance, 2017; Sanitas, 2017). Likewise, in the automotive insurance market, safer driving can be rewarded, and detected through in-vehicle sensors, retrofit dongles and smartphone applications that measure dangerous driving patterns, such as speeding and swerving (AXA, 2017).

The advent of fully-autonomous vehicles brings the promise of a new era of traffic safety, where the frequency of road accidents can be drastically reduced, and potentially eliminated entirely (Fagnant and Kockelman, 2015). However, even in the most advanced markets, it will take decades to make this vision a reality. More specifically, recent predictions indicate that it is unlikely that the majority of the light-duty vehicle fleet in the U.S. will be capable of full self-driving automation by the year 2045 (Bansal and Kockelman, 2017). As such, semi-autonomous vehicles will become the dominating paradigm in the coming years, and traffic accidents from the manual driving of these vehicles will persist as a key insurance issue (Albright et al., 2016). As semi-autonomous vehicles becoming increasingly common, customers will expect a shift from the traditional model-based approaches of accident prediction and 'rough proxies' of exposure (Sheehan et al., 2017), to a crowd-sourcing approach using the real-time collection and analysis of data from these vehicles. Semi-autonomous vehicles are enabled by a wealth of technology capabilities, including network connectivity, high precision sensor data from the vehicle's Controller Area Network (CAN) Bus, GPS, high definition cameras, and LIDAR systems (Mannering and Washburn, 2012). These technologies can be leveraged to detect accidents and 'near miss incidents', such as heavy braking and evasive manoeuvres, otherwise known as Critical Driving Events (CDEs). The data from these incidents could be employed by insurers to determine existing and emerging locations of high accident probability and enable more proactive business models. For example, usage-based incentives for taking 'safe routes' could be provided and pay-how-you-drive insurance plans could be optimised (Sheehan et al., 2017). Furthermore, with the knowledge that a potentially dangerous location is ahead, a semi-autonomous vehicle might drive in a more cautious mode to reduce risk or hand over control to the driver to transfer insurance liability. In addition, recent research studies of in-vehicle warning systems have shown that drivers themselves can be encouraged to adapt their driving behaviour at potentially dangerous locations (Kazazi et al., 2015; Ryder et al., 2017; Tey et al., 2011; Werneke and Vollrath, 2013). Ultimately, insurance companies can collaborate with road authorities to improve the road infrastructure that contribute to dangerous locations, and manufactures to better understand vehicle capabilities and advance safety focused offerings (Sheehan et al., 2017).

As such, if insurers wish to accurately measure and encourage safer driving and progress further toward more preventative business models, then the ability to identify locations on the road network that carry a high risk of accident occurrence, so called 'blackspots', 'sites with promise', or 'hotspots' (Cheng and Washington, 2005), is of utmost value. In this regard, there are two common measures of a roads perilousness that have been extensively researched. The first is Crash Frequency, the actual number of accidents that have occurred on a road section during a specified period of time (Deacon et al., 1974). The second is Crash Rate, a measure of accident exposure risk of vehicles on a road segment, which is typically estimated by normalising the Crash Frequency by the traffic volume, e.g. Average Daily Traffic (ADT) (Hauer and Persaud, 1984). While there is research showing that the situational factors of crashes and near-misses are strongly related, there is limited empirical data on whether the Crash Frequency and Crash Rate of potentially dangerous locations can be reliably identified through analysis of CDEs (Pande et al., 2017).

1.1. Research objectives

Whereas low levels of semi-autonomous vehicle adoption currently make the large-scale collection of CDE data a challenge, we have made preliminary analysis possible through the installation of a retrofit system analogous to such advanced vehicles. As a result, the research at hand provides early results from a nationwide field study of 72 cars covering over 690,000 km in Switzerland. These vehicles were equipped with a system that collected sensor information from each car's CAN Bus in order to gather data synonymous with that available in semi-autonomous vehicles. The study presented can be split into two parts. In a first step, we analyse an ideal setting for a sub-region of Switzerland, where data for both Crash Frequency and the volume of vehicle traffic, i.e. ADT, were available. This enabled us to determine the Crash Rate of these locations. In this setup, we investigate our first research question:

RQ1: To what extent can the Crash Rate of a location be predicted by CDE information from that location, assuming that Crash Frequency and ADT data are available?

In practise, however, regions where both Crash Frequency and ADT are available are exceedingly rare. While government bodies are now regularly collecting police-recorded Crash Frequency data across whole road networks, accurate measures for traffic frequency remain only partially available and come most often from counting stations. Since traffic frequency is necessary to calculate exposure measures such as Crash Rate, we are commonly limited in traffic safety analysis to utilising Crash Frequency as a dependent variable for all road segments in this sparse setting. However, to reliably estimate the Crash Frequency of a location on the basis of CDEs, the relationship between traffic frequency of the utilised vehicle fleet and total traffic frequency (of the overall vehicle population) must be known for each location. This dilemma can be addressed by fulfilling a key assumption: The fleet should be 'representative' of the overall population of vehicles. Therefore, in the second stage of our analysis we build upon the assumption that a 'representative' fleet of vehicles is available for data capture. More specifically, we assume that there is a well-defined relationship between traffic frequency of the overall population (TF_{Pit}) and traffic frequency of the sample, i.e. field study fleet (TF_{Sit}), which is independent of location *i* for a given timeframe *t* of the analysis. From this we can test the relationship between fleet-generated CDEs and the Crash Frequency variable, tackling our second research question:

RQ2: To what extent can the Crash Frequency of a location be predicted by CDE information from that location, assuming we have sparse data coverage (only Crash Frequency information is available) and a well-defined relationship between $\mathrm{TF}_{\mathrm{Sit}}$ and $\mathrm{TF}_{\mathrm{Pit}}$, i.e. a 'representative' fleet?

To address this question, we first leverage the partially available $\mathrm{TF}_{\mathrm{Pit}}$ data to test to what extent the field study fleet utilised for our analysis fulfils the 'representative fleet' assumption. We then conclude our analysis by estimating Crash Frequency on the basis of CDEs for the majority of Switzerland, where $\mathrm{TF}_{\mathrm{Pit}}$ was unavailable, utilising spatial regression.

1.2. Research implications and paper structure

This research has implications for automotive insurance companies and other industry players both today and in the near future. The presented approach empowers insurers with access to data from semi-autonomous vehicles to rapidly identify areas of high accident exposure and to improve their management of customer risk. Furthermore, we demonstrate the importance of such a fleet satisfying the 'representative fleet' assumption, and how this enables analysis where traffic frequency measurements are unavailable. Moreover, this information may have potential not only for insurers, but also for policy makers in this long-standing field, where understanding the new capabilities and the reliance of findings from recent automotive technology advances will be vital for determining suitable traffic safety approaches and strategies.

The remainder of this paper is structured as follows. In the next section, we outline the background research and proceed with an explanation of the data collection process and the field study setting. Next, we present the results of our study with regard to the relationship between Crash Rate and CDEs. We follow this with an exploration of fleet data from the field study to explain Crash Frequency, and finally demonstrate the relationship between Crash Frequency and CDEs through nationwide spatial regression, for situations in practise where traffic frequency data is unavailable. The final sections discuss the limitations of our research and how they could be addressed in future work, and the paper closes with our general conclusions.

2. Background

2.1. Accident analysis

The topic of road traffic accident analysis has been extensively researched over the past sixty years, with various methods developed to assess the need for, and impact of, road improvements (Hauer, 1996). The most common approaches have historically been non-spatial techniques, considering traffic accidents which occurred on locations defined by the underlying road structure. The so-called Crash Frequency method (Deacon et al., 1974) is probably the most fundamental identification technique of this type. In this approach, the number of accidents which occurred during a specified period determines a road segment's perilousness. Estimations for Crash Frequency typically utilise count data models at selected locations (Anastasopoulos and Mannering, 2009; Bhat et al., 2014). Moreover, other examples of this approach have examined how the frequency of highway accidents is impacted by roadway geometries (e.g. horizontal and vertical alignments), weather, and seasonal effects on the basis of a multivariate analysis (Shankar et al., 1995). Additionally, the impact of horizontal curvature and an auxiliary lane has recently been investigated (Pande et al., 2017). The Crash Rate method is similar to the concept of Crash Frequency, but provides a measure of accident exposure of vehicles as it takes traffic volume into account (Hauer and Persaud, 1984). However, there are some evident drawbacks of both the Crash Frequency and Crash Rate methods, such as not considering random fluctuations of the number of accidents and the general lack of traffic volume data for generating the Crash Rate (Yu et al., 2014). The most common models for predicting accident data, that is count data, are negative binomial and Poisson regression models and their variants (Gianfranco et al., 2017; Greibe, 2003; Lord and Mannering, 2010; Miaou, 1994; Quddus, 2008). Following this, negative binomial and hierarchical Bayesian models have been utilised to investigate the prominently discussed topic of the relationship between speed and accidents, where it was found that average speed is not associated with accident rates, when controlling for traffic volume and road geometry, however, speed variations are positively associated with accident rates (Quddus, 2013). The effects of land use on accident rates has also been modelled, with a uniform grid with $0.259\,\mathrm{km}^2$ cells, resulting in statistically significant results using negative binomial models to predict the number of accidents (Kim et al., 2006). Finally, classification analysis has been applied to create a learning model based on attributes such as road, weather and traffic conditions and social factors (Park et al., 2016). These models fit count data well, however they do not consider spatial or temporal effects.

Spatial autocorrelation is a property found across geographic space, according to which random variables at certain distances from each other are either "more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of observations" (Legendre, 1993). Thus, observations are dependent on their spatial location, and data points can be aggregated to areas such as those used in census tracts or other zoning schemes (Quddus, 2008), grid based representations, or divided into regular or irregular polygons (Wang and Kockelman, 2013). As such, models for accident analysis have been proposed to account for this spatial relationship. For example, traditional count data models have been compared to spatial lag and error models, as well as spatial Bayesian hierarchical models (Quddus, 2008). These new models address spatial heterogeneity in geographical data and have been determined to perform well across the different model types.

Alternative approaches for identifying hazardous locations include measuring injury severity and cost estimates for crashes, where various statistical methods have been investigated, such as the ordered logit and the ordered probit models (Kockelman and

Kweon, 2002; O'Donnell and Connor, 1996). These models are suitable when the dependent variable has multiple possible outcomes and a natural ordering, for example, injury severity can be encoded by "no injury (0), minor injury (1), severe injury (2), and fatal injury sustained by driver (3)" (Kockelman and Kweon, 2002). Throughout the literature, locations identified with higher values for risk measures, e.g. Crash Frequency, Crash Rate and Crash Severity, are regarded to be more hazardous than locations with lower values. There are, however, common issues in many of these studies, which are associated with the difficulties in using crash data for the analysis of traffic safety (Lord and Mannering, 2010).

Traditionally, traffic accident analysis is based on historic crash data and is restrictive in many ways, typically suffering from issues including under and over-dispersion, small sample size and underreporting of traffic accidents (Mannering and Washburn, 2012). Additionally, where it is provided, historical accident data is often only available on a deferred basis. In Switzerland, for example, such data is only made available once per year. Therefore, and related to the aforementioned problem of data scarcity, accident analyses based on such data can be severely out-of-date. Finally, traffic frequency measurements, e.g. ADT, are vital to accurately calculate exposure measures, such as Crash Rate. However, such traffic frequency measurements are expensive to generate and hence, to the best of our knowledge, mostly suffer under a very low spatial resolution or cover only small fractions of a road network.

A promising solution to the described drawbacks lies in the post-processing and aggregation of data available from the advanced sensors of semi-autonomous and other highly connected vehicles. Naturalistic driving data offers both researchers and practitioners "the potential [to] greatly expand the scope of statistical modelling and the inferences that can be drawn" when compared to the restrictive analysis possible with sparse data collected after an accident has occurred (Mannering and Bhat, 2014). As such, the modern advancements in data quantity and quality from these vehicles has focused many research endeavours toward analysing critical driving situations and near-accidents. Many problems associated with crash data can be overcome by identifying 'crash potential' of road sections, and conventional methods can be augmented by insights provided from vehicle data (Guo, Klauer, McGill, et al., 2010; Klauer et al., 2006). Moreover, estimations of traffic frequency measurements, such as ADT, can be generated from the frequency and trip details of vehicles travelling through these locations, transmitted via modern telematics, and used to estimate accident exposure measures.

2.2. *Insights from vehicle data*

Data which can be extracted from the vehicle itself to measure driving activities have been considered for various purposes by researchers over the years. Since access to standardised On-Board Diagnostics (OBD-II) data is mandatory for all vehicles manufactured or sold in the USA from 1996, many of these parameters have been widely used in research. This data gives insights into features such as vehicle speed, and has been used to detect hazardous driving behaviour (Imkamon et al., 2008). For example, one study predicted passenger ratings of driving behaviour by assigning current hazardousness to a range of values between driving safely and driving dangerously (Castignani et al., 2015). The authors developed classification models of these levels by combining OBD-II sensor data, 3-axis accelerometer data and Fuzzy Logic, a technique suitable when a binary distinction is not appropriate. In other work, the behaviour of different driver groups in the US State of Georgia was studied by collecting data from vehicle OBD-II systems as well as GPS location data (Jun et al., 2007). It was noted that clusters of hard deceleration events were co-located with clusters of historical accident data. Moreover, deeper insights into a vehicle's operation is available on the CAN Bus of a variety of vehicles. The CAN Bus provides access to specific unstandardized data from sensors in the vehicle, and as such has been used in recent research studies of driver behaviour. For example, characteristics of aggressive and calm driving have been identified with access to CAN Bus data (Karaduman et al., 2013).

In the last years, several researchers have built upon the idea that driving behaviour displayed in avoiding a crash is similar to that of a crash event itself, and hence, if detected, can be considered in the identification of crash potential. A naturalistic driving study of 100 cars equipped with camera, radar and OBD-II sensors was conducted in order to establish a surrogate safety metric for crashes and evaluated accident 'near-misses' (Dingus et al., 2006). In addition, a 'near-miss' has been defined operationally as "any circumstance that requires a rapid, evasive manoeuvre by the participant vehicle, or any other vehicle, pedestrian, cyclist, or animal, to avoid a crash. A rapid, evasive manoeuvre is defined as steering, braking, accelerating, or any combination of control inputs that approaches the limits of the vehicle capabilities" (Guo, Klauer, McGill, et al., 2010). Near-misses that occured in the study were determined to have a very strong correlation with actual crash events. It was found that the percentage of driver reaction is much lower for crashes than for near-crashes, suggesting that often the reaction of the driver is the deciding factor between a crash and a near-miss.

While the study by Guo and colleagues was specific with respect to defining a near-miss, in general, multiple factors have been identified as a 'near-miss' and critical driving event. As such, there are several methods of detecting CDEs that have been considered in recent literature. For example, various studies have explored the insights that can be gained through smartphone accelerometer data (Johnson and Trivedi, 2011). Unfortunately, this option comes with difficulties, such as drivers interacting with the phone during the journey, triggering high acceleration values and thus leading to false positive events. In very recent work, a small naturalistic driving study of university staff members, using only data gathered from a GPS logger, went beyond smartphone data and modelled accident data on US Highway 101 (Pande et al., 2017). In the study, a total of 39 segments were analysed using percentage of high jerk (rate of change of acceleration) events, slope, average daily traffic, and other explanatory variables in a negative binomial model. As a result, they found a promising relationship between jerk events and accidents which strongly motivates further research.

In summary, by extending existing work to detect high jerk events with a large enough fleet of vehicles and validating this approach with existing traffic accident analysis techniques, it becomes theoretically possible to identify crash potential and dangerous locations before an accident occurs. While traditional vehicles require retrofitting to deliver such insights, the growing number and popularity of semi-autonomous vehicles offer a unique opportunity to detect the locations of real-time CDEs. These events can act as a predictor for rarely occurring traffic accidents, and hence address issues of data scarcity and time delay (Guo, Klauer, Hankey, et al., 2010; Klauer et al., 2006). A prominent example of how this could be applied in practise addresses traffic queues forming on highways, where many lives are lost each year from drivers approaching these unforeseen queues too quickly (Li et al., 2013). By detecting such hazardous situations early enough, insurers can provide loss prevention services, such as triggering in-vehicle warnings to encourage drivers to approach with caution, and fully- or semi-autonomous vehicles can take risk reducing measures, such as adapting their speed gradually rather than abruptly.

3. Material and methods

3.1. Average daily traffic dataset

The amount of traffic flow, i.e. the number of cars travelling on a particular stretch of road, has long been associated with traffic accidents (Hauer and Persaud, 1984). The argument for this is that with a fixed probability of a traffic accident occurring, sections of road with higher traffic flow will see a higher number of traffic accidents per year. Therefore, in order to assess a specific location's Crash Rate or risk exposure based on the number of traffic accidents, it is important to account for traffic frequency. In order to incorporate this into the analysis, traffic data was obtained from the Swiss Road Authority (FEDRO). This dataset is comprised of the average number of vehicles per day passing a variety of counting stations across the Swiss road network. However, as is the case in many countries, in Switzerland there is only partial location coverage of ADT counting stations. This data was filtered to cover the same 18-week time period as the field study data. As such, the final set of observations were constructed from locations where measurements were available for this period, resulting in 194 counting stations and ADT measurements. The distribution of the ADT counts at these locations is shown in Fig. 1.

3.2. Traffic accident dataset

Traffic accident data for six years was additionally obtained from FEDRO in order to generate the dependent variables for our analysis. This dataset contained GPS locations, as well as contextual information on the causes, on over 268,000 traffic accidents which occurred in Switzerland between 2011 and 2016. Since accident locations can change over time due to improved road infrastructure, time-relevant Crash Rate and Crash Frequency dependent variables were generated by first sampling only the traffic accidents which occurred during the 18-week field study from the overall dataset, a total of 25,493 accidents across the country. Secondly, a naive grid-count approach was applied to count the traffic accidents which occurred within a 1 km^2 grid of each of the 194 traffic counting stations from Section 3.1, where the centre of the grid was each traffic frequency counting station, similar to a previous traffic accident study (Yang and Kim, 2003). Over the 18-week period, this came to a total of 972 traffic accidents, with an arithmetic mean of 5.01 and a geometric mean of 3.29 accidents per counting station location. Fig. 2 shows the distribution of traffic accidents within the bounding grid of the counting stations.

3.3. Field study dataset

72 professional drivers were recruited for an 18-week field study. Each of the drivers worked for the same roadside assistance company across a variety of locations, and all drove Chevrolet Captiva's of similar make and model. During the 18-week period drivers drove for approximately four hours, with an average of 144 km travelled per driver per day, resulting in over 690,000 km driven using the system in total. All drivers were male, with the exception of one participant, with a mean of 40.3 and a median of

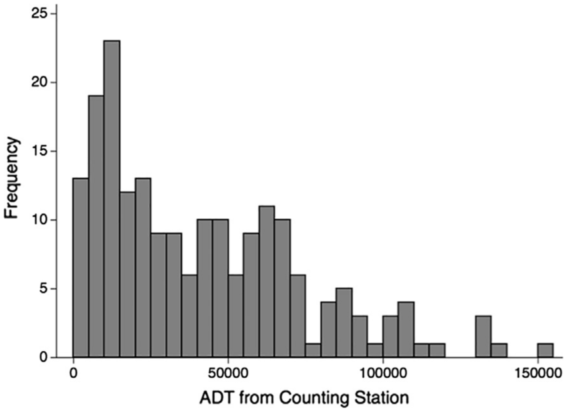

Fig. 1. Histogram showing the distribution of ADT from counting stations during the 18-week field study, with a bin size of 5000.

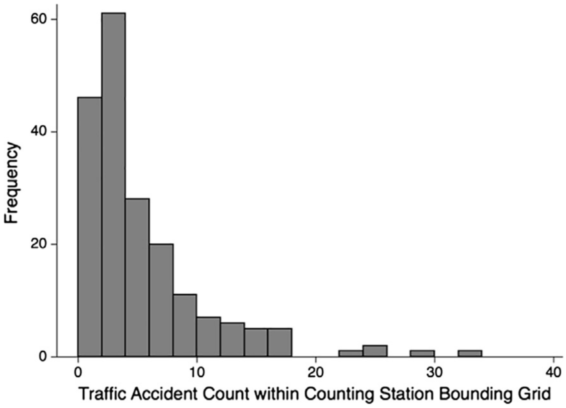

Fig. 2. Histogram showing the distribution of traffic accidents within counting station bounding grids during the 18-week field study, with a bin size of 2.

39 years of age. Vehicle data, representative of that available in semi-autonomous vehicles, was collected by accessing the CAN Bus of the Chevrolet Captiva's involved in the study via an OBD-II dongle. This dongle was paired via Bluetooth with a smartphone in the vehicle. The smartphone then transmitted various CAN Bus signals in real-time to a server, augmented with the GPS location and timestamp from the phone itself. Fig. 3 shows the routes in Switzerland which were travelled by the participants during the study. Since the location data acquisition relies on the GPS position of the smartphone, in areas where there was no GPS signal, such as in tunnels, no location data was captured. Therefore, these locations are not included in this analysis.

Following a recent study, high-jerk events were considered for classifying near-miss and CDEs (Pande et al., 2017). For the purpose of this analysis the jerk event behaviour was calculated from the vehicle's longitudinal acceleration sensor, which was sampled at 10 Hz. Post-processing of the acceleration values enabled the generation of jerk events for each participant of the field study. These values were calculated for each measurement time step based on the formula in Eq. (1), where *da* is the change acceleration (m/s^2), and *dt* is the change in time (s).

$$j = \frac{da}{dt} \tag{1}$$

The thresholds suggested for identifying near-miss jerk events significantly vary between research projects (Aichinger et al., 2016). These thresholds range from very strong events between -9.9 m/s^3 and -12.6 m/s^3 (Bagdadi and Várhelyi, 2011), which occur very infrequently in our fleet of professional drivers, with less than 600 events in the 18-week period, to very low events between -0.15 m/s^3 and -0.61 m/s^3 (Pande et al. 2017), which were triggered for almost every braking event that occurred in our study. As such, a threshold value of -2 m/s^3 was chosen to classify an event as high jerk. These jerk events were then limited to situations where the vehicles were decelerating, resulting in a dataset of roughly 912,000 geo-located jerk events over the course of

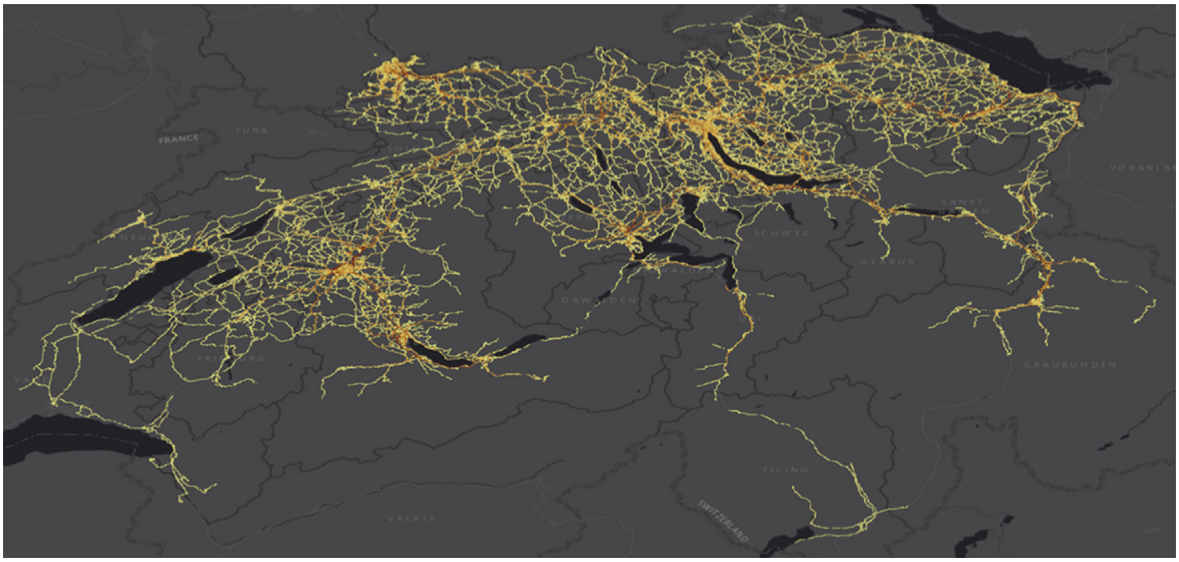

Fig. 3. Locations in Switzerland where naturalistic field study driving data was collected over the 18-week period.

the field study. Therefore, deceleration jerk events were calculated using the pseudocode shown in Algorithm 1, where successive jerk events were counted as one single event.

Algorithm 1. Pseudocode for deceleration jerk event generation.

```
DecelerationJerkEventGeneration (A, J)
  set successive_jerk_event = false
  set jerk_event_list = []
  for each value A' in acceleration values list A:
    if (A' < $0\,m/s^2$):
      // vehicle is decelerating
      J' = jerk values list J[position of A' in A]
      if (J' < $-2\,m/s^3$):
        // vehicle jerk rate is under threshold
        if (successive_jerk_event == false):
          add to jerk_event_list
          set successive_jerk_event = true
      else if (J' > $0\,m/s^3$):
        set successive_jerk_event = false
    else:
      // vehicle is accelerating
      set successive_jerk_event = false
  return jerk_event_list
```

3.4. Crash frequency, crash rate and traffic frequency

The first two datasets, the daily traffic dataset and the traffic accident dataset outlined in Sections 3.1 and 3.2 respectively, are provided by the Swiss Road Authority FEDRO. On the basis of these datasets, the following definitions are put forward:

- Let CF_{Pit} (Crash Frequency) be the number of accidents within location *i* from population *P* in timeframe *t*
- Let TF_{Pit} (Traffic Frequency) be the number of transits of location *i* of population *P* in timeframe *t* – in our case the average ADT over the period of the field study multiplied by the length of the study

A standardized measure of roadway safety and long-standing alternative to analysing Crash Frequency is to consider the exposure risk of locations, such as Crash Rate. In research and practise, Crash Rate is often reported and analysed in number of accidents per 1-million or even 100-million vehicle miles travelled. There are two fundamental reasons for the context-specific scaling of Crash Rates. Firstly, very small decimal numbers, as well as very high numbers, are hard to communicate. Secondly, the scaling of Crash Rate enables the application of well-proven techniques, such as established count regressions. Based on the scaling of a previous study (Anastasopoulos et al., 2012) we define Crash Rate in Eq. (2), where CR_{Pit} is the number of accidents per 100-million vehicle transits of population *P* in timeframe *t* for location *i*.

$$CR_{Pit} = \frac{CF_{Pit}}{TF_{Pit}} * 10^8 \tag{2}$$

The third dataset is the connected vehicle fleet dataset from the field study. In accordance to the definitions above we put forward the following definitions on the basis of fleet data:

- Let JF_{Sit} (Jerk Frequency) be the number of high jerk events within location *i* from sample (fleet) *S* in timeframe *t*
- Let TF_{Sit} (Trip Frequency) be the number of transits of location *i* from sample (fleet) *S* in timeframe *t* – in our case the number of trips through that location during the period of the field study

Finally, in Eq. (3) we define Jerk Rate (JR_{Sit}) as the number of high jerk events per vehicle transit of sample (fleet) *S* in timeframe *t* for location *i*.

$$JR_{Sit} = \frac{JF_{Sit}}{TF_{Sit}} \tag{3}$$

4. Results

In the following section we present the results of four sequential sets of analysis, with the underlying theme of investigating the link between CDEs and traffic accidents. In the analyses of Sections 4.1–4.3 we primarily consider locations where ADT measurements from FEDRO are available. In Section 4.4 we extend our approach to the majority of the Swiss road network, where the population traffic frequency measurements are unavailable. As such, we proceed by addressing the following problems:

- Section 4.1 examines an optimal scenario for determining accident exposure on the basis of driving data, in that we consider the subset of locations with available ADT measurements. This enables an investigation into the relationship between the Crash Rate of the population and the Jerk Rate of vehicles at these sites, and further motivates the subsequent sections.
- Section 4.2 follows by investigating the assumption that the Trip Frequency of the field study fleet is representative of the overall population traffic frequency, i.e. ADT, through these locations. By satisfying this, we can start to model the more practical situation where traffic frequency data is unavailable.
- Section 4.3 then continues by utilising the assumption from Section 4.2 and testing of the relationship between the Crash Frequency of the population, and the Jerk Rate and Trip Frequency of the field study fleet at these locations.
- Section 4.4 concludes our analyses by demonstrating the same relationship from Section 4.3 through nationwide spatial regressions covering the majority of the Swiss road network. We additionally explore the robustness of this model by iteratively including explanatory variables well-known to be associated with impacting the likelihood of traffic accidents occurring.

4.1. Crash rate and jerk rate

In a first step, we focus our analysis on an ideal setting, where Crash Frequency as well as traffic frequency are available for all locations. More specifically, locations were defined following a naive grid-count approach, i.e. traffic accidents and traffic volumes were determined within a 1 km^2 grid of each of the 194 traffic counting stations. Since our dependent variable in the initial analysis is $\mathrm{CR_{Pit}}$, i.e. the expected number of traffic accidents of the grid per 100-million transits, we are effectively analysing a count variable as a measure of exposure. We see large differences between the mean (183.3) and the variance (123,340.2) of the variable, indicating over-dispersion. In addition, when considering the $\mathrm{JR_{Sit}}$ independent variable, we see that it also follows a Poisson distribution. As such, we transform this via the commonly used natural log function to normalise the variable (Quddus 2008), and utilise negative binomial regression, thus assuming the relationship defined in Eq. (4) between $\mathrm{CR_{Pit}}$ and $\mathrm{JR_{Sit}}$.

$$CR_{Pit} = e^{a1} * (JR_{Sit})^{b1} \tag{4}$$

Regressions were run for the whole set of counting stations (Model 1.1), and additional models generated where counting stations were excluded based on a minimum value of $\mathrm{TF_{Sit}}$. As shown in Table 1, we see that the correlation between Crash Rate and Jerk Rate is highly significant in all models and has a similar coefficient, ranging between 0.293 and 0.383. In addition, the over-dispersion constant, i.e. ln alpha, is significant for all models, confirming the suitability of negative binomial regression. The best fitting model is where counting stations were excluded with 20 or less trips ($\mathrm{TF_{Sit}}$) through the surrounding bounding grid, with a pseudo R^2 of 0.137. As one might expect, this indicates that there is a minimum number of times an area should be driven through before an assessment should be conducted on the likelihood of traffic accidents occurring based on driving data. However, the model fit drops when limiting with higher number of trips, most likely due to decreased observations in the model.

In order to answer our initial research question: To what extent can the Crash Rate of a location be predicted by CDE information from that location, assuming that Crash Frequency and ADT data are available, we test the null hypothesis that $\mathrm{CR_{Pit}}$ of a location is not explained by $\ln(\mathrm{JR_{Sit}})$. Through testing the χ^2 values for each of the generated models, we are led to reject this null hypothesis, with $p < 0.001$ for all models. Since negative binomial regression utilises the log-link, and $\mathrm{JR_{Sit}}$ is natural log transformed, the interpretation of the best-fitting model allows us to conclude that with a 10% increase in the Jerk Rate we would expect a 3.7%

Table 1
Negative binomial regression results for the Crash Rate ($\mathrm{CR_{Pit}}$) dependent variable and the independent variable Jerk Rate ($\mathrm{JR_{Sit}}$), sampled by number of trips through counting station bounding grid ($\mathrm{TF_{Sit}}$).

	Negative binomial regression models, sampled by minimum $\mathrm{TF_{Sit}}$ Dependent Variable: $\mathrm{CR_{Pit}}$					
Independent variables	Model (1.1) $\mathrm{TF_{Sit}} > 0$	Model (1.2) $\mathrm{TF_{Sit}} > 5$	Model (1.3) $\mathrm{TF_{Sit}} > 10$	Model (1.4) $\mathrm{TF_{Sit}} > 15$	Model (1.5) $\mathrm{TF_{Sit}} > 20$	Model (1.6) $\mathrm{TF_{Sit}} > 25$
ln ($\mathrm{JR_{Sit}}$)	0.293^{***} (3.86)	0.338^{***} (4.43)	0.354^{***} (4.56)	0.362^{***} (4.40)	0.385^{***} (4.74)	0.383^{***} (4.66)
Constant	20.03^{***} (273.63)	20.09^{***} (266.12)	20.12^{***} (260.01)	20.14^{***} (251.74)	20.16^{***} (252.18)	20.16^{***} (252.45)
ln alpha						
Constant	-0.326^{***} (−4.22)	-0.352^{***} (−4.30)	-0.377^{***} (−4.39)	-0.375^{***} (−4.29)	-0.393^{***} (−4.43)	-0.396^{***} (−4.44)
Observations	194	172	159	152	149	147
Pseudo R^2	0.079	0.108	0.121	0.120	0.137	0.133
χ^2	14.90	19.66	20.82	19.36	22.48	21.71

t statistics in parentheses.
$^{*}p < 0.05$.
$^{**}p < 0.01$.
$^{***}p < 0.001$.

($1.10^{0.385} = 1.037$) increase in the Crash Rate within the surrounding area.

However, for practitioners and researchers alike, while $\mathrm{CF_{Pit}}$ can often be obtained, such ideal data coverage for $\mathrm{TF_{Pit}}$, e.g. ADT, is rarely available to generate $\mathrm{CR_{Pit}}$. It is therefore important to consider more common situations in traffic analysis where we can only leverage Crash Frequency as a dependent variable for the whole road network. Therefore, while Crash Rates have been used in the first part of this analysis, the remainder of the paper focuses on Crash Frequency, additionally motivated by previous work investigating traffic accident exposure (Hauer, 1995).

4.2. Representative fleet assumption

In the second stage of our analysis we build upon the following assumption to determine Crash Frequency: That there has to be a well-defined relationship between traffic frequency of the population ($\mathrm{TF_{Pit}}$) and traffic frequency of the fleet sample ($\mathrm{TF_{Sit}}$), which is independent of location *i* for a specific timeframe *t*, i.e. a 'representative fleet'. A roadside assistance fleet, as employed in our study, could be considered representative, since they are called to locations where typical drivers have broken down or require assistance. Thus, in order to test our second research question regarding the relationship between fleet-generated events and the Crash Frequency variable, we must first test our representative assumption on the fleet dataset at hand. We can test this assumption by considering the relationship between $\mathrm{TF_{Pit}}$ and $\mathrm{TF_{Sit}}$. Since $\mathrm{TF_{Pit}}$ for the set of stations is generated by multiplying the ADT of the location by the length of the field study, we are estimating a count variable which follows a Poisson distribution. Additionally, by considering the mean (41,964) and variance (1.10e + 09) of $\mathrm{TF_{Pit}}$, we see that the variance is much larger, indicating that the variable is over-dispersed and that negative binomial regression is a suitable technique for analysis.

When considering $\mathrm{TF_{Sit}}$, the traffic frequency of our sample of fleet drivers, we also observe a Poisson distribution of the data, and thus apply the commonly used normalisation technique of natural logarithm transformation to the variable (Quddus, 2008). As in the previous section, we additionally assume that there is a trade-off with our sample, where there is a minimum number of times an area should be driven through for a reliable assessment on the basis of driving data. As such, we expect that removing grids with less trips from the fleet of drivers ($\mathrm{TF_{Sit}}$) will increase the model fit, until a certain point where too many data points have been removed. The results of these regressions are shown in Table 2.

We see that the natural logarithm of $\mathrm{TF_{Sit}}$ has a highly significant coefficient with $\mathrm{TF_{Pit}}$. Through testing the null hypothesis that $\mathrm{TF_{Pit}}$ of a location is not explained by $\ln(\mathrm{TF_{Sit}})$ of our field study fleet, we are able to determine whether overall population traffic volume can be explained by fleet traffic volume on the basis of the assumed negative binomial relationship between $\mathrm{TF_{Pit}}$ and $\mathrm{TF_{Sit}}$ as shown in Eq. (5).

$$TF_{Pit} = e^{a2} * (TF_{Sit})^{b2} \tag{5}$$

Through testing the χ^2 values for each of the generated models, we are led to reject this null hypothesis, with $p < 0.001$ for all models. The best fitting model was where counting stations were excluded with 15 or less trips through the surrounding bounding grid and with a pseudo R^2 of 0.327. Finally, Fig. 4 shows the distribution of the natural logarithm transformed $\mathrm{TF_{Sit}}$ independent variable, and Fig. 5 the relationship between $\mathrm{TF_{Pit}}$ and $\mathrm{TF_{Sit}}$. Both figures feature the full dataset, and a cut-off demonstrating the data used in the negative binomial regression model with the best fit, i.e. only including grids with greater than 15 trips.

Table 2
Negative binomial regression results for the Population Traffic Frequency ($\mathrm{TF_{Pit}}$) dependent variable and the independent variable Fleet Traffic Frequency ($\mathrm{TF_{Sit}}$) sampled by number of trips through counting station bounding grid ($\mathrm{TF_{Sit}}$).

	Negative binomial regression models, sampled by minimum $\mathrm{TF_{Sit}}$ Dependent Variable: $\mathrm{TF_{Pit}}$					
Independent variables	Model (2.1) $\mathrm{TF_{Sit}} > 0$	Model (2.2) $\mathrm{TF_{Sit}} > 5$	Model (2.3) $\mathrm{TF_{Sit}} > 10$	Model (2.4) $\mathrm{TF_{Sit}} > 15$	Model (2.5) $\mathrm{TF_{Sit}} > 20$	Model (2.6) $\mathrm{TF_{Sit}} > 25$
ln ($\mathrm{TF_{Sit}}$)	0.203^{***} (5.78)	0.269^{***} (5.30)	0.342^{***} (7.19)	0.425^{***} (10.14)	0.424^{***} (9.28)	0.428^{***} (9.04)
Constant	14.49^{***} (74.18)	14.13^{***} (49.44)	13.72^{***} (51.08)	13.24^{***} (56.77)	13.24^{***} (51.51)	13.22^{***} (49.50)
ln alpha						
Constant	-0.482^{***} (−4.97)	-0.642^{***} (−5.30)	-0.740^{***} (−6.25)	-0.808^{***} (−6.46)	-0.792^{***} (−6.33)	-0.792^{***} (−6.25)
Observations	194	172	159	152	149	147
Pseudo R^2	0.204	0.241	0.279	0.327	0.309	0.302
χ^2	33.43	28.12	51.77	102.8	86.11	81.72

t statistics in parentheses.
$^{*}p < 0.05$.
$^{**}p < 0.01$.
$^{***}p < 0.001$.

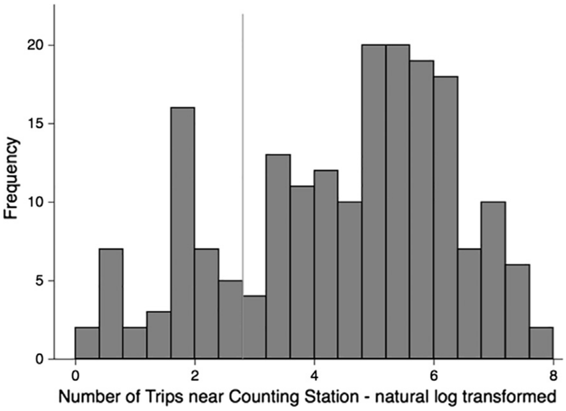

Fig. 4. Histogram showing log-normal distribution of field study fleet trips through counting station bounding grids, with a bin size of 0.4. Vertical line indicates 15 trips cut off.

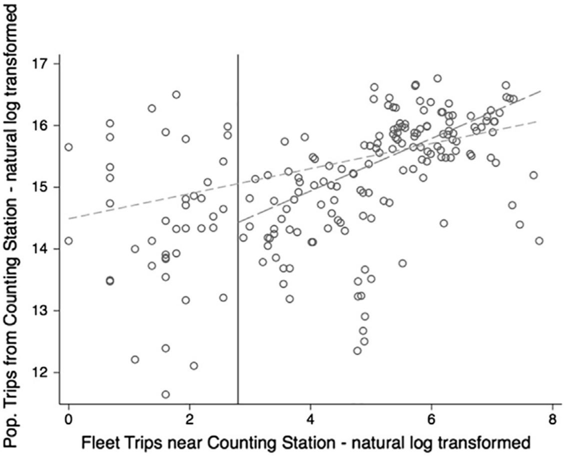

Fig. 5. Scatter plot showing relationship between TF_{Pit} and TF_{Sit} Dashed line is Model (2.1). Long dashed line is Model (2.4). Vertical line indicates 15 trips cut off.

4.3. Crash frequency and jerk frequency

Building upon the evidence that we have a well-defined relationship between TF_{Sit} and TF_{Pit}, we can estimate the Crash Frequency of the locations based on the Jerk Rate and Trip Frequency of our sample, substituting TF_{Pit} with TF_{Sit} and thereby eliminating the need for ADT information. More specifically, based on the previous definitions and assumptions we can conclude:

(from Equation 2) $CR_{Pit} = \frac{CF_{Pit}}{TF_{Pit}} * 10^8$

(from Equation 4) $CR_{Pit} = e^{a1} * (JR_{Sit})^{b1}$

(from Equation 5) $TF_{Pit} = e^{a2} * (TF_{Sit})^{b2}$

$$\frac{CF_{Pit}}{TF_{Pit}} * 10^8 = e^{a1} * (JR_{Sit})^{b1} \tag{6}$$

$$CF_{Pit} = e^{a1-\ln(10^8)} * (JR_{Sit})^{b1} * TF_{Pit} \tag{7}$$

$$CF_{Pit} = e^{a1-\ln(10^8)} * (JR_{Sit})^{b1} * e^{a2} * (TF_{Sit})^{b2} \tag{8}$$

$$CF_{Pit} = e^{a1+a2-\ln(10^8)} * (JR_{Sit})^{b1} * (TF_{Sit})^{b2} \tag{9}$$

Thus, through a Poisson or negative binomial regression with Crash Frequency CF_{Pit} as the dependent variable, we can derive the coefficients *b1* as well as *b2*. If *a2* is available from a previous analysis determining the relationship between TF_{Pit} with TF_{Sit}, i.e. through the analysis in Section 4.2, we can also determine *a1*. With estimates for *a1* and *b1* we can ultimately also predict CR_{Pit} on the

Table 3
Negative binomial regression results for the Crash Frequency (CF_{Pit}) dependent variable and the independent variables Jerk Rate (JR_{Sit}) and Fleet Trip Frequency (TF_{Sit}), sampled by number of trips through counting station bounding grid (TF_{Sit}).

	Negative binomial regression models, sampled by minimum TF_{Sit} Dependent Variable: CF_{Pit}					
Independent variables	Model (3.1)	Model (3.2)	Model (3.3)	Model (3.4)	Model (3.5)	Model (3.6)
	$TF_{Sit} > 0$	$TF_{Sit} > 5$	$TF_{Sit} > 10$	$TF_{Sit} > 15$	$TF_{Sit} > 20$	$TF_{Sit} > 25$
ln (JR_{Sit})	0.330^{***}	0.367^{***}	0.405^{***}	0.371^{**}	0.380^{***}	0.384^{***}
	(4.88)	(5.18)	(5.31)	(4.92)	(5.02)	(4.93)
ln (TF_{Sit})	0.232^{***}	0.259^{***}	0.314^{***}	0.378^{***}	0.367^{***}	0.363^{***}
	(6.43)	(5.20)	(6.15)	(7.95)	(7.54)	(7.30)
Constant	0.479^{**}	0.333	0.0171	−0.366	−0.302	−0.276
	(2.61)	(1.25)	(0.06)	(−1.49)	(−1.19)	(−1.06)
ln alpha						
Constant	-0.929^{***}	-0.905^{***}	-0.960^{***}	-1.020^{***}	-1.012^{***}	-1.004^{***}
	(−6.38)	(−5.83)	(−5.84)	(−5.82)	(−5.74)	(−5.68)
Observations	194	172	159	152	149	147
Pseudo R^2	0.281	0.275	0.303	0.336	0.331	0.322
χ^2	56.77	49.49	62.09	73.27	69.24	64.32

t statistics in parentheses.
$^{*}p < 0.05$.
$^{**}p < 0.01$.
$^{***}p < 0.001$.

basis of $CR_{Pit} = e^{a1} * (JR_{Sit})^{b1}$.

Considering the CF_{Sit} dependent variable, we continue to see a Poisson distribution where the variance (28.20) is higher than the mean (5.01), indicating over-dispersion of the variable and thus we apply negative binomial regression. Results of the regressions, sampled by number of trips (TF_{Sit}) are shown in Table 3. When compared to the CR_{Pit} set of models, we see improved model fit, with pseudo R^2 ranging from 0.281 to 0.336. In addition, we observe the same pattern seen in the previous regressions of Section 4.2, where model fit improved when grids with less than 15 trips were excluded from the analysis. This reaffirms that there should be a minimum number of fleet transits of an area before analyses are conducted on the basis of driving data.

We can now proceed to answer our second research question: To what extent can the Crash Frequency of a location be explained by CDE information from that location, assuming we have sparse data coverage (only Crash Frequency information is available) and a well-defined relationship between TF_{Sit} and TF_{Pit}, i.e. a 'representative' fleet. This is achieved by separately testing two null hypotheses, i.e. that CF_{Pit} of a location is not explained by either ln(JR_{Sit}) or by ln(TF_{Sit}) of our field study fleet. Through independently testing the χ^2 values for each variable of the generated models, we are led to reject both null hypotheses, with $p < 0.001$ for both variables and all models. Regarding the *b1* coefficient, which gives the proportional increase in both Crash Rate, and Crash Frequency in our models, we observe that in the CF_{Pit} regressions it falls within a similar range (0.330–0.405) to the coefficient in the CR_{Pit} set of models (0.293–0.383). In addition, we observe that the *b2* coefficient also falls in a similar range (0.232–0.378) to the coefficient in the TF_{Pit} models (0.203–0.428). From this we can conclude that, with a fleet which can be demonstrated to be representative of the population, we can estimate Crash Frequency, and from the coefficient make an approximation for the Crash Rate exposure measure.

4.4. Crash frequency – spatial regression

In practise, such an analysis would not be run on a small subset of grids, but over the whole road network. In order to provide insights into a nationwide approach, we now consider this problem for the whole field study dataset, covering the majority of the Swiss road network. Following the same approach as the first part of the analysis, the country was divided into regular square grid cells (Kim et al., 2006) of 1 km^2 in order to perform the data analysis. Crash Frequency was generated within each grid, along with the Jerk Frequency, and Traffic Frequency of the fleet, to generate new values for the variables CF_{Pit}, JR_{Sit}, and TF_{Sit}. Based on the results of the previous sections, grids with a value for TF_{Sit} less than 15 were excluded from the analysis, resulting in 4197 km^2 of the country's road network being considered. Of the 25,493 traffic accidents which occurred during the 18-week period, 15,450 were included in the grids covered by the field study and fulfilled the minimum number of transits by the utilised fleet.

Urban areas, as well as highways, often show very specific accident patterns (Kim et al., 2006; Pande et al., 2017). These variables can be obtained for locations through map-matching services and open data sources. Therefore, in order to include these as explanatory variables in the regression, each grid was further enriched with two binary variables, one for 'Urban' or 'Rural' and another for 'Highway' or 'Non-highway' through a professional map-matching service provider. The data set is split roughly in half between urban and rural areas, as shown in Fig. 6, while Fig. 7 shows the distribution of highway and non-highway roads. As mentioned above, significant differences are expected between the areas and road-types due to the difference in both population and traffic frequency.

Spatial autocorrelation is where measurements are dependent on their location and surroundings, i.e. not independent and

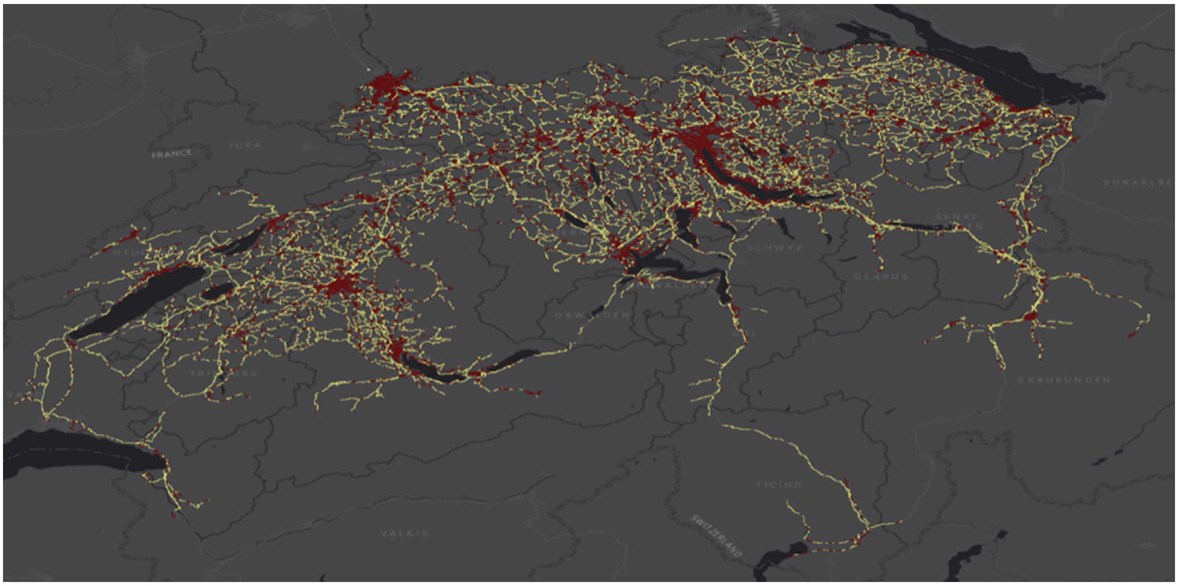

Fig. 6. Distribution of Urban (red) and Rural (yellow) locations in Switzerland where naturalistic field study driving data was collected. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

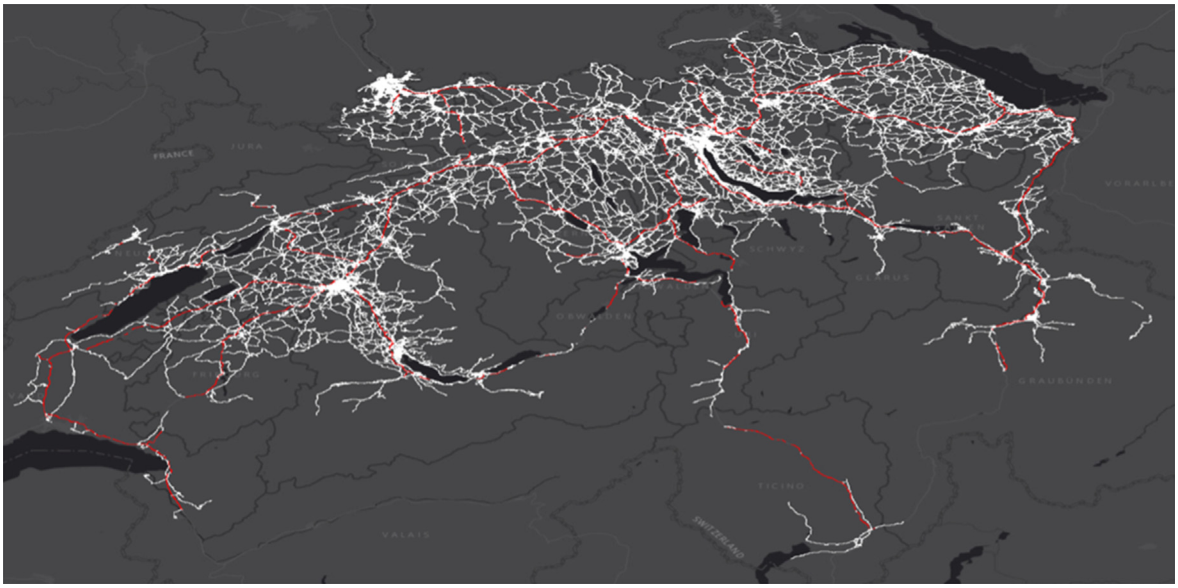

Fig. 7. Distribution of highway (red) and other (white) locations in Switzerland where naturalistic field study driving data was collected. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

identically distributed. Thus, it can be defined as a property found across geographic space, where variables are either more similar or less similar at certain distances from each other than would be expected for randomly associated pairs of observations (Legendre, 1993). This spatial autocorrelation can be tested with various indicators, where the most commonly used measure is the Moran's I statistic (Anselin and Rey, 2014; Bivand et al., 2013). Moran's I can be interpreted as a regression coefficient for autocorrelation, where the value falls between -1 and 1. A value of 1 indicates a strong positive spatial autocorrelation, -1 indicates a strong negative spatial autocorrelation and 0 means that observations are random, and thus not spatially auto-correlated (Anselin, 1996). For our dataset and a grid size of 1 km^2, the Moran's I value is 0.426 ($p = 0.000$), hence we find a statistically significant spatial autocorrelation which should be accounted for in our model.

Another important concept for statistical tests is homo- and heteroscedasticity. Homoscedasticity is the assumption of constant error variance (Anselin and Rey, 2014). That is, the variances around the regression line are equal for all values of the independent variable. If the distribution of values is heteroskedastic, the variance of the independent variable cannot be assumed to be constant. Due to the spatial distribution of accidents we expect heteroscedasticity of accidents and vehicle data along different roads sections.

Table 4

Spatial combo regression results for the Crash Frequency (CF_{Pit}) dependent variable and iteratively added independent variables, with grids limited to those with a fleet Trip Frequency greater than 15 (TF_{Sit}).

	Spatial combo regression models dependent variable: ln (CF_{Pit})			
Independent variables	Initial model (4.1)	Urban model (4.2)	Highway model (4.3)	Full model (4.4)
ln (JR_{Sit})	0.3176^{***}	0.2948^{***}	0.3224^{***}	0.3043^{***}
	(35.7135)	(39.8947)	(34.1915)	(31.2257)
ln (TF_{Sit})	0.1060^{***}	0.1304^{***}	0.0857^{***}	0.0766^{***}
	(7.6174)	(12.2520)	(5.1154)	(4.6636)
Urban		0.1417^{***}		0.1651^{***}
		(6.8906)		(7.0837)
Highway			0.0292	0.1250^{***}
			(4.9491)	(3.6352)
Constant	-0.8754^{***}	-0.9776^{***}	-0.8694^{***}	-0.9376^{***}
	(-13.0789)	(-15.4511)	(-14.0448)	(-17.3417)
Spatial lag	0.0618	0.0840^{*}	0.1183^{***}	0.2022^{***}
	(1.4026)	(6.8906)	(3.0585)	(6.0833)
Spatial error	0.6330^{***}	0.5994^{***}	0.5559^{***}	0.4296^{***}
	(25.9786)	(24.0536)	(21.4148)	(14.6869)
Observations	4197	4197	4197	4197
Pseudo R^2	0.3899	0.4017	0.3913	0.4083

z statistics in parentheses.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

To address the problem of heteroscedasticity, spatial regression models can include local and global spatial relationships in order to mitigate or reduce the effect. To utilise these spatial effects, spatial weights are used (Anselin and Rey, 2014; Bivand et al., 2013). For our grid-based data set, contiguity weights are the most logical choice, where only bordering cells are expected to have a direct effect on the dependent variable. Other techniques, such as threshold and inverse distance weights, are impractical as they use different distances for diagonal than horizontally or vertically adjacent cells due to the calculation using the centre point of the square. Contiguity weights also allow higher order weights for cells that are further away than the cell directly adjacent.

Lagrange Multiplier tests showed that our dataset is susceptible to both spatial lag and spatial error, where tests are highly significant for both lag and error models for the general and the robust specification. Therefore, the spatial combo model is used as it includes both spatial error and spatial lag. The dependent variable, CF_{Pit}, is transformed with the natural logarithm, to replicate the log-link relationship from the negative binomial regressions. For this analysis we iteratively developed four models, presented in Table 4, in order to test the robustness of the relationships when controlling for factors that traditionally explain large parts of the variance in Crash Frequency, specifically Urban and Highway environments. The first of these, model (4.1), uses just the natural logarithm transformed JR_{Sit} and TF_{Sit} as dependent variables, extending our results from the previous sections to a country-wide setting and accounting for spatial autocorrelation. In model (4.2) we add to the regression the Urban binary variable, and in model (4.3) the binary variable for Highway. We finally add both binary variables to our combined model (4.4).

Considering the initial model (4.1), as seen in the negative binomial regressions from the previous sections, we observe highly significant coefficients for both ln(JR_{Sit}) and ln(TF_{Sit}). Here the coefficient of ln(JR_{Sit}) fall within a similar range as the results from the previous section, and so with an increase of 10% in Jerk Rate from our fleet at a location we would expect an increase of 3.1% ($1.10^{0.3176} = 1.031$) in the Crash Frequency at that location. On the other hand, while the number of trips the fleet makes through the location also shows a significant proportional increase in the crash frequency, with a 10% increase in trips contributing a 1.0% ($1.10^{0.1060} = 1.010$) increase in the number of accidents, the coefficient is lower than in the previous models. Here we potentially see the importance of incorporating spatial factors in such an analysis, since the spatial error is highly significant in all of the models there is an indication that the error of an observation affects the errors of its neighbours. As the individual trips of the fleet naturally pass through neighbouring grids, this behaviour can potentially be accounted for in the spatial model.

Model (4.2) sees the added significance of the Urban binary variable added into the regression, and the spatial lag additionally becomes significant. Since spatial lag is the variable which captures the influence of neighbouring observations on the dependent variable, we can determine that controlling for Urban and Rural locations highlights the similarity of neighbouring observations in the dataset. Model (4.3) adds the binary variable for whether the location is a Highway, here the variable itself does not have a significant effect on the number of crashes within that grid. However, we see that the coefficient and the significance of the spatial lag increase, indicating higher importance of the influence of neighbouring observations on the number of crashes.

Finally, the combined and best fitting model (4.4), with a pseudo R^2 of 0.4083, shows the significance of both the binary variables of Urban and Highway, and the spatial lag and error for these models remain significant. With regard to the impact of the binary variables, we primarily consider the results of this model. In including these variables, we see minor changes in the coefficients of ln(JR_{Sit}), from 0.3176 to 0.3043, and ln(TF_{Sit}), from 0.1060 to 0.0766, with the significance of both variables remaining. The binary variable coefficients in the model indicate the direct impact on the dependent variable of being in urban and highway areas respectively. Where the impact of being in an urban location increases the number of accidents at that location by 16.5% when compared to rural areas. Moreover, locations which were highways had an increase of 12.5% in the Crash Frequency when compared to other non-highway roads.

5. Discussion, limitations and future work

The research at hand has several implications for research, policy, and industry. At the core of our research is the hypothesis that the rate of CDEs at a specific location is related to the Crash Rate and Crash Frequency of that location. Thus, we conducted an explorative nationwide study to provide early evidence that the Crash Rate can be modelled by the Jerk Rate of vehicles that are representative of a semi-autonomous fleet. Here, we observe the significance of coefficients that would be expected from the literature and see promising quality metrics with regard to pseudo R^2 and χ^2. In addition, model fit generally increased where grids were only considered for analysis when they had a higher number of transits by the utilised fleet. This demonstrates that with more trips the measure of Jerk Rate becomes more representative of the population Crash Rate. However, in reality, we are commonly limited in traffic safety analysis to estimating Crash Frequency as a dependent variable, since the necessary traffic frequency data needed to calculate exposure measures, such as Crash Rate, is rarely available for whole road networks.

In order to reliably estimate the Crash Frequency of a location on the basis of Jerk Rate (without the corresponding data for overall population traffic frequency), the relationship between the traffic frequency of the field study fleet and total traffic frequency of the population has to be known. The results indicate that a road assistance fleet, such as the one used in this field study, can address this requirement. More specifically, we provide evidence of a well-defined relationship between the number of trips made by the fleet and the overall traffic frequency measure, ADT. In practise, it may be easier for insurance companies to use a professional fleet to model both Crash Rate and Crash Frequency, rather than recruiting and collecting data from typical drivers within the general population. After validating the assumption that the fleet in our field study was representative of the population, we developed models to estimate Crash Frequency, and from the coefficient make an approximation for the Crash Rate exposure measure. Through these steps, we contribute to the existing body of research by going beyond the most recent studies that consider individual highway segments (Pande et al., 2017), applying our analysis to a subset of locations with ADT data, and providing model quality measures to help other researchers and practitioners validate their approaches. Furthermore, in practise such an analysis would not be run merely on a small subset of grids where traffic frequency information is available, but over the whole road network where only Crash Frequency is typically collected. Therefore, our study, which covers the majority of the Swiss road network, provides first insights into a nationwide approach. We highlight the importance of controlling for spatial lag and error effects, and draw attention to how urban and highway variables obtained from map-matching services impacted Crash Frequency at those locations and improved model fit. Finally, the models presented in each section form a concrete starting point for automotive insurance companies looking to exploit the growing opportunity for active loss prevention, such as providing in-vehicle warnings and safe-routing services. These services, which build upon an accurate knowledge of dangerous locations on the road network, will be further enabled by the growing popularity and advanced features of semi-autonomous vehicles in the coming decades.

Whilst this study demonstrates how traffic accident exposure can be modelled by naturalistic driving data, the results should be seen in the light of their limitations. While there are many jerk threshold values suggested in the related research, as well as other variables to detect CDEs, in this early stage we consider just one threshold of jerk in order to generate our independent variables. We believe that to gain greater insights into road safety there is a strong need to develop clear operationalisations of critical driving event variables, and reliably validate these for multiple vehicle and driver demographics, across different road types. In addition, spatial point data can be analysed in numerous ways, and it is important to mention that improvements in the analysis could be made by considering non-grid based spatial regression techniques. For example, map-matching traffic accidents, trips and CDEs onto a representation of the road network and dividing this into network segments could be a promising future endeavour (Pande et al., 2017). Moreover, including attributes for these road segments, e.g. road curvature, as seen in other studies, could provide deeper insights into the relationship between CDEs and Crash Frequency.

6. Conclusions

While there is existing research showing that the situational factors of crashes and near-misses are strongly related, and that naturalistic driving data can provide promising insights for traffic accident analysis, there is limited empirical data on whether the Crash Frequency and Crash Rate of locations can be reliably identified through analysis of CDEs. With the rise of semi-autonomous vehicles, and the technology capabilities facilitating their innovative features, a unique opportunity to address this issue has arisen. While low levels of adoption currently make investigating this potential a challenge, preliminary analysis is made possible with the installation of a retrofit system analogous to such advanced vehicles. Using such a setup, jerk events were detected from 72 vehicles equipped with a system that collected sensor information from each car's CAN Bus, synonymous with data available in semi-autonomous vehicles. The research at hand presents results from an 18-week naturalistic driving field study of these vehicles, which covered over 690,000 km in Switzerland. The contributions of this paper can be summarised as follows:

- To the best of our knowledge, this is the first work to incorporate a national level traffic accident and traffic volume dataset covering the same period as a nationwide naturalistic driving field study. We demonstrate with negative binomial regressions on a sub-sample of data where ADT was available that the rate of high jerk events of vehicles has a proportional relationship with the Crash Rate of the population.
- Since traffic frequency measures, such as ADT, are expensive to collect and rarely available over the whole road network, we

further demonstrate that the number of trips made by the connected fleet in the field study has a proportional relationship with the population traffic frequency. From this, Crash Frequency is modelled through negative binomial regressions with the rate of high jerk events and the number of trips made by the fleet.

– Finally, we apply spatial regression analysis on Crash Frequency across locations covering over 4000 km^2 of the Swiss national road network and demonstrate the relationship with the rate of high jerk events and fleet traffic frequency, along with urban and highway explanatory binary variables for these locations.

In conclusion, this early-stage research shows the potential for traffic safety analysis that could be achieved with the driving data collected from semi-autonomous vehicles. With the ongoing shift in the insurance industry that is moving away from purely 'reactive' business models toward active loss prevention, the ability to identify high risk locations on the road network is of utmost value. Through this, preventative measures can be incorporated into insurance products to encourage drivers to adapt their driving behaviour at potentially dangerous locations. Additionally, semi-autonomous vehicles approaching these locations might drive more carefully to reduce risk or hand over control to the driver to transfer insurance liability. We believe that understanding the new practical capabilities and the reliance of findings from recent automotive technology improvements has become vital for determining suitable traffic safety approaches and strategies, both for insurers and policy makers in this long-standing field of traffic accident prevention.

Acknowledgment

The authors would like to acknowledge the feedback, comments, and suggestions from two anonymous reviewers, whose input has greatly improved this paper. This work was supported by the Bosch IoT Lab at the University of St. Gallen, Switzerland. We would also like to thank our colleagues on our Connected Car project, Prof. Dr. Markus Weinberger, and Timo Gessmann.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.tra.2018.05.007.

References

Aichinger, C., Nitsche, P., Stütz, R., Harnisch, M., 2016. Using low-cost smartphone sensor data for locating crash risk spots in a road network. Transportation Res. Procedia 14, 2015–2024. http://dx.doi.org/10.1016/j.trpro.2016.05.169.

Albright, J., Bell, A., Schneider, J., Nyce, C., 2016. Automobile Insurance in the Era of Autonomous Vehicles, KPMG. (https://home.kpmg.com/content/dam/kpmg/pdf/2016/05/kpmg-automobile-insurance-in-era-autonomous.pdf, accessed August 28, 2017).

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159 (10.1016/j.aap.2008.10.005).

Anastasopoulos, P.C., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012. A study of factors affecting highway accident rates using the random-parameters Tobit model. Accid. Anal. Prev. 45, 628–633. http://dx.doi.org/10.1016/j.aap.2011.09.015.

Anselin, L., 1996. The moran scatterplot as an ESDA tool to assess local instability in spatial association. Spatial Anal. Perspect. GIS 111–125.

Anselin, L., Rey, S.J., 2014. Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL, GeoDa. Press LLC.

AXA. 2017. "AXA DriveSafe." (https://www.axa.ie/car-insurance/young-drivers-insurance/products/drivesave/, accessed August 28, 2017).

Bagdadi, O., Várhelyi, A., 2011. Jerky driving - an indicator of accident proneness? Accid. Anal. Prev. 43 (4), 1359–1363 (10.1016/j.aap.2011.02.009).

Bansal, P., Kockelman, K.M., 2017. Forecasting Americans' long-term adoption of connected and autonomous vehicle technologies. Transportation Res. Part A: Policy Practice 95, 49–63. http://dx.doi.org/10.1016/j.tra.2016.10.013.

Berger, R., 2015. Internet of Things and Insurance. (http://theinternetofthings.report/Resources/Whitepapers/783de5d8-cbe5-45f1-9170-824454460543_Roland_Berger_Internet_of_Things_and_insurance_20150513.pdf, Accessed August 28, 2017).

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. Anal. Methods Accident Res. 1, 53–71. http://dx.doi.org/10.1016/j.amar.2013.10.001.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., 2013. Applied spatial data analysis with R. Use R, Vol. 1. https://doi.org/10.1007/978-0-387-78171-6.

Castignani, G., Derrmann, T., Frank, R., Engel, T., 2015. Driver behavior profiling using smartphones: a low-cost platform for driver monitoring. IEEE Intell. Transp. Syst. Mag. 7 (1), 91–102 (10.1109/MITS.2014.2328673).

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prev. 37 (5), 870–881 (10.1016/j.aap.2005.04.015).

CSS Insurance, 2017. Health account and health account bonus. https://www.css.ch/en/home/privatpersonen/krankenkasse/zusatzversicherung/gesundheitskonto.html#gesundheitskonto, (Accessed August 28, 2017).

Deacon, J.A., Zegeer, C.V., Deen, R.C., 1974. Identification of hazardous rural highway locations. Transportation Res. Board.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knipling, R.R., 2006. The 100-Car Naturalistic Driving Study Phase II – Results of the 100-Car Field Experiment. Dot Hs 810 593 (April), No. HS-810 593. https://doi.org/DOT HS 810 593.

Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transp. Res. Part A: Policy Practice 77, 167–181. http://dx.doi.org/10.1016/j.tra.2015.04.003.

Gianfranco, F., Soddu, S., Fadda, P., 2017. An accident prediction model for urban road networks. J. Transportation Saf. Security 1–19.

Greibe, P., 2003. Accident prediction models for urban roads. Accident Anal. Prevention 273–285. http://dx.doi.org/10.1016/S0001-4575(02)00005-2.

Guo, F., Klauer, S., Hankey, J., Dingus, T., 2010a. Near crashes as crash surrogate for naturalistic driving studies. Transportation Res. Rec.: J. Transportation Res. Board 2147, 66–74. http://dx.doi.org/10.3141/2147-09.

Guo, F., Klauer, S., McGill, M., Dingus, T., 2010. Evaluating the relationship between near-crashes and crashes: can near-crashes serve as a surrogate safety metric for crashes?. Contract No. DOT HS (811:October), p. 382.

Hauer, E., Persaud, B.N., 1984. Problem of identifying hazardous locations using accident data. Transportation Res. Record ((No. HS-03:975) 49.

Hauer, E., 1995. On exposure and accident rate. Traffic Engineering & Control 36, 134–138.

Hauer, E., 1996. Identification of sites with promise. Transportation Res. Rec.: J. Transportation Res. Board 1542, 54–60. http://dx.doi.org/10.3141/1542-09.

Imkamon, T., Saensom, P., Tangamchit, P., Pongpaibool, P., 2008. Detection of hazardous driving behavior using fuzzy logic. In: 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, vol. 2, pp. 657–660. https://doi.org/10.1109/ECTICON.2008.

4600519.

Johnson, D.A., Trivedi, M.M., 2011. Driving style recognition using a smartphone as a sensor platform. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 1609–1615, https://doi.org/10.1109/ITSC.2011.6083078.

Jun, J., Ogle, J.H., Guensler, R., 2007. Relationships between crash involvement and temporal-spatial driving behavior activity patterns using GPS instrumented vehicle data. In: Annual Meeting of the Transportation Research Board, pp. 1–17.

Karaduman, O., Eren, H., Kurum, H., Celenk, M., 2013. An effective variable selection algorithm for aggressive/calm driving detection via CAN bus. In: 2013 International Conference on Connected Vehicles and Expo, ICCVE 2013 - Proceedings, pp. 586–591. https://doi.org/10.1109/ICCVE.2013.6799859.

Kazazi, J., Winkler, S., Vollrath, M., 2015. Accident prevention through visual warnings: how to design warnings in head-up display for older and younger drivers. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 1028–1034.

Kim, K., Brunner, I., Yamashita, E., 2006. Influence of land use, population, employment, and economic activity on accidents. Transp. Res. Rec. 1953 (1953), 56–64 (10.3141/1953-07).

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J., Ramsey, D.J., 2006. The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. https://doi.org/DOTHS810. Analysis 594.

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34 (3), 313–321 (10.1016/S0001-4575(01)00028-8).

Koenig, K., Lanzilotta, C., LaValle, S., Pande, R., Vaidya, M., 2016. The Internet of Things in Insurance. http://www.ey.com/Publication/vwLUAssets/EY_-_The_internet_of_things_in_insurance/$FILE/EY-the-internet-of-things-in-insurance.pdf (Accessed August 28, 2017).

Krueger, R., Rashidi, T.H., Rose, J.M., 2016. Preferences for shared autonomous vehicles. Transportation Res. Part C: Emerging Technol. 69, 343–355. http://dx.doi.org/10.1016/j.trc.2016.06.015.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 1659–1673. http://dx.doi.org/10.2307/1939924.

Li, Z.B., Chung, K., Cassidy, M.J., 2013. Collisions in freeway traffic: influence of downstream queues and interim means to address them. Transp. Res. Rec. 5003 (2396), 1–9 (10.3141/2396-01).

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Res. Part A: Policy Practice 44 (5), 291–305 (10.1016/j.tra.2010.02.001).

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods Accident Res. 1, 1–22. http://dx.doi.org/10.1016/j.amar.2013.09.001.

Mannering, F.L., Washburn, S.S., 2012. Principles of Highway Engineering and Traffic Analysis. John Wiley and Sons, vol. 1. https://doi.org/10.1017/CBO9781107415324.004.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions. Accid. Anal. Prev. 26 (4), 471–482 (10.1016/0001-4575(94)90038-8).

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accid. Anal. Prev. 28 (6), 739–753 (10.1016/S0001-4575(96)00050-4).

Pande, A., Chand, S., Saxena, N., Dixit, V., Loy, J., Wolshon, B., Kent, J.D., 2017. A preliminary investigation of the relationships between historical crash and naturalistic driving. Accid. Anal. Prev. 101, 107–116. http://dx.doi.org/10.1016/j.aap.2017.01.023.

Park, S., Kim, S., Ha, Y., 2016. Highway traffic accident prediction using VDS big data analysis. J. Supercomputing 72 (7), 2815–2831 (10.1007/s11227-016-1624-z).

Quddus, M., 2013. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and GIS. J. Transportation Saf. Security 5 (1), 27–45 (10.1080/19439962.2012.705232).

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accid. Anal. Prev. 40 (4), 1486–1497 (10.1016/j.aap.2008.03.009).

Reifel, J., Hales, M., Pei, A., Lala, S., Bhardwaj, N., 2014. The Internet of Things: Opportunity for Insurers. https://www.atkearney.com/documents/10192/5320720/Internet+of+Things+-+Opportunity+for+Insurers.pdf/4654e400-958a-40d5-bb65-1cc7ae64bc72 (Accessed August 28, 2017).

Ryder, B., Gahr, B., Egolf, P., Dahlinger, A., Wortmann, F., 2017. Preventing traffic accidents with in-vehicle decision support systems - the impact of accident hotspot warnings on driver behaviour. Decis. Support Syst. 99, 64–74. http://dx.doi.org/10.1016/j.dss.2017.05.004.

Sanitas, 2017. Sanitas Active App. https://www.sanitas.com/en/index/private-customers/bewegung/sanitas_active_app.html?cmpID=print_sanmag_16 (Accessed August 28, 2017).

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Anal. Prevention 27 (3), 371–389 (10.1016/0001-4575(94)00078-Z).

Sheehan, B., Murphy, F., Ryan, C., Mullins, M., Liu, H.Y., 2017. Semi-autonomous vehicle motor insurance: a bayesian network risk transfer approach. Transportation Res. Part C: Emerging Technol. 82, 124–137. http://dx.doi.org/10.1016/j.trc.2017.06.015.

Tey, L.S., Ferreira, L., Wallace, A., 2011. Measuring driver responses at railway level crossings. Accident Anal. Prevention 43 (6), 2134–2141 (10.1016/j.aap.2011.06.003).

Wang, Y., Kockelman, K.M., 2013. A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accid. Anal. Prev. 60, 71–84. http://dx.doi.org/10.1016/j.aap.2013.07.030.

Werneke, J., Vollrath, M., 2013. How to present collision warnings at intersections? - a comparison of different approaches. Accid. Anal. Prev. 52, 91–99. http://dx.doi.org/10.1016/j.aap.2012.12.001.

Yang, B.-M., Kim, J., 2003. Road traffic accidents and policy interventions in Korea. Injury Control Saf. Promotion 10 (1–2), 89–94 (10.1076/icsp.10.1.89.14120).

Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prev. 66, 80–88. http://dx.doi.org/10.1016/j.aap.2014.01.017.

Glossary

ADT:	Average Daily Traffic
CAN:	Controller Area Network
CDE:	Critical Driving Event
CF_{Pit}:	Number of accidents within location i from population P in timeframe t
CR_{Pit}:	Number of accidents per 100-million vehicle transits of population P in timeframe t for location i
FEDRO:	Swiss Road Authority
GPS:	Global Positioning System
JF_{Sit}:	Number of high jerk events within location i from sample (fleet) S in timeframe t
JR_{Sit}:	Number of high jerk events per vehicle transit of sample (fleet) S in timeframe t for location i
OBD-II:	On-Board Diagnostics
TF_{Pit}:	Number of transits of location i of population P in timeframe t – in our case the average ADT over the period of the field study multiplied by the length of the study
TF_{Sit}:	Number of transits of location i from sample (fleet) S in timeframe t – in our case the number of trips through that location during the period of the field study