Exploring the relationships between drivers’ familiarity and two-lane rural road accidents. A multi-level study

EXPLORING THE RELATIONSHIPS BETWEEN DRIVERS' FAMILIARITY AND TWO-LANE RURAL ROAD ACCIDENTS. A MULTI-LEVEL STUDY.

Paolo Intini*^

paolo.intini@poliba.it

p.intini88@hotmail.it

Nicola Berloco*

nicola.berloco@poliba.it

Pasquale Colonna*^

pasquale.colonna@poliba.it

Vittorio Ranieri*

vittorio.ranieri@poliba.it

Eirin Ryeng**^

eiring.ryeng@ntnu.no

*Department of Civil, Environmental, Building Engineering and Chemistry

Technical University of Bari

Via Orabona 4, 70125 Bari, Italy

**Department of Civil and Environmental Engineering

Norwegian University of Science and Technology

Høgskoleringen 7, 7030 Trondheim, Norway

^Corresponding Authors

Final draft of the paper published in Accident Analysis and Prevention:

Intini, P., Berloco, N., Colonna, P., Ranieri, V., & Ryeng, E. (2018). Exploring the relationships between drivers' familiarity and two-lane rural road accidents. A multi-level study. *Accident Analysis & Prevention*, *111*, 280-296.

DOI: https://doi.org/10.1016/j.aap.2017.11.013

ABSTRACT

Previous research has suggested that the familiarity of drivers with a given route may affect their performances. Some road safety-related pitfalls were highlighted for both the familiar (distraction and more dangerous behaviours) and the unfamiliar drivers (expectations not matching the reality). Moreover, the interactions between those two categories are a safety issue because of behavioural differences. Previous studies focused on relationships between drivers' familiarity and road crashes, showing differences in the definition of familiarity. In this article, a novel measure of familiarity was introduced based on the distance from residence; overcoming some previous limitations.

The relationships between familiarity and accidents were searched based on a traffic and accident database, referred to rural two-way two-lane sections of two important arterial Norwegian highways (E6, E39). A multi-level strategy of analysis, from a macro perspective to more detailed levels, was employed. In the macro analyses, the comparison of accident rates between different seasons and different values of summer traffic variation were taken as a reference variable. At the second level, a logistic regression model was used to explain the familiarity/unfamiliarity of drivers involved in crashes, considering a list of variables retrieved from the database. In the last step, an in-depth analysis of the relationships between familiarity and different accident types and dynamics was performed. The detailed analyses seem to suit better than the macro analysis the aims of the study, since no differences were found between accident rates in the different considered conditions. Conversely, some traffic and accident-related factors were related to familiar and/or to unfamiliar drivers: seasonal summer traffic variations, speed limits, vehicle types (heavy vehicles), travel purposes (commuters/work travelers), drivers' age (young drivers). To a minor extent, some indications arise from the crash in-depth analyses about types and dynamics, especially for familiar drivers.

Keywords: Route Familiarity, Accident Analysis, Accident Rate, Logistic Regression, Accident Type, Interactions between Drivers

1. INTRODUCTION

The strong influence of driving behavior on road crashes has been recognized since decades. Human, vehicle, road, environment and traffic are the five categories of contributing factors to accidents occurring (see e. g. Colonna, 2002). Anyway, their relative incidence is completely disproportionate in favor of human factors (see e.g. Treat et al. 1979, Singh, 2015). Therefore, road engineers and traffic safety researchers, have the urgent need of considering this challenging matter in both the design and the safety-based maintenance activities (see e.g. Campbell et al., 2012).

Singh (2015) estimated that the most frequent driver-related critical errors (more of 90 % of the total) are the recognition errors, which account for 41 % and are related to drivers' inattention, distraction and inadequate surveillance. This is confirmed by several other works which recognized driver distraction as a crucial causal factor in the crash occurring (e.g. Sandin, 2009; Staubach, 2009; Regan et al., 2008; Klauer et al., 2006; Young and Salmon, 2012). Moreover, this is coherent with the fourth law of accident causation proposed by Elvik (2006), the "law of cognitive capacity": the more the cognitive capacity approaches its limits, the greater is the increase in the accident rates. Therefore, as long as distraction and inattention affect negatively the cognitive capacity, accident rates can increase.

1.1 Familiar drivers and accident risk

The issue widely studied of driver distraction can be strongly linked to the drivers' route familiarity, which however is a topic less frequently considered. The route familiar drivers are road users who frequently travel on the same route, having perfect knowledge of the road environment and of all of its characteristics. This comes directly from the definition of the adjective *familiar*: well known from long or close association (Oxford, 2016), as long as no clear definition of "route familiar driver" was found in literature. In fact, the concept of route familiarity was considered in different ways: on a time-based scale (drivers categorized as familiar if the road was traveled at least once a week by Liu and Ye, 2011; or once a month by Beijer et al., 2004, and Bertola et al., 2012); on a distance-based scale (town limits of the driver's residence were used as a boundary for defining familiarity e.g. by

Rosenbloom et al., 2007) or on a more complex way, by considering that drivers can be familiar with a route only during specific times of the day or roadway conditions (Lotan, 1997).

Anyway, a typical example of route familiar user is a driver repeating almost daily its travel from home to work, which is also a frequent driving condition (about a third of the vehicle miles traveled related to private vehicles are for commuting according to: AASHTO, 2013). In this case, if no other unexpected events arise and if the user is enough experienced with the driving process itself (excluding novice and very-low mileage drivers), then the driving process is in the "habituation" stage (Colonna et al., 2016). This is a low-energy consumption state of the driver in which the response to external stimuli is reduced, coherently with Malleable Attentional Resource Theory (MART) by Young and Stanton (2002), the dual-process theory (Rankin et al., 2009) and the external and internal risk model (Colonna and Berloco, 2011). In fact, driving on a familiar route is mostly an automatic process, in which skill-based tasks are unconscious (Rasmussen, 1986). Therefore, route familiarity can lead to distraction and inattention by favoring mind wandering: the mind is occupied by thoughts not concerning the driving task and consequentially, responses to external stimuli are potentially slowed down. Thus, route familiarity can be involved in the same problems related to accident proneness discussed above while considering distraction. This theoretical and logical assumption is supported by some research. Yanko and Spalek (2013) found that route familiar users (who had driven on the simulated route four times before the test) needed greater reaction times than the unfamiliar (who drove on the experimental route for the first time during the test) in order to respond to unexpected external stimuli introduced in the scenarios: pedestrians crossing the road or lead vehicle suddenly braking. These results are similar to what found by Martens and Fox (2007) from another similar study based on driving simulation. In this case, priority road signs were modified in the last driving test, in which drivers could have been considered route familiar due to test repetitions. Only 2 out of 12 drivers noticed a change in signs, indicating possible inattention for familiar drivers.

Therefore, route familiarity can cause inattention. Anyway, this is not the only measurable output of a familiarization process. Familiarity with a given road environment can be a synonymous of more self-confidence and more risk-taking behaviors especially for more aggressive drivers (Colonna et al., 2015). Rosenbloom et al. (2007) observed the driving behavior of a sample of female drivers in both familiar and unfamiliar locations. They found that drivers performed more traffic violations, dangerous behaviors and speeding while driving in more familiar locations. The same tendency of speed increasing for familiar drivers was found by Colonna et al. (2016) from an on-road test. They also highlighted that this tendency is roughly independent from road geometry, being more related to the drivers' attitude to risk (even if a similar experiment conducted by Intini, 2014; but on a different road environment, with a smaller sample and a different measuring apparatus, did not reveal the same speed increase over days). A driving simulator study conducted by Bertola et al. (2012) revealed that drivers who acquired familiarity with the test route increased their speed and mean standard deviation of lateral position. Moreover, a pilot study by Colonna et al. (2016) inquiring into changes in chosen curve trajectories with the acquired route familiarity revealed that familiar drivers are more prone to curve-cutting behavior and encroachments, highlighting also the role of drivers' attention at horizontal curves (Charlton, 2007). Therefore, familiar drivers may try to maximize their mobility benefits in terms of reduction of travel time, but this leads to an increase of the accident risk due to the speed increase and to more dangerous behaviors (Noland, 2013; Nilsson, 2004; Intini et al., 2016). These findings show that the drive-related measurable parameters speed and lateral position can change with the acquired route familiarity towards a less safe scenario. Considering again the American statistics by Singh (2015), the second most frequent driver-related critical errors related to crashes are the decision errors (speeding, false assumptions of others' actions, illegal manoeuvers and misjudgment of gap and others' speeds) accounting for the 33 % of the total driver-related accidents. Therefore, familiarity can be involved also in this other group of errors.

1.2 Unfamiliar drivers and accident risk

Based on what stated above, the logical conclusion could be that unfamiliar drivers are safer than familiar drivers in respect to a given route. This is because it is expected that unfamiliar drivers should be in the road “studying” phase, where the attentional capacity is almost entirely devoted to the acquisition of the information related to the road environment. Therefore, they should be less inclined to distraction and less prone to speeding and risk-taking behaviors because the road is not well known. However, these conclusions do not take into account other important features. In the road design guidelines, it is commonly followed this good practice principle: road design should be thought for users who are driving on a roadway for the first time and who have no familiarity with its features (Milliken et al., 1998), This need is also coherent with the concept of the self-explaining roads (Theeuwes and Godthelp, 1995; Charlton et al., 2010; Mackie et al., 2013). A sudden sharp curve after a long straight section of road is unexpected and dangerous for all drivers, because the reality (the unexpected curve) does not match the expectations built up during the previous long stretch of straight road. However, the curve is truly unexpected only for the unfamiliar drivers who never or rarely traveled on that road and could lead to errors in speed and steering. This simple example easily explains why also route unfamiliar drivers can show some weakness regarding road safety.

Moreover, the Highway Capacity Manual (TRB, 2000) suggests to take into account the vehicular composition of traffic flow with regard to route familiarity by the introduction of a coefficient (driver population factor) in the calculation of the equivalent flow rate Vp for highways and freeways:

$$Vp = \frac{V}{PHF * N * f_{HV} * f_p} \tag{1}$$

where: Vp = 15-minute passenger-car equivalent flow rate (pcphpl); V = hourly volume (pc/hr); PHF = Peak Hour Factor; N = number of lanes in one direction; f_{HV} = heavy-vehicle adjustment factor; fp = driver population adjustment factor, variable between 0.85 (strong presence of recreational users such as tourists in the traffic flow) and 1 (flow mainly composed of regular users such as commuters). This means that other conditions being equal, in the context of uninterrupted flows, a decrease in fp due to the presence of unfamiliar drivers, corresponds to an increase in both the Vp (equivalent traffic flow rate) and the car density (equivalent passenger cars/km), and to a worsening in the level of service of the road. Therefore, according to this, the interaction between familiar and less familiar drivers in the traffic flow could potentially be related to an increase in the accident risk, since multi-vehicle accidents increase with density (Lord et al., 2005). Moreover, according to Wang et al., 2015; the traffic speed variance could be affected by an increase in density (for values far from the congestion) and this effect could increase the accident risk too (Garber and Gadiraju, 1989).

1.3 State of the art about relationships between accidents and familiarity

On summarizing, familiar drivers seem to be prone to inattention and risk-taking behavior, while unfamiliar drivers could be involved in errors due to unexpected road features or unexpected interaction situations. All these phenomena can be related to an increase in the accident risk.

Relationships between familiarity and road accidents have been investigated in literature by using two types of strategies: post-crash surveys and accident data analysis. In studies based on post-crash surveys, familiarity is one of the several factors considered for the analysis of a specific crash type. In studies based on accident data, the focus was exclusively on familiarity. It was mainly analyzed by considering the relative differences: urban/rural environments and foreign/local drivers; identified through zip codes or driving licenses. Findings related to those studies are summarized in Table 1.

(Table 1 here)

The studies considered in Table 1 show that there are several factors (related to road, driver and environment) to be controlled for, when comparing accidents to familiar and unfamiliar drivers. Moreover, the different perspectives on the definition of familiarity (frequency-based or distance-based) do not easily allow the consideration of the results.

1.4 Research questions

Considering the sum of all these issues, these research questions arise: 1) is route familiarity positively or negatively related to road safety in terms of occurring accidents? 2) Since familiar and unfamiliar drivers can be potentially related to accident risk through different mechanisms, can route familiarity be associated to particular types of crashes?

Previous studies found in the literature can only partially answer to these research questions. However, on summarizing the findings shown in the previous sub-sections, and trying to catch their underlying common thread, three hypotheses are considered in this article:

- Both familiarity and unfamiliarity can be risk factors in respect to accidents (see 1.1 and 1.2);
- Considering the different subjective, environmental and road characteristics related to crashes, they can be differently associated to the familiar and/or to the unfamiliar drivers.
- Familiar/unfamiliar drivers show different relative involvement in different accident types;

These hypotheses are tested based on a combined analysis of a vehicle crash database and of traffic counts. The analysis was focused on two-way two-lane rural highway segments in Norway in order to consider only a particular road environment. The remainder of the paper summarizes the process of data collection (Section 2), the methods employed for data analysis (Section 3) and the presentation and discussion of results (Section 4). Conclusions are drawn in Section 5.

2. METHODS

2.1 Database

The employed database is composed by two separate archives: one including traffic volumes and the other with crash data. Both of them were provided by the Norwegian Public Road Agency (NPRA). Traffic data refer to a period of 10 years, from 2005 to 2014. Traffic counts relate to two important Norwegian arterial roads: 102 counts from the E6 (from Trelleborg, Sweden to Kirkenes, Norway; Norwegian itinerary from the southern Swedish boundary to Kirkenes, for a length of 2628 km) and 77 counts from the E39 (from Trondheim, Norway to Aalborg, Denmark; Norwegian itinerary from Trondheim to Kristiansand, for a length of 1140 km). The information about traffic include the Annual Average Daily Traffic (AADT) and the average daily traffic in the months of June, July and

August (named SDT, Summer Daily Traffic). Among all the 179 traffic count stations, some of them have been implemented only recently, resulting in the presence of missing data. However, for about 90 per cent of the stations, traffic volumes are available for at least three years out of ten in the inquired 10-years period (for all the measures considered). Further, a continuous series of 10-years traffic values are available for 76 sites out of 179. Based on these complete time series, a yearly rate of traffic growth was computed (e.g. from 2005 to 2006, 2006 to 2007) and averaged over the samples belonging to the two different roads. Then, these traffic growth rates (estimated overall mean rate is 2.4 % per year) were used to reconstruct the missing data (about one fourth of the entire sample).

Accident data refer to the same period of 10 years, from 2005 to 2014 and to the same roads E6 and E39 in the Norwegian territory. In particular, the accident database is composed by 6992 traffic accidents, which involved 13126 vehicles and 17108 persons. The database includes only fatal and injury accidents in which at least one vehicle was involved. It is formed by three connected spreadsheets: the first reporting information about the accident; the second reporting about the units involved in each accident, and the third reporting about the persons involved. The presence of both coordinates (longitude/latitude) and the meter of road section in which each crash occurred (codified according to the Norwegian road registry) allowed the localization along the two studied road trunks.

2.2 Selection of road sites

Since accident data were analyzed in combination with traffic data, the selection of road sites to be inquired along the considered routes (E6 and E39) was based on the available traffic data. The study is focused on the two-way two-lane undivided rural road segments, so only traffic data belonging to this road category were taken into account. Values of traffic counts can be considered constant along the road segment if no significant intersections are present on it. Therefore, the preliminary operation of the selection process consisted in associating each traffic count to a two-lane road segment included between two intersections, which represent the limit bounds of the segment (see Fig. 1). As long as both the E6 and the E39 belong to the national road network “Riksvei” (first-level roads), the connections between the inquired roads and roads of the same level of importance or with roads belonging to the county road network "Fylkesvei" (second-level roads), were considered as "intersections". Driveways or other minor intersections were not considered for the further analyses as intersections on the segments, since their influence on the traffic variation along the road sections is considered to be not significant. The study is focused on accidents on road segments and so, distances equal to 150 m taken on the sections near intersections were not considered as part of the road segments to be inquired, as a safety margin to remove the influence of intersections on accidents occurring. Furthermore, the minimum length of the sections associated to a traffic count was fixed in 1 km, discharging very short sections connecting two consecutive intersections. Given all the criteria above explained, very short road segments, two-lane sections divided by median barriers, segments in urban or suburban environments (in which the noticeable influence of urban settlements could alter the hypothesis of rural segments) were not included in the dataset. If the traffic count station is placed on a segment included between an intersection and a urban center, then the section considered for the analyses starts 150 m after the end of the urban center and it ends 150 m before the intersection. The road sections meeting all the requirements above defined are henceforth referred to as "road sites".

(Figure 1 here)

The final dataset is composed by 84 road sites (E6: 37, E39: 47) and 633 road crashes. For each road site, a traffic count station provides traffic data. Details about the sites are summarized in Table 2.

(Table 2 here)

2.3 Procedure

Three different and complementary strategies were used in this study, reported as follows.

2.3.1 First Level: Macro-Scale Analysis

Since SDT (Summer Daily Traffic) data were available for all the considered road sites, then seasonal variations related to the summer traffic volumes were assessed, by using the AADT values as a term of comparison. The overall mean SDT/AADT ratio (see Table 2) is equal to 1.30 (st. dev. = 0.21). Summer traffic volumes are therefore considerably higher than the yearly volumes. It is reasonable to assume that a part of this increase in traffic during summer months (a higher amount from 10 % to 50 %) can be due to recreational drivers or tourists, barely familiar with the road. Therefore, a first measure of the influence of familiarity on accidents can be obtained by comparing: 1) accident rates in summer months with accident rates in the other months; 2) within the sample of summer accidents, the rates at sites with high seasonal traffic variation with rates at sites with low seasonal traffic variation. If significant differences will be revealed, then a macro-scale effect of familiarity on accidents will be highlighted, by considering only the number of accidents and the traffic volumes. This procedure implicitly assumes that accident rates are constant with the increase in the traffic volume. Actually, this can be considered valid for the type of roads considered (AASHTO, 2010).

2.3.2 Second Level: Analysis based on the Accident Database

A crash database including information about drivers, road, vehicle and environmental features was available. Based on this, the relations between familiarity and accidents were analyzed in more detail. The persons involved in the accidents occurred at the inquired road sites were classified into familiarity classes ("familiar", "unfamiliar", "transition") according to distance-based measures related to drivers (see 2.4: Measures). Thereafter, statistical analyses were performed in order to identify which variables can significantly be related to each of the so defined categories of drivers' familiarity (see Section 3: Data Analysis Techniques). However, as discussed in 1.2, the interaction between familiar and unfamiliar drivers could also be potentially related to the crash occurrence. Hence, for studying these phenomena in more detail, a further level of analysis was developed and described as follows.

2.3.3 Third level: Detailed Analysis of Accident Types and Dynamics

In this part of the study, both the possible over-involvement of familiar/unfamiliar drivers in different types of accidents and the diverse role played by them in the accident dynamics were analyzed in detail, to reveal particular accident patterns related to drivers' familiarity. This aim was pursued through the use of further statistical analyses (see Section 3). The information about accident types and dynamics were deduced from the database, whereas the familiarity of drivers was based on distance-related classification previously mentioned.

2.4 Measures

In this section, the measures used for the analyses are defined, differentiated for each level of analysis.

2.4.1 First Level Analysis

Distance of drivers from residence. The distance of drivers from the place of residence is a crucial measure for the development of this study, since it is used to define the familiarity of involved drivers with the road sites and to classify them. The distance of the drivers from the place of residence is based on the zip code associated to them. In fact, the distance between the exact point in which the accident occurred (taken from the coordinates in the database) and the center of the town/city associated to the zip code, was identified as the distance from the residence for each driver involved in the accidents. Each distance was computed by considering the road itinerary characterized by the shortest time travel connecting the residence and the accident site among all the possible alternatives.

SDT/AADT. The SDT (Summer Daily Traffic) is the average daily traffic during the months: June, July, August. As explained above, the ratio between the SDT and the AADT values can be used as a surrogate measure of the share of unfamiliar drivers in the traffic flow, since it is almost impossible to exactly quantify this share in a road section. For the road sites considered, it varies between 0.99 and 1.91 (mean: 1.30 ± 0.21). A preliminary analysis consisted in plotting the average distance of drivers from residence for each site (y-axis) against the SDT/AADT ratio of that site (x-axis). The regression line showed in Fig. 2 clearly indicates an increase of the average distance with the summer seasonal traffic variation. This means that, on average, at road sites with higher rates of traffic variation in summer months, drivers involved in the accidents are more likely to be further from home than at sites with lower traffic variation. Since higher average distances from the place of residence can be reasonably related to a greater presence of unfamiliar drivers in the traffic flow, then the SDT/ADT ratio can be used to categorize road sites based on the summer seasonal traffic variation.

(Figure 2 here)

Accident rates. Two strategies were chosen in this paper for identifying macro-scale effects of familiarity on accidents: 1) comparing the accident rates in June, July and August (summer months as defined in the traffic data) with the accident rates in the other months at all sites, 2) comparing the accident rates computed in summer months, between sites showing high and low seasonal traffic variation. The measures of the accident rates computed for each site are defined as follows:

$$ACC.RATE_{summer,\ site\ i} = \frac{N_{accidents,\ occurred\ in\ summer\ months\ over\ the\ 10\ years\ period\ at\ site\ i} * 10^6}{2.5\ years * 365 * mean\ SDT_{(over\ the\ 10\ years)} * Length_{\ (road\ section)}} \left[\frac{accidents}{MVKT}\right] \quad (2)$$

$$ACC.RATE_{other\ seasons,\ site\ i} = \frac{N_{accidents,\ occurred\ in\ all\ the\ other\ seasons\ over\ the\ 10\ years\ period\ at\ site\ i} * 10^6}{7.5\ years * 365 * mean\ OSDT_{(over\ the\ 10\ years)} * Length_{\ (road\ section)}} \left[\frac{accidents}{MVKT}\right] \quad (3)$$

All the terms present in the equations 2 and 3 are coherent with the description given in the Section 2.1 (Databases), except for the OSDT (Other Seasons Daily Traffic), which is the estimate of the average traffic volume in the other seasons except for summer, obtained by inverting the equation 4. It is a simple proportion for distributing the annual traffic in the different considered year periods.

$$AADT\ (mean, over\ the\ 10\ years) = \frac{OSDT\ (mean,\ \ over\ the\ 10\ years) * 9 + SDT\ (mean,\ \ over\ the\ 10\ years) * 3}{12} \quad (4)$$

For the same reason, the period used for the computation of accident rates in summer over the ten years period is 2.5 years (one quarter of the total) and the complementary period of 7.5 years (three quarters of the total) was used for the computation of accident rates in the other seasons.

Road sites showing high/low summer traffic variation. The strategy used for the first level analysis implies the differentiation between road sites with high summer traffic variation (supposed to be partly due to recreational drivers more unfamiliar with the road) and the road sites with low summer traffic variations (where a minor presence of unfamiliar drivers are expected in summer due to a constant trend of traffic volumes over the year). There are no objective measures to define if a site shows high or low traffic variation. In fact, there are no exact definitions of thresholds after which an average seasonal traffic variation (in comparison with AADT values) can be considered as "high". For this reason, sites were classified into traffic variation classes by using cluster analysis (see Section 3) in order to avoid assigning labels to sites based on a-priori deterministic thresholds.

2.4.2 Second Level Analysis

"Familiar" and "unfamiliar" drivers. In order to perform statistical analyses, drivers involved in the accidents present in the dataset were divided into classes based on their familiarity with the road sites investigated. The definition of the familiarity or unfamiliarity of drivers with given road sections is a complex matter to address, as explained in the introductory section. It should be based on behavioural differences: the human brain acts in different ways if the subject lies in a "habituation" stage, typical of the confidence with a given situation. However, in order to individuate some relationships between an accident already occurred and the drivers' familiarity, some measurable indicators have to be necessarily considered, which are able to reveal those relationships. Anyway, this process is surely affected by some errors. In this article, the familiarity of drivers with the place of the traffic accident was based on the measure: "distance of drivers from residence" above explained. In fact, it is reasonable to assume that, on average, drivers spend most of their annual mileage on roads near the place of residence and then, they are familiar with them (e.g. drivers are surely familiar with the roads traveled for home-work commuting). For the same reason, it is more likely that drivers are unfamiliar with roads very far from the place of residence. Anyway, the definition of a universal boundary distance after which a driver can be considered unfamiliar with the road is a complex matter to address. In a similar study, Rosenbloom et al. (2007) used the town limits of the driver's residence as a boundary for defining familiarity. In this study, literature data about mean travel patterns were used to give this definition rather than using arbitrary a-priori values. Anyway, the concept of familiarity cannot be reasonably regarded as a strict binary variable. This means that it is hard to define a threshold distance from the residence, beyond which the driver can become unfamiliar with the roads placed out of this boundary. Therefore, in this paper, the area surrounding the place of residence of the generic driver was divided into three concentric zones (see Fig. 3), identified as follows:

- The "familiar zone", the area in which the generic driver is assumed to be familiar with the roads present on it, is represented by the circular area in Fig. 3 defined by a radius of 20 km. This distance was set based on Norwegian data (Hjorthol et al., 2014) about mean distances of car commuting travels (15.8 km for car drivers and 21.7 km for car passengers). It is comparable to the value proposed by Litman (2003) about home-work trips (about 24 km) and with the estimated average time of about 1 hour needed for mobility for each person per day (Colonna, 2009). Furthermore, the distance of 20 km from the residence corresponds to the point for which the cumulative frequency curve of distances diverges from the initial linear tendency (Fig. 4). Drivers involved in accidents occurred in their so-defined familiar zone are slightly more than one third.

(Figures 3 and 4 here)

- The "unfamiliar zone", the area in which the generic driver is assumed to be unfamiliar with the roads present on it, is the area represented in Fig. 3, further than 200 km from the residence. Drivers involved in accidents occurred in their so-defined unfamiliar zone are about 15 % of the total. This distance was set based on the consideration that long trips are rarer for drivers than commuting trips and that, for them, other means of transport can be chosen rather than the car. Based on the same Norwegian data (Hjorthol et al., 2014), a “long trip” is defined as a travel for distances greater than 100 km, while Norwegian travelers prefer planes for distances greater than 300 to 400 km (Hjorthol, 2014; Thrane, 2015). In this study, an intermediate distance of 200 km was chosen since most travelers not choosing plane are likely to be unfamiliar with roads at distances included between 200 and 400 kilometers. Moreover, defining a “long trip” as a travel of minimum 100 km could reasonably include also a not negligible share of drivers familiar with those roads. This rule could be potentially violated by the professional drivers of long-distance bus or trucks, since they can be familiar with roads placed at long distances from the residence. However, these vehicles account for less than 8 % of the total vehicles involved in the accidents of the database. Anyway, the influence of the variable “heavy vehicle” was taken into account in the statistical analyses performed and further discussed. Moreover, drivers identified as foreign drivers in the database (no zip code present, but indication about the nationality) were assimilated to unfamiliar drivers. In fact, apart from the considerations about distances (they are likely to come from more than 200 km from the place of the accident in several cases), they could be considered as unfamiliar mainly because of their possible ignorance of the road environment in a foreign country. Some previous studies focused on these differences, by considering discrepancies between residents and foreigners (e.g. Yannis et al., 2007). However, in this case, the definition of a separate category of foreign drivers was avoided, because of their scarcity (41 out of 1091, accounting for only about the 4 %).

- The “transition zone” is the circular crown represented in Fig. 3, included between the familiar and the unfamiliar zones. In this area, the authors assumed that it is not possible to define the familiarity of the generic driver with the roads present in it, with a reasonable margin of error. Therefore, this area was not considered for this analysis.

Based on the assumptions explained above, the familiarity/unfamiliarity with the place of the accident can be defined for each driver involved (univocally related to each vehicle), based on the distance from his/her residence. After, for each crash, two binary variables were defined as follows. The variable "familiarity" was set to: 1 if at least one driver was distant 20 km or less from the residence, 0 if all drivers were distant more than 20 km. The variable "unfamiliarity" was set to: 1 if at least one driver involved was distant 200 km or more from the residence, 0 if all drivers were distant less than 200 km. The two variables above defined were used to develop separate models (see Section 3).

2.4.3 Third Level Analysis

Accident Types. The detailed accident types present in the database were clustered into these main categories: run-off, rear-end, lateral/angle, head-on and other crash (e.g.: with pedestrians, animals). Thereafter, rear-end and lateral/angle crashes were grouped together since the lateral/angle category was the less numerous (only 33 crashes, 5.2 %) and some of them were described as resulting in a rear-end. The 22 accidents (3.5 %) classified as "other" were removed from this third level analysis, because hard to be interpreted.

Familiarity categories of accidents. In order to deepen the study of the relations between accident types, familiarity and the interactions between different familiarity categories, the same measures of familiarity defined in 2.4.2 were used. Since in this case the interactions were considered, crashes were divided into the following seven categories according to the familiarity of drivers involved:

- only Familiar: crashes having involved only drivers distant 20 km or less from residence;
- only Unfamiliar: crashes having involved only drivers distant 200 km or more from residence;
- only Transition: crashes having involved only drivers distant from 20 to 200 km from residence;
- interaction Familiar/Unfamiliar: crashes having involved a combination of at least one familiar driver and at least one unfamiliar driver;
- interaction Familiar/Transition: crashes having involved a combination of at least one familiar driver and at least one transition driver;

- interaction Unfamiliar/Transition: crashes having involved a combination of at least one unfamiliar driver and at least one transition driver;
- interaction Familiar/Unfamiliar/Transition: crashes having involved a combination of at least one unfamiliar driver, at least one familiar driver and at least one transition driver.

The first three categories include crashes in which interactions between different familiarity categories of drivers did not occur (“no interactions”). The last four categories include crashes with interactions between the familiarity categories (“with interactions”). This separation will be useful for the analyses conducted. The accidents considered were 521 (373 without and 148 with interactions between categories), since crashes involving at least one vehicle with missing zip code were removed.

Accident Dynamics (assigned to each vehicle). The different detailed accident dynamics present in the database were clustered into four categories: “moving” (being involved in the crash while moving on the roadway), “stationary” (being involved while stationary on the roadway after braking or before turning), “out of control” (having lost control of the vehicle), “maneuvering” (being involved while turning into a driveway/minor intersection or while overtaking), “other dynamics” (other cases, e.g.: a vehicle hit while parked). No data about dynamics were found for 240 vehicles (22.0 %). Those vehicles were excluded, together with 20 (1.8 %) vehicles classified as “other” dynamics, hard to be commented.

Familiarity categories of vehicles. The relationships between dynamics and familiarity were inquired at a more detailed level too. In this case, for each vehicle involved in the crash, the role played in the accident dynamics can be defined (moving, stationary, out of control, maneuvering, other). Then, according to the measures used in 2.4.2, a familiarity class was associated to each driver (familiar, unfamiliar or transition) and then to each vehicle (there is only one driver for each vehicle). The final cases considered were 776, after the vehicles showing missing zip codes were removed.

3. DATA ANALYSIS TECHNIQUES

Details about the analysis techniques used at the different levels of study are given in this section. A summary of the strategies and techniques of analysis employed in this article is given in Table 4.

3.1 Macro-Scale Analysis (First Level)

In order to meet the objectives explained in Section 2.3.1, the classification of road sites was based on the technique presented in Section 3.1.1. Furthermore, the analyses described in Section 3.1.2 allowed to make comparison between accident rates.

3.1.1 Cluster Analysis of Road Sites according to their SDT/AADT ratios.

Cluster analysis was carried out in order to group road sites into clusters characterized by similar average summer traffic variation rates. The two-step algorithm described by Chiu et al. (2001) was performed, by adopting the log-likelihood distance. Since the variable SDT/AADT ratio is continuous, then it is considered as normally distributed in the log-likelihood distance calculation. The optimal number of cluster was automatically defined by using a two-stage estimator, based on the Bayesian Information Criteriorn (BIC). Since the final solution may depend on the order of the cases, they were firstly arranged in random order to minimize the effect. Cluster analysis was carried out using the SPSS software, as well as all the other further analyses conducted in this study.

3.1.2 Statistical tests of Accident Rates

Statistical tests were performed in order to compare accident rates at inquired road sites in the diverse conditions considered. Data of accident rates were firstly analyzed for testing the normality and homoscedasticity assumptions. The normality assumption was verified using the Kolmogorov-Smirnov and the Shapiro-Wilk tests. These tests revealed that the normality assumptions can be rejected at the 5 % level of significance for both the distributions of accident rates (in summer, and at sites showing high seasonal traffic variation). Given that data distributions are not normal, accident rates were compared by non-parametric tests. The Mann-Whitney U test, a rank-based nonparametric test was used to determine if there are differences between two groups on the dependent variable. The determination of the differences is based on the medians of the groups if the shapes of their distribution are similar, otherwise it is based on mean ranks (Lehmann, 2006; Laerd Statistics, 2015). In detail, in the first analysis, the possible differences between the distribution of summer accident rates between the two groups (sites showing low and high seasonal traffic variation) were assessed. In the second analysis, the test concerned the difference in the distribution of accident rates at high traffic variation sites between the two groups: summer and other seasons.

3.2 Second Level Analysis

Based on 2.3.2, the relations between crash features and drivers' familiarity were inquired by using logistic regression, commonly used in similar cases (e.g. by Kim et al., 2012; Al-Ghamdi, 2002).

3.2.1 Binary Logistic Regression

The logistic regression explains a categorical dependent variable based on a set of variables either categorical or continuous. In the binary form, the dependent variable assumes only two values (i.e. one or zero, representing namely the presence or the absence of a given attribute in the sample of cases, see Harrell, 2015). It preserves many features of linear regressions, by requiring less assumptions and remaining robust even if they are not met (Sreejesh et al., 2014). It is based on the equation:

$$p(k) = \frac{e^k}{1+e^k} \tag{5}$$

where p(k) is estimated through linearization, involving the natural log of the odds of the event (logit):

$$logit(k) = ln\left(\frac{p(k)}{1-p(k)}\right) = ln(e^k) = k = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon = X\beta + \varepsilon \tag{6}$$

where ε is the error variable, X is a vector of variables and β is a vector of coefficients.

This is an exploratory study aimed at finding variables related to drivers' familiarity barely selectable a-priori based on previous similar analyses. Hence, a stepwise selection of variables was used to select the model variables of the vector X among the initial set. This approach is based on this iterative two-stage procedure: 1) a process of forward selection starting from the null model, choosing the variable showing the highest score statistic among all and including it in the model if its significance is lower than the cut-off value; 2) a process of backward elimination, choosing the variable showing the lowest score statistics and removing it from the model if its significance is higher than the cut-off value. The iterative process ends if no further variables can be added according to this algorithm. The vector of coefficients β is estimated through the maximum likelihood estimation. The Wald statistic is used as score statistic, with cut-off p-values of 0.05 for the entry and of 0.10 for the removal of variables.

The global significance of the logistic regression model is evaluated through the likelihood ratio test. In non-linear regressions, the explained variance cannot be estimated through the R^2 indicator. Hence, the pseudo-R^2 by Nagelkerke (1991), covering the full range: [0, 1], was used.

For the aims of the study, two logistic regression models were performed. A single model was developed for each of the two accident-related binary variables defined in 2.4.2: "familiarity" and "unfamiliarity". Each of these two variables is the dependent variable of each single model.

The considered explanatory variables are the same for the two models and they are described in Table 3. They are divided into three categories according to the information contained in the database (see e.g. Montella et al., 2013, for a dissertation about the structure of crash databases): variables related to the accident (including road, environment and traffic), variables related to the traffic units (vehicles and pedestrians) and variables related to the units involved.

(Table 3 here)

(Figure 5 here)

As long as more than one traffic unit and more than one person can be associated to each single accident, then the variables related to the traffic units and to the persons involved have to be rearranged. In fact, since the modalities of the dependent variables of each of the two models are univocally defined for each accident (e.g. the variable “familiar” assumes the value 1 if at least one driver involved in the accident was distant less or equal than 20 km from residence); then also the explanatory variables have to be univocally defined for each accident. Hence, for example, the unit-related variable “young persons involved (< 24 years)” is univocally determined for each accident by considering the following rule in defining dummy variables: it assumes value 1 if there is at least one young person involved in the accident and value 0 if no involved person is young. All other units and persons-related variables are transformed into dummy variables in the same way.

All the explanatory variables were modeled as categorical. Only the variables: traffic volume, traffic seasonal variation, time of the day, presence of additional lane, road width and speed limits could have been considered as metric. However, coherently, for the variable “traffic seasonal variation”, the same categorical classification used for the first level analysis was chosen. Apart from “hour of the day”, in this case meaningless as continuous; the other cited variables are characterized by a range of values strongly over-represented (traffic volume < 5,000; traffic seasonal variation < 1.40; road width included between 7 and 9 meters; speed limits ≥ 80 km/h). Hence, the differences between those categories and the extreme ones were inquired, rather than modelling them as continuous variables. Furthermore, the speed limits are not a continuous range of values, but discrete instead.

Once the binary logit model based on categorical predictors is estimated, the exponent of each β_i coefficient: $\exp(\beta_i)$, represents the odds ratio. It is the ratio of the odds of exhibiting the presence of the attribute (i.e. value “1” of the independent variable) to the odds of exhibiting the absence of the attribute (i.e. value “0” for the independent variable), for a given modality of the variable x_i with respect to the modality of reference not included in the model, controlling for the other variables. This is valid also for the not dichotomous explanatory variables (i.e.: showing j + 1 modalities). In fact, they are before converted to j dummy variables as well as in linear regressions, allowing the interpretation of the coefficients in terms of odds ratio similar to the binary variables.

Two out of the twenty-four variables extracted from the database were excluded. In fact, considering to have more than 30 cases for the categories of each variable, the variables "lanes" (presence of additional lanes) and "pedestrians/cyclists involved" did not meet this requirement. The person-related variables: "commuting traveler", "work traveler", "pleasure traveler" and "driving under influence", are affected by a number significantly higher than 30 missing data (no information about travel purposes or driving influenced for all drivers involved). However, the categories to be inquired (e.g.: at least one commuter or driver under influence involved) had more than 30 cases. Hence, for these variables, a further modality was added for accounting missing data, so avoiding their removal. According to the structure of the variables, the significance of the effect of a unit/person-related variable on the dependent one can be explained as follows. Let us suppose that the variable "young persons involved" significantly affects the "unfamiliarity", with odds ratio greater than 1. This could mean that, if at least one young driver was involved in the accident, then it was more likely to find an unfamiliar driver involved. However, this result can only indicate a possible relationship between young and unfamiliar drivers in the accident occurring. To clarify these relationships, further chi-square tests were conducted for each unit/person-related variable included in the final models, in order to definitely associate them with the drivers' familiarity. For example, a chi-square test was conducted between the variable unfamiliarity (groups: unfamiliar, not unfamiliar) and the variable young age (groups: young, not young), including the overall sample of drivers.

3.3 Micro-Scale Analysis (Third Level)

Based on 2.3.3, the relationships between the accident characteristics (types and dynamics described in 2.4.3) and the drivers' familiarity was inquired through chi-square tests of independence.

3.3.1 Statistical tests of Accident Types

For this analysis, the whole sample of accidents was used. The accidents were divided into categories according to familiarity of drivers as described in sub-section 2.4.3. However, an important distinction has to be made between the accidents in which interactions occurred or not occurred between drivers belonging to different familiarity categories. This is coherent with the possible importance of these interactions (see 1.2), and it could allow a better interpretation of results by avoiding to confound effects from different categories. Hence, two tests about the association between familiarity and accident types were performed: one considering the accidents in which all drivers were "familiar", "unfamiliar" or "transition" (first three categories defined in 2.4.3) and the other considering the accidents in which an interaction occurred between different drivers (categories from fourth to seventh). The last category (interactions between familiar, unfamiliar and transition drivers) was excluded from this analysis since it is composed of only 7 items (1 %) and it depends on the interactions between all the categories, being impossible to distinguish them. All the crash types (run-off, rear-end/angle and head-on) were considered in the first analysis. Run-off accidents were excluded from the second analysis since, in almost all cases except for 1, they were described as single-vehicle: no interactions were possible. Moreover, since in the second analysis the interactions between all drivers are considered, the transition drivers were included.

3.3.2 Statistical tests of Accident Dynamics

For the two chi-square tests presented in the previous sub-section, a familiarity category was assigned to each accident. However, a description of the role played by each vehicle in the accident dynamics was present in the database for almost all the vehicles, being related to the individual vehicle. Hence, a last chi-square test of independence was performed to test if there is association between familiarity and accident dynamics, considering in this case the sample of vehicles. The same sample was not used to search for associations between vehicles and accident types because it should not be considered as a set of independent measures: more than one vehicle/driver can be involved in each crash. Conversely, accident dynamics can be independently associated to each vehicle involved.

(Table 4 here)

4. RESULTS AND DISCUSSION

4.1 Macro-Scale Analysis (First Level)

4.1.1 Road sites clustering

Road sites were divided into two clusters (see 3.1.1), based on the variable SDT/AADT ratio. Two clusters were identified as the best solution through the clustering procedure. Hence, cluster 1 can be considered as composed of the sites showing low summer traffic variations and cluster 2 as composed of those showing high summer traffic variations. More details are shown in Table 5 and Figure 6.

(Table 5 here)

(Figure 6 here)

Most of the inquired road sites (64 out of 84, 76 %) belong to cluster 1, showing summer traffic variation with respect to the AADT less or equal than 40 %. There are only 20 sites (cluster 2, 20 out of 84 road sites, 24 %), for which the summer traffic volume is from 1.4 to almost 2 times the AADT.

4.1.2 Results of statistical tests on accident rates

Descriptive statistics about accident rates in the diverse considered conditions are shown in Table 6.

(Table 6 here)

A Mann-Whitney U test was performed to determine if there are differences in summer accident rates between sites showing low and high traffic variation. Distributions of accident rates for the two groups of sites were similar. Median summer accident rates for high variation sites (0.065) and low variation sites (0.065) were not statistically significantly different, $U = 596.5$, $z = -0.462$, $p = .644$.
A similar test was performed to determine if there are differences in accident rates at sites showing high summer traffic variation between summer and other seasons. Distributions of the accident rates for the two groups were similar. At high variation sites, median accident rates in summer (0.065) and in the other seasons (0.152) were not statistically significantly different, $U = 251$, $z = 1.394$, $p = .167$.

These statistics were aimed to test if an higher share of unfamiliar drivers in the traffic flow, indirectly measured through the SDT/AADT ratios, could be related to a difference in accident rates, a macro indicator. In fact, higher SDT/AADT ratios were supposed to be related to higher shares of unfamiliar drivers (see 2.4.1). Results from tests indicate that there is no statistical evidence of this influence. In fact, in the first analysis, no difference can be noted in summer between accident rates at road sites characterized by different shares of unfamiliar drivers in the traffic flow (high variation sites, SDT/AADT ratio: 1.42 - 1.91; and low variation sites, SDT/AADT ratio 0.99 -1.40). This indicates that shares of unfamiliar drivers higher than the mean (based on SDT volumes) seem not to be related to different accident rates, even if the number of sites are not equally represented: there is a small number of sites with high summer traffic variation. Also in the second analysis, in which only the high traffic variation sites were considered, no statistical difference were found between accident rates during the rest of the year and the summer (when the traffic can also be almost 2 times than the average volume, and it could be populated by a considerable share of recreational/unfamiliar drivers). These results could lead to three different arguments. The first is that using the SDT/AADT ratio for indirectly measure the share of unfamiliar drivers could be inappropriate. Anyway, the tendency shown in Fig. 2, coherently with logical deductions, can lead to reject this first hypothesis. The second argument is that the presence (and the relative volume) of unfamiliar drivers is not related to safety. The third hypothesis is that using a macro-scale indicator as the accident rate for measuring differences in safety performances between familiar and unfamiliar drivers could be unsuitable. In fact, these differences act at a personal level and, they can influence safety from different perspectives (see Introduction). The validity of these two last hypotheses will be assessed in the rest of the study.

4.2 Detailed Analysis (Second Level)

A binary logistic regression was performed to reveal the effect of the accident variables reported in Table 3 on the likelihood that at least one driver involved in the crash was familiar. A second similar regression was performed to reveal the effect of the same variables on the likelihood that at least one driver involved in the accident was unfamiliar.

4.2.1 Familiar model: Results

Among the 633 crashes in the sample, 357 (56.4 % of the total) did not involve familiar drivers (modality of reference), 246 (38.9 % of the total) involved at least one familiar driver, 30 (4.7 % of the total) showed missing zip codes for all units involved, being excluded from the analysis.

The logistic regression model was statistically significant, $\chi^2(15) = 137.836$, $p < .001$. The model explained 30.9% (Nagelkerke R^2) of the variance in familiarity and correctly classified 72.6% of cases. Only 8 out of the 22 variables resulted significant in the final model: traffic volume, traffic seasonal variation, season, section type, speed limit, vehicle age, commuting traveler, work traveler.

(Table 7 here)

The odds of finding at least one familiar driver involved in an accident is 2.30 times higher for traffic volumes greater than 11,000 than for volumes smaller than 5,000. A site showing high traffic seasonal variation is strongly less like to experience crashes involving familiar drivers than the other sites (OR = .27). The odds of finding a familiar driver inolved is 2.43 times higher in autumn and 2.72 times higher in winter than in summer. Crashes involving familiar drivers are more likely to happen at minor intersections (OR = 3.19) than at road segments. At sections with speed limits greater or equal than 80 km/h, familiar drivers can be more likely involved in crashes (OR = 1.89) than at sections with lower limits. If at least one vehicle aged 15 or more was involved in a crash, than the probability of having familiar drivers involved is higher, compared with the condition of all more recent vehicles (OR = 1.56). Having at least one commuter involved is strongly associated to at least one familiar driver involved (OR = 3.64). The opposite is noted for having at least one work traveler (OR = .50). The structure of the variables does not allow to draw direct conclusions about the influence of the vehicle and driver-related variables (in this case: vehicle age, commuting and work travel), as explained before (3.2.1). Thus, additional chi-square tests were performed for the vehicle/driver-related significant variables. Statistically significant associations were found between commuting and familiarity. More familiar drivers than the expected were involved in accidents while commuting wih respect to the other drivers ($\chi^2(1) = 31.994$, $p < 0.001$, count of familiar drivers involved in accidents while commuting = 34, adjusted residual with respect to other drivers = 5.7). Statistically significant associations were found between work travel and familiarity. Less familiar drivers than the expected were involved in accidents while traveling for work with respect to the other drivers ($\chi^2(1) = 10.146$, $p = .001$, count of familiar drivers involved in acccidents during work travels = 20, adjusted residual with respect to other drivers = -3.2). No association was found between familiarity and the vehicle age ($\chi^2(1) = 2.853$, $p = .091$), not confirming the results of regression if considering drivers directly. Hence, this variable was excluded from further discussions.

4.2.2 Unfamiliar model: Results

Among the 633 accidents in the sample, 446 (70.5 % of the total) did not involve unfamiliar drivers (modality of reference), 157 (24.8 % of the total) involved at least one unfamiliar driver, 30 (4.7 % of the total) showed missing zip codes for all units involved and they were excluded from the analysis. The logistic regression model was statistically significant, $\chi^2(12) = 118.518$, $p < .001$. It explained 29.1% (Nagelkerke R^2) of the variance in unfamiliarity, by correctly classifying 78.4% of cases. 7 out of the 22 predictors resulted significant in the final model: traffic volume, traffic seasonal variation, season, accident type, heavy vehicles involved, commuting traveler, young drivers involved.

(Table 8 here)

A site showing high traffic seasonal variation is strongly more likely to experience accidents involving unfamiliar drivers than the other sites (OR = 8.21). The odds of finding an unfamiliar driver inolved is 0.41 times lower in autumn than in summer. Accidents involving unfamiliar drivers are more likely to be head-on (OR = 2.47) and rear-end/angle crashes (OR = 1.75) than run-off road crashes. If at least one heavy vehicle (truck, light truck, camper, van, bus, vehicle with trailer) was involved in an accident, than there is a higher probability that unfamiliar drivers are involved in the accident compared to the situation including only light vehicles (OR = 2.04). The conditions of having at least one commuter and one young driver involved (≤ 24 years old) are less associated to have at least one unfamiliar driver involved (namely: OR = .40, OR = .63).

Also in this case, a chi-square test for each vehicle/driver-related variable was further performed. Statistically significant associations were found between unfamiliarity and all the variables investigated. More unfamiliar drivers than the expected were involved in accidents while driving heavy vehicles with respect to the other drivers ($\chi^2(1) = 21.864$, $p < 0.001$, count of unfamiliar drivers involved in accidents while driving heavy vehicles = 61, adjusted residual with respect to other drivers = 4.7). Less unfamiliar drivers than the expected were involved in accidents while commuting with respect to the other drivers ($\chi^2(1) = 9.694$, $p = .002$, count of unfamiliar drivers involved in acccidents while commuting = 2, adjusted residual with respect to other drivers = -3.1). Less unfamiliar drivers than the expected were involved in accidents if aged 24 or less with respect to the other drivers ($\chi^2(1) = 9.809$, $p = .002$, count of unfamiliar drivers 24 or less years old involved in acccidents = 27, adjusted residual with respect to other drivers = -3.1).

4.2.3 Discussion

The logistic regressions were aimed to highlight some influential variables on the occurrence of accidents to familiar/unfamiliar drivers.

Differently from what emerged from the first-level analysis, seasonal effects seem to have a great influence on the odds of familiar/unfamiliar drivers to be involved in crashes. Sites at which the SDT/AADT ratio is more than 1.40 have a probability about 8 times greater of experiencing accidents to unfamiliar drivers than the other sites, while coherently the opposite tendency was noted for the familiar ones. Moreover, an unfamiliar driver can be more likely involved in an accident during summer than in autumn, while the opposite tendency (autumn/winter) occurs for familiar drivers. These effects clearly revealed by this analysis were expected because the presence of recreational drivers in the traffic flow, unfamiliar with the sites, is greater in summer and at sites showing high traffic seasonal variation. However, there are also other variables associated with familiarity/unfamiliarity, which seem specific for the two categories.

It is more likely to find familiar drivers involved in accidents at sites with high traffic volumes (>11,000) than at sites with low volumes (< 5,000). This could be explained by the fact that those volumes can be reached on two-lane roads close to important towns/cities and near them it is likely that the traffic flow is mainly composed of familiar drivers. However, the same tendency was noted for the unfamiliar drivers too. In this case, a higher traffic volume can be related to more interactions between drivers. This seems to confirm what explained in the introduction: the interactions between drivers could be more demanding for the unfamiliar drivers with respect to the condition of low volume and minor interactions. However, it must be specified that this variable is referred to the annual volume and not to that at the moment of the crash, representing only an indicator of that value. The sample is composed of rural roads, therefore it can be possible that minor intersections and segments with lower speed limits cluster near urban centers. This could explain the odds of finding more familiar drivers in accidents at driveways/minor intersections compared to normal road segments and in accidents at road segments with speed limits lower than 80 km/h. However, speed limits lower than 80 km/h could be associated also to specific rural segments (i.e. the 22 % of the sites with posted speeds of 70 km/h, see Table 3) not close to urban centers. Therefore, the odds of finding familiar drivers involved in accidents could be greater (about 2 times higher) at these segments also because they can show more dangerous behaviors and take higher risks as they well know the road environment (see Introduction). Furthermore, by looking at chi-square tests too, it seems logical to find more familiar drivers involved in accidents while commuting to/from work or school than the other drivers, because of the definition itself of “familiar driver” used in this work based on distance. Coherently, there are less unfamiliar drivers associated to the variable “commuting travel”. On the other hand, a less immediate and interesting association was found with the traveling for work. Familiar drivers, basing also on chi-square tests, are less likely to be involved in accidents while traveling for work than the other drivers. However, in this case, the condition of work traveler cannot be clearly expected to be related to familiar or unfamiliar drivers: professional drivers or workers using vehicles for job-related reasons can cover short or long distances. Therefore, interestingly, there could be a higher probability to find familiar drivers involved in accidents while commuting and a lower probability while traveling for work. It could be argued that the driving style of familiar drivers is safer while traveling for work reasons then while commuting. This could be explained by the role played by attention while driving. In fact, a familiar driver could be more affected by the phenomenona of distraction and inattention while commuting, peculiar to them. On the other hand, a work travel could require more attention. An attentive familiar driver who knows well the road could be safer than other drivers. The variable vehicle age is not further discussed since a direct connection with the familiarity was excluded.

In the case of unfamiliar drivers, the relationship with different accident types can be noted: odds almost 2 times higher of finding unfamiliar drivers in rear-end/angle than run-off crashes and 2.5 times higher in head-on crashes than run-off crashes. Since run-off accidents are mostly single-vehicle crashes, it could be argued that unfamiliar drivers can be more likely found in crashes with interactions with other drivers. This seems to confirm what expected about this matter (see 1.2). On the other hand, unfamiliar drivers seem less prone to run-off crashes, coherently with Liu and Ye (2011). This could mean that unfamiliar drivers are more careful in more dangerous sections were run-off road crashes are more probable. However, for what concerns the interactions with other drivers, the structure of the variable "unfamiliarity" (at least one unfamiliar driver involved or no unfamiliar drivers) does not allow to consider them. Therefore, an in-depth analysis of the accident types and dynamics was conducted in the next stage, before drawing conclusions about interactions. Anyway, an influence was noted at this level, considering also the significance of the variable traffic: higher volumes cause higher interactions, and this was associated to the presence of at least one unfamiliar driver involved. In a crash involving a heavy vehicle (including trucks, light trucks, campers, vans, buses, vehicles with trailer), at least one unfamiliar driver can be more likely involved (as confirmed by the chi-square test conducted). Actually, more unfamiliar drivers than the other ones involved in the crashes were driving these vehicles. However, the work travel was not a significant variable in this regression, even if this could be explained by the presence of missing data. Hence, the association with heavy vehicles could not be only related to professional truck drivers, but also to recreational drivers/tourists driving campers/vans. Finally, in a crash involving young drivers aged 24 or less, it is less likely to have at least an unfamiliar driver involved. This suggests that young drivers could be involved in less crashes then the other drivers when far from home (as confirmed by the chi-square test). They could be more prudent/alert while driving on unknown roads, perhaps because of their minor degree of familiarity with the act of driving. This could mean that the general greater crash risk for young drivers (Jonah, 1986) is more present at familiar than unfamiliar routes.

4.3 In-Depth Analysis of Accident Types and Dynamics (Third Level)

The accidents and the involved vehicles/drivers divided into the familiarity categories are shown in Figures 7 and 8, together with the proportions of the associated accident types and dynamics.

(Figure 7 here)

(Figure 8 here)

4.3.1 Drivers' Familiarity and Accident Types

A chi-square test of independence was performed between accident type and familiarity (interactions excluded). There is a statistically significant association between accident type and familiarity, $\chi^2(4) = 13.836$, $p = 0.008$. However, the association is small (Cohen, 1988), Cramer's V = .136.

As it emerges from the first test, basing on adjusted residuals highlighted in Table 9, more familiar drivers were involved in rear-end or angle accidents and less familiar drivers were involved in head-on accidents than the expected. More transition drivers were involved in head-on accidents than the expected. No clear indications can be obtained for unfamiliar drivers, not confirming the influence found from logistic regression, where the interactions between different categories were not excluded.

Furthermore, the chi-square test was repeated between accident type and familiarity, considering only the interactions between the different familiarity categories of drivers. In this case, no statistically significant association between accident type and familiarity was found, $\chi2(2) = 0.170$, $p = 0.918$.

(Table 9 here)

In most of the accidents (69 %), no interactions between different categories of drivers' familarity occurred. This can be related to the high share of run-off crashes in the sample, typically involving only one vehicle (44 %). Among these "no interactions" accidents, familiar drivers were more associated to rear-end/angle accidents and less associated to head-on accidents than the expected. No differences between the diverse types of accidents were found for the "interactions" accidents (in which familiar and unfamiliar drivers were both involved).

The finding that familiar drivers are more associated to rear-end crashes than run-off crashes could be explained by their proneness to more dangerous behaviors due to their possible over-confidence with the road environment (see also Intini et al., 2017). They could be disposed to wait until the last possible moment to brake before turning (for example into a driveway) but also disposed to greater speeds and closer car following, behaviours possibly related to rear-end crashes (both striking or being struck). The head-on accidents were less associated to the involvement of only familiar drivers. They can be equated with run-off crashes because they are generally caused by a vehicle initially losing control and eventually invading opposite lane. The familiarity with a given road seems to prevent this type of accident due to possible errors in speed and steering, thanks to the knowledge of the road features. This is coherent with results from Wilks et al. (1999) based on a crash database analysis: international drivers are over-represented in head-on collisions with respect to the local ones. Anyway, in this case, no specific association was found with unfamiliar drivers for head-on crashes (but with transition drivers).

For what concerns the interactions between the familiarity categories, one should expect that the interactions familiar/unfamiliar drivers would lead to safety problems, as explained in Section 1. However, in this case, no particular association was found between the different interactions between the various familiarity categories and the accident type. Nevertheless, it must be stated that in this work, only two-lane rural roads were analyzed. In order to observe clearer effects of familiarity on the road crashes, multi-lane highways should be probably analyzed with techniques similar to those used here. In fact, on multi-lane roads, the number of interactions between vehicles could be greater due to higher traffic volumes and more lanes. Moreover, on those roads, differences between recreational drivers and commuters were previously highlighted (see TRB, 2000).

4.3.2 Drivers' Familiarity and Accident Dynamics

A chi-square test of independence was performed between accident dynamics and familiarity. There is a statistically significant association between accident dynamics and familiarity, $\chi 2(6) = 14.632$, $p = 0.023$. However, the association is small (Cohen, 1998), Cramer's V = .097.

As it emerges from the test, based on adjusted residuals highlighted in Table 10, more familiar drivers were involved in accidents being stationary after braking or for turning than the expected. Less unfamiliar drivers lost control in the accident than the expected. Less transition drivers were involved in accidents being stationary after braking or for turning than the expected.

(Table 10 here)

Unfamiliar drivers seem to be less prone to lose control of their vehicles during the accidents. Actually, most of the run-off road crashes are likely to be caused by losing control. Unfamiliar drivers were more involved in the run-off accidents than the expected, even if this tendency cannot be considered significant (Table 9) and it should be taken only as an indication. However, by merging these findings, it could be suggested that in other crash types different from the run-off, unfamiliar drivers can be less prone to being out of control. Anyway, it is difficult to compare this with previous studies. In fact, the latter were more focused on foreigners than on more general unfamiliarity (Wilks et al., 1999; Yannis et al., 2007; Kim et al., 2012).

Familiar drivers were more often involved in accidents while stationary after braking or before turning. This is coherent with previous results from Baldock et al. (2005) even if based on a small sample of interviews: drivers traveling with a daily frequency on a given road were more likely struck on that road, in comparison with daily drivers striking. In this study, familiar drivers were associated to rear-end crashes when excluding the interactions between different categories of familiarity (Table 9). However, in light of this result about dynamics, since most of the vehicles being stationary during an accident are normally involved in rear-end accidents, one should conclude that the familiar drivers are often the ones being struck in rear-end crashes rather than striking.

5. CONCLUSIONS

The relationships between accidents and drivers' familiarity were inquired in detail. They were searched by using an integrated approach, composed of three levels of analysis, from a macro-analysis to more detailed levels considering specific accident, vehicle and person-related variables. Drivers' familiarity was defined based on the distance between the crash site and the drivers' residence.

The macro-scale analyses did not reveal the hypothesized relationships. In fact, a macro indicator as the accident rate could be inappropriate to reveal the specific detailed relationship between the familiarity of drivers and the occurrence of accidents, because of the several other factors hidden in the accident rate besides the drivers' familiarity. However, the method used for the macro-analysis considering summer seasonal traffic variations and different seasons could be verified in further studies. In fact, as expected, the effects of traffic seasonal variations and different periods of the year were further revealed as influential by detailed analyses at a more disaggregate level. It was noted how the traffic summer seasonal variation was strongly positively related with the average distance from residence of drivers involved in the accidents. This was confirmed by the significance of the variables "traffic seasonal variation" and "season" in the logistic regression model for unfamiliarity. Therefore, it can be clearly suggested that crashes to unfamiliar drivers cluster at sites showing high summer traffic variation (ranging from 1.4 to almost 2 times the AADT) and are more frequent in summer months. The initial hypotheses were the following: 1) familiarity and unfamiliarity can be both risk factors for accident occurring; 2) familiar/unfamiliar drivers can be differently related to human, environmental and road factors related to crashes; 3) familiar and unfamiliar drivers can show different relative involvements in different types of crashes. Some important results were highlighted from the statistical analyses, revealing possible relationships. The initial hypotheses can be confirmed as a whole, even with some limitations, as summarized below.

- Familiarity was confirmed as a risk factor. This may due to the tendency of familiar drivers to distraction, inattention and possible more dangerous behaviours due to the over-confidence. This was argued considering that it was more likely to find familiar drivers involved in accidents at road sections with speed limits lower than 80 km/h and while commuting. Considering the accidents involving only familiar drivers, without interactions with other drivers, the rear-end/angle crashes are significantly over-represented among the crash types. Moreover, by cross-checking data of types and dynamics, it was deduced that they are more likely to be struck than strike in rear-end accidents, by considering the entire sample of accidents (with or without interactions). Among the other variables, it seems that familiar drivers are safer when traveling for work purposes.
- Unfamiliarity was assumed as a risk factor considering their possible weakness due to the ignorance of the road layout and because of the possible negative interactions with other drivers not unfamiliar. In this case, the variables resulted significant from the analysis did not reveal these supposed greater risks. Further results about the accident types and dynamics do not definitely clarify those relationships. In fact, firstly the rear-end/angle accident was identified as the crash type in which an unfamiliar driver can be more likely found. This seemed to indicate that interactions with other drivers could be the most recurrent problem. However, when specificately further looking at interactions and no-interactions accidents, no relevant findings were highlighted, even if it could have been expected by the significance of the variable “traffic volume”. Among the other variables considered, a relationship was noted between the unfamiliar drivers and the heavy vehicles involved (likely both due to the

professional long-distance drivers and to the tourists driving vehicles different from cars). Moreover, young drivers (under 24) seem to drive safer than others when far from home.

Findings from this study can suggest interesting practical remarks for road safety engineering. Particular attention should be paid to sites with traffic seasonal variation greater than 40 % in summer. Measures allowing the safe traveling in particular for the drivers unfamiliar with the road should be taken. However, traditional measures such as those related to pavement or lighting (see also Yannis et al., 2007) could have not influence on the occurrence of crashes to unfamiliar drivers as deduced from the devoted regression model. Some solutions such as particular road signs or variable message signs (VMS) (see e.g. Erke et al., 2007) could help unfamiliar drivers in the safe travel, as well as a design in compliance with the self-explaining theory (Theeuwes and Godthelp, 1995; Charlton et al., 2010; Mackie et al., 2013). For what concerns familiar drivers, their possible vulnerability in sub-urban contexts (as emerged from their association with high volumes sections and driveways/minor intersections) and at sections with lower speed limits was highlighted. In this case, considering the proneness of familiar drivers to distraction but also to more dangerous behaviours, the necessity for efficient traffic enforcement measures could arise (see also Martens and Fox, 2007). However, the drivers' perception of the level of enforcement (Ryeng, 2012) and the possible adaptation to it (Montella et al., 2015) should be taken into account.

It has to be noted that his study is solely based on Norwegian two-lane two-way rural road segments. For the aims of this research, it is necessary to enlarge the study to multi-lane highways and to other countries. In fact, on those other road types, traffic volumes are considerably higher and the number of interactions between drivers is surely greater also due to the increased number of lanes. In these conditions, perhaps near touristic places, allowing to observe consistent summer traffic variations, the specific matter of interactions between familiar and unfamiliar drivers could be better addressed. Traffic volumes were considered in this study, in order to control crash data for exposure measures. However, it should be noticed that average traffic volumes were used, rather than the volumes at the moment of the accident. Therefore, they were used mainly to have indications about the specific road site in terms of annual volumes and seasonal variations. Future studies could benefit from real-time data at the moment of the accident, better if together with information from plates recognition, useful for inquiring into shares of unfamiliar drivers in the traffic flow. Moreover, the study is only based on Norwegian data. It is important to highlight again that familiarity is a subjective matter, concerning the human brain, and it can affect different persons in different ways. Therefore, as cultural and population variables can affect driver behavior, also the effects of familiarity can vary in different countries. It could be useful to repeat a similar study for other countries. In this work, familiarity was measured based on distance of drivers from the residence. This is an indirect measure of the drivers' familiarity, which cannot guarrantee the correct classification of drivers into classes based on a personal feature, rather difficult to be directly measured instead. Therefore, it could be affected by some errors even if it was believed to be reasonable for the aims of this study.

6. ACKNOWLEDGMENTS

The Norwegian University of Science and Technology, the Technical University of Bari and the European Union (Erasmus+ Project) are acknowledged for their funding and support which have made possible the stay of the first author in Norway, essential for the realization of this article. The authors would also like to acknowledge the Norwegian Public Road Administration for providing the datasets.

7. REFERENCES

Agresti, A. (2013). Categorical data analysis (3rd Edition). John Wiley & Sons, Hoboken.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention, 34(6), 729-741.

American Association of State Highway and Transportation Officials (AASHTO). Highway safety manual (1st ed.), Washington, DC., 2010.

American Association of State Highway and Transportation Officials (AASHTO). Commuting in America 2013. The National Report on Commuting Patterns and Trends. Washington, DC., 2013.

Baldock, M. R. J., Long, A. D., Lindsay, V. L. A., & McLean, J. (2005). Rear end crashes. Centre for Automotive Safety Research.

Beijer, D., Smiley, A., & Eizenman, M. (2004). Observed driver glance behavior at roadside advertising signs. Transportation Research Record: Journal of the Transportation Research Board, (1899), 96-103.

Bertola, M. A., Balk, S. A., & Shurbutt, J. (2012). Evaluating driver performance on rural two-lane horizontal curved roadways using a driving simulator (No. FHWA-HRT-12-073).

Blatt, J., & Furman, S. M. (1998). Residence location of drivers involved in fatal crashes. Accident Analysis & Prevention, 30(6), 705-711.

Brown, J., Fitzharris, M., Baldock, M., Albanese, B., Meredith, L., Whyte, T., & Oomens, M. (2015). Motorcycle In-depth Crash Study (No. AP-R489-15).

Campbell, J. L. et al. (2012). Human factors guidelines for road systems (Vol. 600). Transportation Research Board.

Charlton, S. G. (2007). The role of attention in horizontal curves: A comparison of advance warning, delineation, and road marking treatments. Accident Analysis & Prevention, 39(5), 873-885.

Charlton, S. G., Mackie, H. W., Baas, P. H., Hay, K., Menezes, M., & Dixon, C. (2010). Using endemic road features to create self-explaining roads and reduce vehicle speeds. Accident Analysis & Prevention, 42(6), 1989-1998.

Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge Discovery and Data mining, 263-268.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2^{nd} Ed.). New York, Psychology Press.

Colonna, P. (2002). Proposal for a safety function for evaluating the road efficiency level. In: Proceedings of the Conference on Traffic and Transportation Studies, ICTTS, pp. 1055–1062.

Colonna, P. (2009). Mobility and Transport for our tomorrow roads. European Roads Review (14), 44-53.

Colonna, P., Berloco, N. (2011). External and internal risk of the user in road safety and the necessity for a control process. XXIV PIARC World Road Congress, Mexico City.

Colonna, P., Berloco, N., Intini, P., & Ranieri, V. (2015). Route Familiarity in Road Safety: Speed Choice and Risk Perception Based on a On-Road Study. In Transportation Research Board 94th Annual Meeting (No. 15-2651).

Colonna, P., Intini, P., Berloco, N., & Ranieri, V. (2016). The influence of memory on driving behavior: How route familiarity is related to speed choice. Safety science, 82, 456-468.

Colonna, P., Intini, P., Berloco, N., Perruccio, A. & Ranieri, V. (2016). Repeated Measurements of Lateral Position and Speed at Horizontal Curves on a Very-Low Volume Rural Road. In Transportation Research Board 95th Annual Meeting (No. 16-3426).

Donaldson, A. E., Cook, L. J., Hutchings, C. B., & Dean, J. M. (2006). Crossing county lines: The impact of crash location and driver's residence on motor vehicle crash fatality. Accident Analysis & Prevention, 38(4), 723-727.

Elvik, R. (2006). Laws of accident causation. Accident Analysis & Prevention,38(4), 742-747.

Erke, A., Sagberg, F., & Hagman, R. (2007). Effects of route guidance variable message signs (VMS) on driver behaviour. Transportation Research Part F: Traffic Psychology and Behaviour, 10(6), 447-457.

Garber, N.J., Gadiraju, R., 1989. Factors affecting speed variance and its influence on accidents. 1989-01-01 1213. Transportation Research Record, Washington D.C.

Harrell, F. (2015). Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer International Publishing.

Hjorthol, R., Engerbretsen, Ø., Uteng, T., P. (2014). Den nasjonale reisevaneundersøkelsen 2013/2014 – nøkkelrapport. Transportøkonomisk institutt. Norway.

Intini, P. (2014). Changes in speed behavior due to acquired road familiarity. A comparison between Italy and Norway. Master Thesis. Norwegian University of Science and Technology.

Intini, P., Colonna, P., Berloco, N., Ranieri, V. (2016). Measuring trade-offs between risk and travel time based on experimental speed data. Applied Human Factors and Ergonomics. Proceedings of the 7th International Conference on Applied Human Factors and Ergonomics (AHFE 2016) and the Affiliated Conferences, July 27-31, Orlando, Florida, USA.

Intini, P., Colonna, P., Berloco, N., Ranieri, V., Ryeng, E. (2017). The relationships between familiarity and road accidents: some case studies. In: Transport Infrastructure and Systems: Proceedings of the AIIT International Congress on Transport Infrastructure and Systems (Rome, Italy, 10-12 April 2017). Dell'Acqua, G., and Wegman, F. (Eds.). CRC Press.

Jonah, B. A. (1986). Accident risk and risk-taking behaviour among young drivers. Accident Analysis & Prevention, 18(4), 255-271.

Kim, K., Brunner, I. M., Yanashita, E., Uyeno, R. (2012). Comparative Assessment of Visitor and Resident Crash Risk in Hawaii. Transportation Research Board 91st Annual Meeting (12-2854).

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.

Laerd Statistics (2015). Statistical tutorials and software guides. Retrieved from https://statistics.laerd.com/

Lehmann, E., L. (2006). Nonparametrics: Statistical Methods based on Ranks. Springer, NY.

Litman, T. (2003). Measuring transportation: traffic, mobility and accessibility. ITE Journal, 73(10), 28-32.

Liu, C., & Ye, T. J. (2011). Run-off-road crashes: An on-scene perspective (No. HS-811 500).

Lord, D., Manar, A., & Vizioli, A. (2005). Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accident Analysis & Prevention, 37(1), 185-199.

Lotan, T. (1997). Effects of familiarity on route choice behavior in the presence of information. Transportation Research Part C: Emerging Technologies, 5(3), 225-243.

Mackie, H. W., Charlton, S. G., Baas, P. H., & Villasenor, P. C. (2013). Road user behaviour changes following a self-explaining roads intervention. Accident Analysis & Prevention, 50, 742-750.

Martens, M. H., Fox, M. R. J. Do familiarity and expectations change perception? Drivers' glances and response to changes. Transportation Research Part F: Traffic Psychology and Behavior, Vol. 10, 2007, pp. 476–492.

Milliken, J. G., Council, F. M., Gainer, T. W., Garber, N. J., Gebbie, K. M., Hall, J. W., et al. (1998). Managing speed: review of current practice for setting and enforcing speed limits. Transportation Research Board, Special Report 254. Washington DC.

Montella, A., Andreassen, D., Tarko, A., Turner, S., Mauriello, F., Imbriani, L., Romero, M. (2013). Crash databases in Australasia, the European Union, and the United States: review and prospects for improvement. Transportation Research Record: Journal of the Transportation Research Board, (2386), 128-136.

Montella, A., Imbriani, L. L., Marzano, V., & Mauriello, F. (2015). Effects on speed and safety of point-to-point speed enforcement systems: Evaluation on the urban motorway A56 Tangenziale di Napoli. Accident Analysis & Prevention, 75, 164-178.

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

Nilsson, G. (2004). Traffic safety dimensions and the power model to describe the effect of speed on safety, Doctoral Dissertation, Lund University, Sweden.

Noland B. From theory to practice in road safety policy: Understanding risk versus mobility. Research in Transportation Economics, Vol. 43, 2013, pp. 71-84.

Oxford dictionary of English. Oxford University Press.

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., Coppola, G., Geyer, M. A., Glanzman, D. L., Marsland, S., McSweeney, F. K., Wilson, D. A., Wu C., Thompson, R. F. Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. Neurobiology of learning and memory, 92(2), 2009, pp. 135-138.

Rasmussen, J. Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering, 1986, Elsevier, New York.

Regan, M. A., Lee, J. D. & Young, K. L. (Eds.). (2008). Driver distraction: Theory, effects, and mitigation. CRC Press.

Rosenbloom, T., Perlman, A., Shahar, A. Women drivers' behavior in well-known versus less familiar locations. Journal of Safety Research, vol. 38, Issue 3, 2007, pp. 283-288.

Ryeng, E. (2012). The effect of sanctions and police enforcement on drivers' choice of speed. Accident Analysis & Prevention, 45, 446-454.

Sandin, J. (2009). An analysis of common patterns in aggregated causation charts from intersection crashes. Accident Analysis & Prevention, 41(3), 624-632.

Schepers, P., & den Brinker, B. (2011). What do cyclists need to see to avoid single-bicycle crashes? Ergonomics, 54(4), 315-327.

Singh, S. (2015). Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. (Traffic Safety Facts Crash, Stats. Report No. DOT HS 812 115). Washington, DC: National Highway Traffic Safety Administration.

Sreejesh, S., Mohapatra, S., & Anusree, M. R. (2014). Binary Logistic Regression. In Business Research Methods (pp. 245-258). Springer International Publishing.

Staubach, M. (2009). Factors correlated with traffic accidents as a basis for evaluating Advanced Driver Assistance Systems. Accident Analysis & Prevention, 41(5), 1025-1033.

Theeuwes, J., & Godthelp, H. (1995). Self-explaining roads. Safety science, 19(2), 217-225.

Thrane, C. (2015). Examining tourists' long-distance transportation mode choices using a Multinomial Logit regression model. Tourism Management Perspectives, 15, 115-121.

Transportation Research Board. Highway Capacity Manual. Washington, D.C., 2000.

Treat, J. R., Tumbas, N. S., McDonald, S. T., Shinar, D., Hume, R. D., Mayer, R. E., Stansifer, R. L. & Castellan, N. J. (1979). Tri-level study of the causes of traffic accidents: final report. Executive summary. (Report DOT HS-034-3-535). Washington, DC: NHTSA.

Wang, H., Li, Z., Hurwitz, D., & Shi, J. (2015). Parametric modeling of the heteroscedastic traffic speed variance from loop detector data. Journal of advanced transportation, 49(2), 279-296.

Wilks, J., Watson, B. C., Johnston, K. L., & Hansen, J. A. (1999). International drivers in unfamiliar surroundings: the problem of disorientation. Travel Medicine International, 17(6), 162-167.

Yanko, M. R., Spalek, T. M. Route familiarity breeds inattention: A driving simulator study. Accident Analysis and Prevention, Vol. 57, 2013, pp. 80-86.

Yannis, G., Golias, J., & Papadimitriou, E. (2007). Accident risk of foreign drivers in various road environments. Journal of safety research, 38(4), 471-480.

Young, K. L., & Salmon, P. M. (2012). Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods. Safety science, 50(2), 165-174.

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 44(3), 365-375.

Zhu, X., & Srinivasan, S. (2011). Modeling occupant-level injury severity: An application to large-truck crashes. Accident Analysis & Prevention, 43(4), 1427-1437.

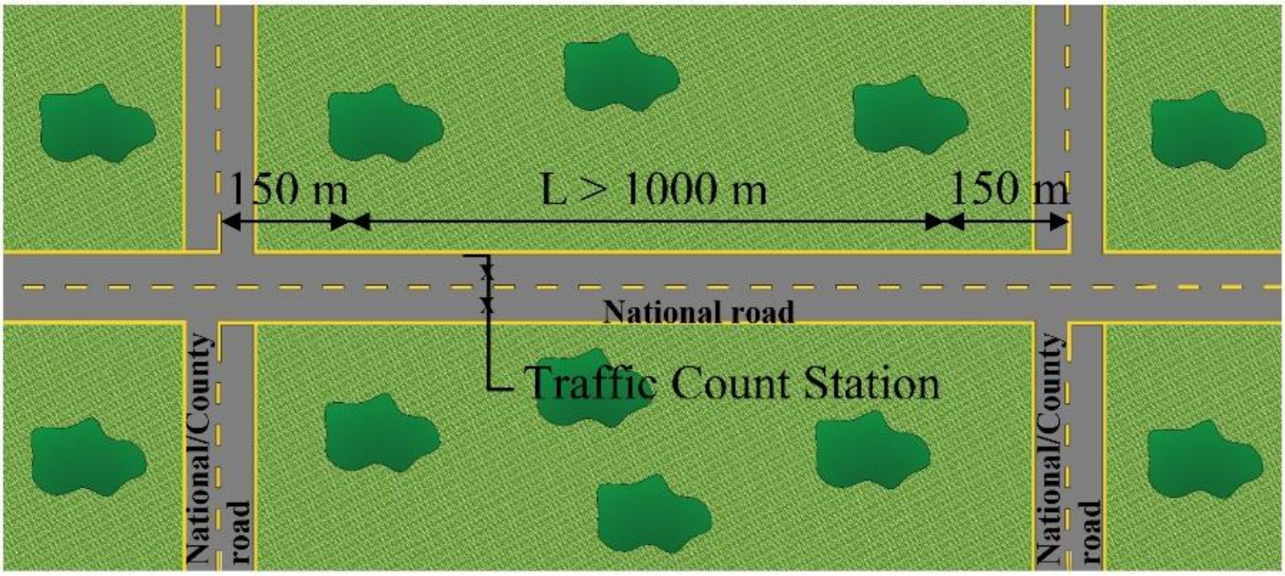

Figure 1 – Individuation of the road sites.

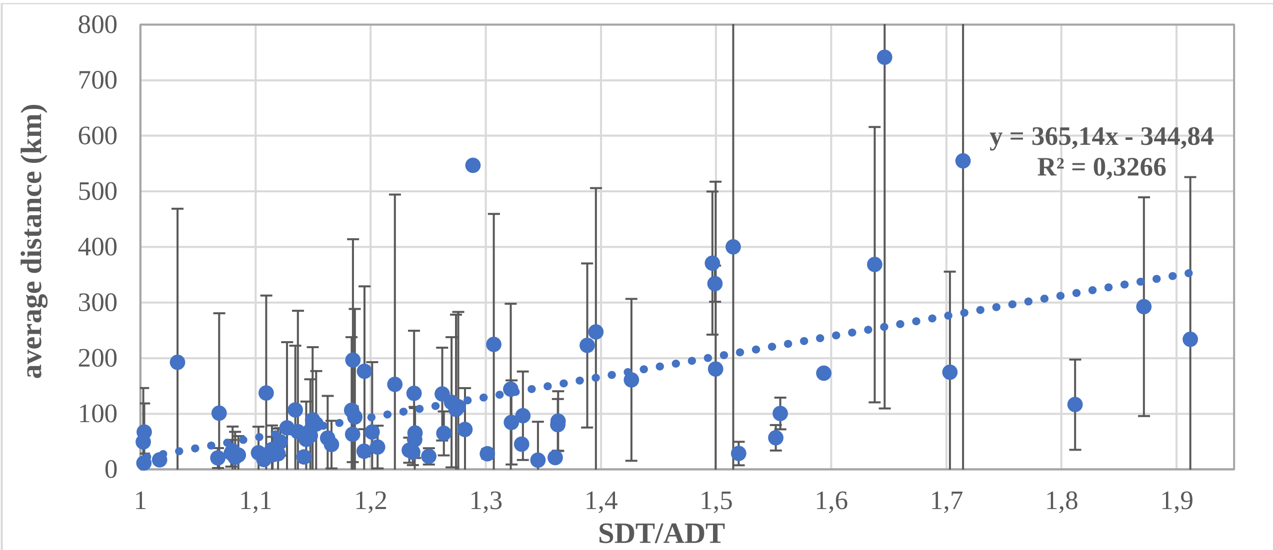

Figure 2 – Relationship between the SDT/AADT ratio and the average distances of drivers from residence for each road site inquired.

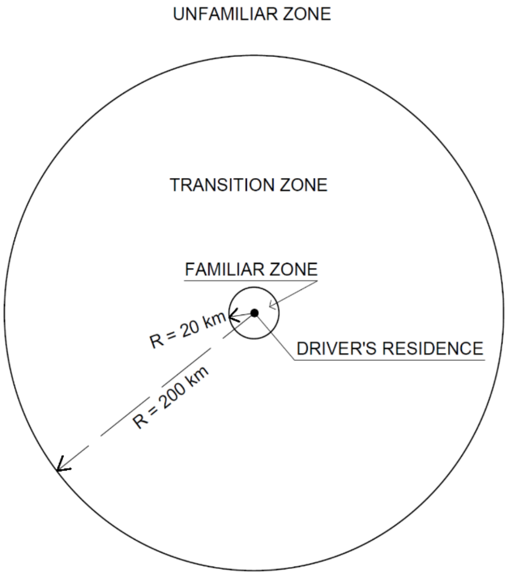

Figure 3 – Scheme of the familiarity zones.

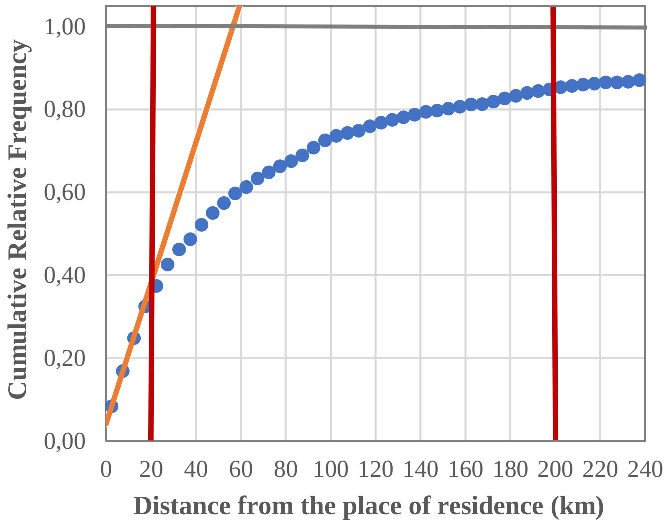

Figure 4 – Particular of the Cumulative Relative Frequency curve of Distances.

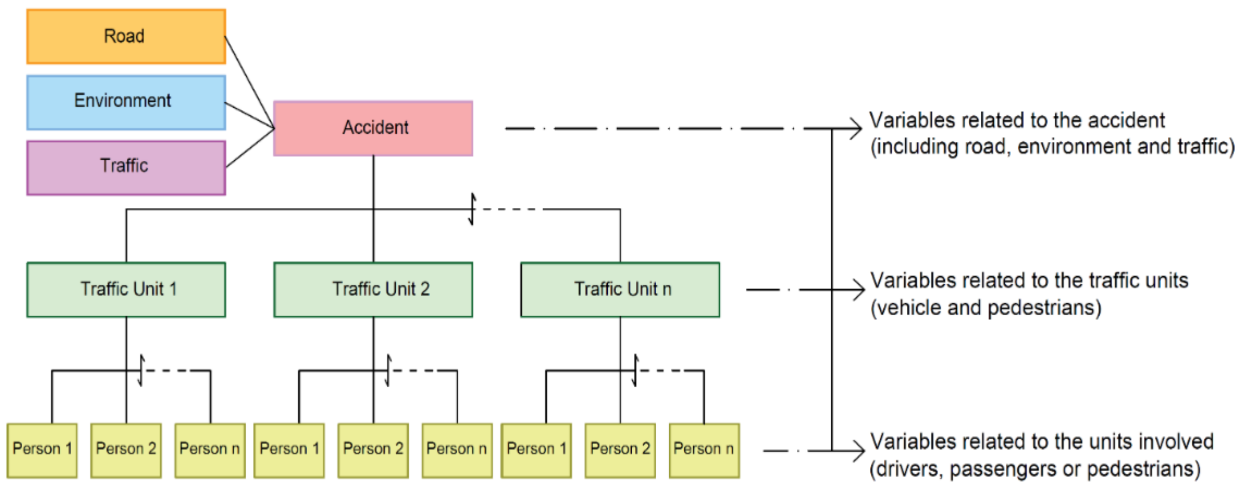

Figure 5 – Levels of the information contained in the accident database.

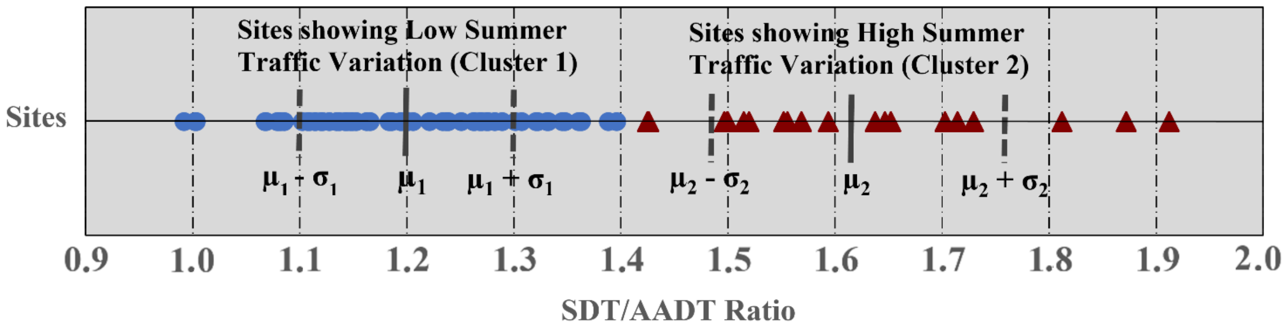

Figure 6 –SDT/AADT ratios of the road sites belonging to the two clusters (μ = mean, σ = st. dev.).

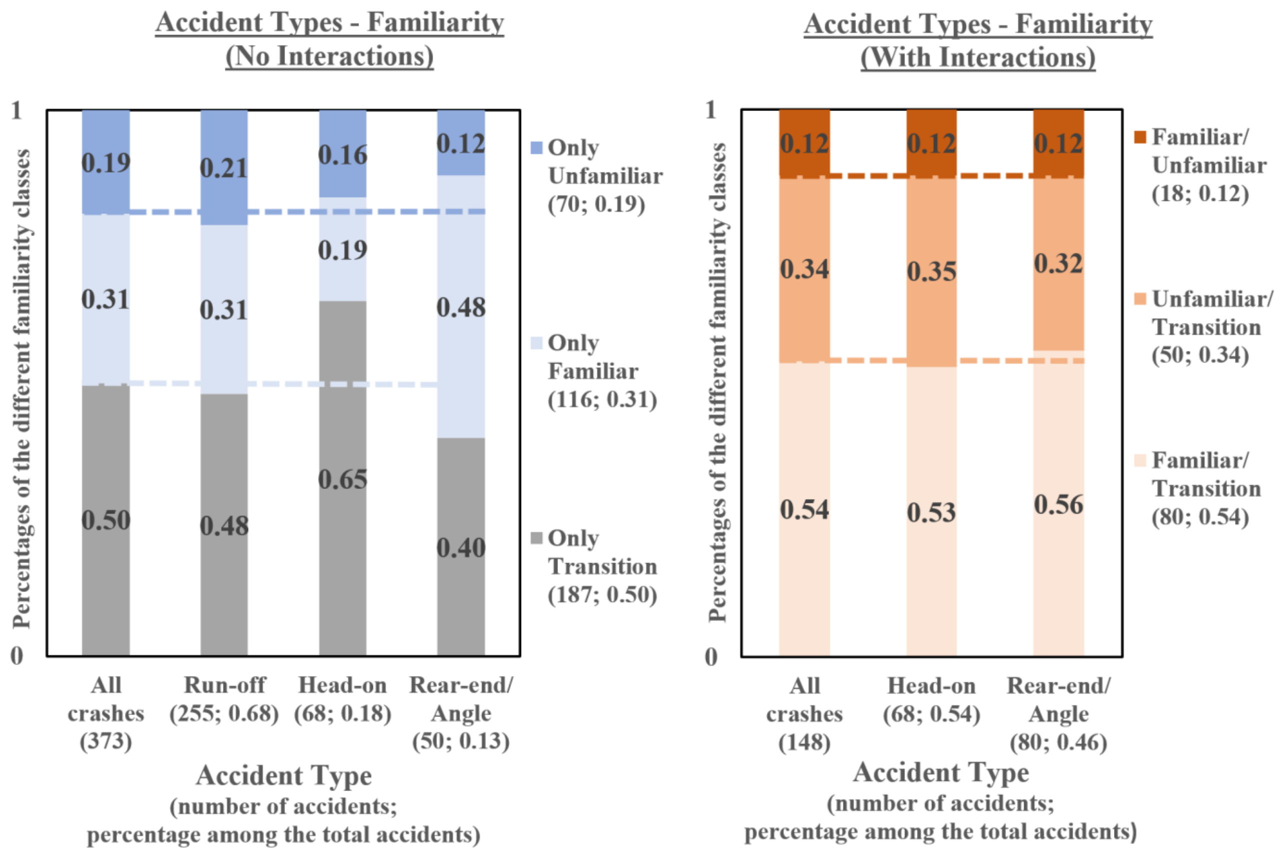

Figure 7 – Percentage distributions of the accidents in the different types and familiarity categories differentiating between crashes without (left) and with (right) interactions between the categories.

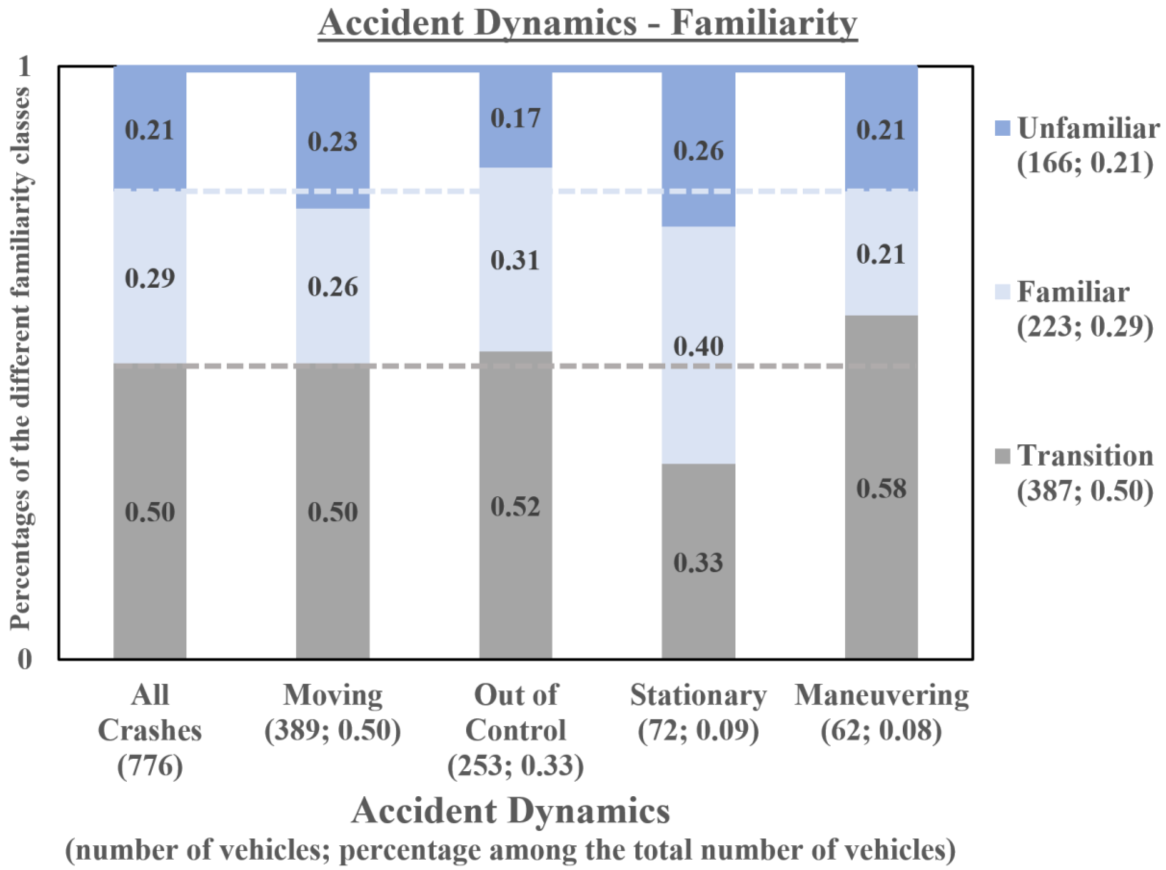

Figure 8 – Percentage distributions of the accidents in the dynamics type and familiarity categories.

Table 1 - Summary of previous findings about relationships between accidents and route familiarity.

Method	Authors	Users/ (Area)	Road Familiarity scale	Focus	Findings regarding familiarity
Post-crash survey	Liu and Ye [13]	Drivers (USA)	-Familiar (daily, weekly or monthly driving) -Unfamiliar (rarely or first time driving)	Type of accident: run-off road	Drivers travelling on familiar roadways are more likely to be involved in run-off road crashes (64 % of single-vehicle crashes occurred to familiar drivers were run-off road, compared to the 54 % of the unfamiliar ones)
Post-crash survey	Baldock et al. [41]	Drivers (South Australia)	-Drive daily -Familiar -Not well known -No knowledge	Type of accident: rear-end crashes	Over-representation of drivers driving daily the road being struck (48 %) in respect of daily drivers striking (30 %) even if based on a small sample of 38 interviews
Post-crash survey	Schepers and den Brinker [42]	Cyclists (Netherlands)	-Familiar with the crash scene -Not familiar with the crash scene	Type of accident: Single-bicycle	Over-representation of drivers unfamiliar with the crash scene (even if not statistically significant), especially for cyclists colliding with bollards or running-off the road in bends
Post-crash survey	Brown et al. [43]	Bikers (Australia)	-Drive daily -2/3 times/week -once a week -once a month -rarely -first time in area	Accident occurrence	Being familiar with the route identified as a crash factor by using a case-control analysis (daily riders were found to be more than seven times more likely to be in the crash sample than the control), unfamiliarity with the route identified as a contributory factor in a small number of crashes by using Haddon Matrix in a qualitative analysis
Crash data analysis	Blatt and Furman [44]	Drivers (USA)	-Rural and small town resident -Suburban, urban, second city residents	Accident occurrence /severity	Over-involvement of rural and small town residents in fatal crashes on rural roads and over-involvement of urban residents in urban crashes
Crash data analysis	Donaldson et al. [45]	Drivers (Utah, USA)	-Urban resident -Rural resident	Accident occurrence /severity	Although the majority of rural crash fatalities involve rural drivers, urban drivers and their passengers have the highest risk of fatality when involved in rural compared to urban crashes, controlling for occupant, behavioral, road and crash characteristics
Crash data analysis	Kim et al. [46]	Drivers (Hawaii, USA)	-Visitor -Resident	Accident occurrence /Type of accident	Visitors are much more likely to be at fault when involved in crashes in Hawaii. At-fault visitors are more likely to cause accidents for improper manoeuver or wrong way and to generate crashes on highways, while these effects are much smaller for at-fault residents based on odds-ratio estimate from logistic regression model

Crash data analysis	Wilks et al. [47]	Drivers (Australia)	-International -Australian	Type of accidents	Over-representation of international drivers in angle, sideswipe and head-on collisions, over-representation of local drivers in fixed-object, pedestrians, parked vehicles and animals collisions. International drivers may be disoriented by different driving conditions and road rules (as the right-side driving)
Crash data analysis	Yannis et al. [48]	Drivers (Greece)	-Greek -Albanian -EU	Accident occurrence	Increased accident fault risk for drivers of foreign nationality, especially for EU drivers. Road infrastructure elements connected with generally higher accident risk of foreign drivers are junctions, while area type and lighting conditions do not seem influential
Crash data analysis	Zhu and Srinivasan [49]	Drivers/ Truck drivers (USA)	-Drive daily -Weekly -Several times per month -Once per month -Rarely -First time	Accident severity	Drivers involved in crashes with trucks are less likely to receive severe injuries on a roadway they are driving for the first time.

Table 2 - Details of traffic and accidents data related to the road sites in the dataset.

	Total				E6				E39			
	Nr./ Mean	St. Dev.	Max. Value	Min. Value	Nr./ Mean	St. Dev.	Max. Value	Min. Value	Nr./ Mean	St. Dev.	Max.	Min.
Sites	84	-	-	-	37	-	-	-	47	-	-	-
Length (m)	6144	5841	35604	1030	6613	7092	35604	1030	5775	-	24225	1110
AADT (veh/day)	5626	4738	18706	544	4745	4798	15807	544	6320	4530	18706	937
SDT (veh/day)	6712	5117	21856	721	6064	5378	18717	721	7222	4758	21856	1277
SDT/AADT	1.30	0.21	1.91	0.99	1.44	0.22	1.91	1.11	1.19	0.11	1.36	0.99
Accidents	633	-	-	-	235	-	-	-	398	-	-	-
Vehicles	1091	-	-	-	406	-	-	-	685	-	-	
Acc./Site	7.5	6.3	35.0	0.0	6.4	6.7	35.0	0.0	8.5	5.8	28.0	0
Acc. Rate (acc./MVKT)	0.103	0.093	0.728	0.000	0.113	0.127	0.728	0.000	0.096	0.053	0.241	0.000

Note: Traffic values of AADT and SDT are averaged on the ten years period 2005-2014. Accidents are referred to the same period. MVKT = MillionVehiclesKilometersTraveled. Accident rates were computed for each site and then means, max., min. and standard deviations are shown here.

Table 3 - Description of the study variables.

Variable	Modalities[1] 1	2	3	4
Type of Variables: Accident-related variables (including Road, Environment and Traffic)				
Traffic Volume (AADT)[2]	Low (< 5,000)	Medium (5,000 – 11,000)	High (> 11,000)	
Traffic Seasonal Variation (SDT/AADT)	Low (≤ 1.40)	High (> 1.40)		
Season[3]	Summer (Jun. – Jul. – Aug.)	Autumn (Sept. – Oct. – Nov.)	Winter (Dec. – Jan. – Feb.)	Spring (Mar. – Apr. - May)
Time of the Day	Morning (6 a. m. – 12 p. m.)	Afternoon (12 p. m. - 6 p. m.)	Evening (6 p. m. – 12 a. m.)	Night (12 a. m. – 6 a.m.)
Week Period	Weekday (Monday to Friday)	Weekend (Saturday – Sunday)		
Accident Type	Run-off	Rear-end/Angle	Head-on	
Section Type	Road segment	Tunnel	Driveway/ Minor intersection	
Road Surface	Dry	Wet	Snowy, icy or otherwise slippery	
Visibility/Weather	Good Visibility, Good Weather	Good Visibility/ Rain	Bad Visibility	
Lighting	Daylight	Dark with road lights/ Twilight	Dark without road lights	
Presence of Additional Lanes	No	Yes		
Road Width	7 - 9	< 7	≥ 9	
Speed Limit[4]	≥ 80 km/h	< 80 km/h		
Type of Variables: Unit-related variables				
Heavy Vehicles Involved[5]	No	At least one		
Motorcycles Involved	No	At least one		
Pedestrians/Cyclists Involved	No	At least one		
Vehicle Age (≥15 yrs)	No	At least one		
Type of Variables: Person-related variables				
Gender: Woman[6]	No	At least one		
Young persons involved (<24 yrs)	No	At least one		
Old persons involved (≥65 yrs)	No	At least one		
Pleasure Travelers[7]	No	At least one		
Commuting Travelers[7] (To/From work or school)	No	At least one		
Work Travelers[7]	No	At least one		
Under Influence[7]	No	At least one		

[1]The first modality of each variable was considered as modality of reference for the analysis.
[2]AADT clusters for the road sites inquired were defined by using the same strategy of cluster analysis described in 3.1.1.
[3]Seasons were defined in 3-months sets to be coherent with the definition of SDT (Summer Daily Traffic) measured from June to August, see 2.1.
[4]Speed limits were clustered into these two groups since most of the road sites inquired had posted speed of 80 km/h (69% of the accidents are related to speed limit set to 80 km/h, 22 % to 70 km/h and accidents of the remaining 9 % are related to limits of 50, 60 and 90 km/h).
[5]The gender variable was so modeled since the most recurring condition is the absence of women involved as drivers/units (about 62% of cases).
[6]In the variable modeling, “heavy vehicles” include trucks, light trucks, campers, vans, buses, vehicles with trailer.
[7]For these variables, an additional modality (“No information”) was added in case of missing data for all vehicles/drivers involved in crashes, in order to control for the high number of missing data about travel purposes and driving under influence,

Table 4 - Summary of the methods used for the three levels of analysis.

Level of Analysis	Objective	Measures	Data Analysis Techniques
First	✓Finding differences between accident rates at road sites considering different seasons and different summer traffic variation rates	✓Distance of drivers from residence ✓SDT/AADT ✓Accident rates ✓Road sites showing high/low summer traffic variation	✓Cluster Analysis ✓Mann-Whitney U Test
Second	✓Finding relationships between familiarity of drivers with road sites and accident characteristics	✓Familiar and unfamiliar drivers	✓Logistic regression ✓Chi-Square Test
Third	✓Finding specific associations between drivers' familiarity and accident types and dynamics, considering the interactions between different drivers	✓Accident Types ✓Familiarity categories of Accidents ✓Accident Dynamics ✓Familiarity categories of Vehicles ✓Distance of drivers from residence	✓Chi-Square Test ✓Kruskal-Wallis H Test

Table 5 - Descriptive statistics about AADT [vehicles/day] and SDT/AADT ratio for the two clusters.

Cluster	Measure	No. of items	Mean	Min. value	Max. value	St. Dev.
1 (Low SDT/AADT ratio)	SDT/AADT	64	1.20	0.99	1.40	0.10
	AADT		6,836	546	18,706	4,812
2 (High SDT/AADT ratio)	SDT/AADT	20	1.62	1.42	1.91	0.14
	AADT		1,911	544	6,353	1,492

Table 6 - Descriptive statistics about accident rates [accidents/MVKT] for different combinations of the conditions at the road sites.

Combination of Traffic Variation and Season	No. of items	Mean	Median	Min. value	Max. value	St. Dev.
Low Summer Traffic Variation - Summer	64	0.076	0.065	0.000	0.390	0.078
High Summer Traffic Variation - Summer	20	0.106	0.065	0.000	0.459	0.124
High Summer Traffic Variation - Other Seasons	20	0.182	0.152	0.000	0.951	0.222

Table 7 - Results of Logistic Regression (Outcome Variable: Familiarity).

	β	S. E.	Wald	df	p	OR	CI for the OR (95 %)	
							Lower	Upper
Traffic Volume			6.099	2	.047			
Medium: 5,000 – 11,000	.186	.245	.574	1	.449	1.204	.744	1.949
High: > 11,000	.831	.337	6.075	1	**.014**	2.295	1.185	4.444
Reference Category – Low: < 5,000								
Traffic Seasonal Variation (High: SDT/AADT > 1.4)	-1.314	.329	15.925	1	**<.001**	.269	.141	.512
Season			12.336	3	.006			
Autumn	.887	.296	8.999	1	**.003**	2.429	1.360	4.336
Winter	1.002	.316	10.049	1	**.002**	2.724	1.466	5.061
Spring	.502	.298	2.844	1	.092	1.653	.922	2.963
Reference Category – Summer								
Section Type			11.636	2	.003			
Tunnel	.531	.388	1.873	1	.171	1.700	.795	3.636
Driveway/Minor Intersection	1.161	.359	10.434	1	**.001**	3.192	1.578	6.455
Reference Category – Road Segment								
Speed Limit (< 80 km/h)	.635	.238	7.087	1	**.008**	**1.886**	1.182	3.009
Vehicle Age	.467	.221	4.458	1	**.035**	**1.595**	1.034	2.459
Commuting Traveler	1.292	.378	11.702	1	**.001**	**3.641**	1.736	7.633
Work Traveler	-.684	.317	4.667	1	**.031**	**.504**	.271	.938
Constant	-1.563	.317	24.307	1	<.001	.209		

Probabilities (p) and Odds Ratios (OR) of statistically significant explanatory variables are reported in boldface.

Table 8 - Results of Logistic Regression (Outcome Variable: Unfamiliarity).

	β	S. E.	Wald	df	p	OR	CI for the OR (95 %)	
							Lower	Upper
Traffic Volume			6.197	2	.045			
Medium: 5,000 – 11,000	.332	.307	1.169	1	.280	1.394	.763	2.546
High: > 11,000	.875	.352	6.194	1	**.013**	2.399	1.204	4.777
Reference Category – Low: < 5,000								
Traffic Seasonal Variation (High: SDT/AADT > 1.4)	2.106	.293	51.791	1	**<.001**	**8.213**	4.629	14.574
Season			8.327	3	.040			
Autumn	-.888	.318	7.809	1	**.005**	**.411**	.221	.767
Winter	-.534	.311	2.960	1	.085	.586	.319	1.077
Spring	-.359	.303	1.403	1	.236	.699	.386	1.265
Reference Category – Summer								
Accident Type			9.364	2	.009			
Rear-end/Angle	.561	.286	3.857	1	**.050**	**1.752**	1.001	3.067
Head-on	.904	.300	9.111	1	**.003**	**2.470**	1.373	4.443
Reference Category – Run-off								
Heavy Vehicles involved	.712	.248	8.228	1	**.004**	**2.039**	1.253	3.318
Commuting Traveler	-.913	.456	4.007	1	**.045**	**.401**	.164	.981
Young drivers involved	-.474	.241	3.865	1	**.049**	**.623**	.388	.999
Constant	-1.686	.340	24.665	1	<.001	.185		

Probabilities (p) and Odds Ratios (OR) of statistically significant explanatory variables are reported in boldface.

Table 9 – Crosstabulation Familiarity (interactions excluded) x Accident Type. Observed counts and adjusted residuals[1] (in brackets).

Familiarity	Accident Type: Run-off	Rear-end/Angle	Head-on	Total
Only Familiar	79 (-0.1)	24 **(2.8)**	13 **(-2.4)**	116
Only Unfamiliar	53 (1.5)	6 (-1.3)	11 (-0.6)	70
Only Transition	123 (-1.1)	20 (-1.5)	44 **(2.7)**	187
Total	255	50	68	373

[1]Adjusted standardized residuals are considered significant and then highlighted in boldface if they are greater than 2 (absolute value), since this is can be considered as a small table (Agresti, 2007 [69]). The expected frequencies are greater than 5.

Table 10 – Crosstabulation Familiarity x Accident Dynamics. Observed counts and adjusted residuals[1] (in brackets).

Familiarity	Accident Dynamics: Moving	Stationary	Out of Control	Maneuvering	Total
Familiar	103 (-1.4)	29 **(2.3)**	78 (0.9)	13 (-1.4)	223
Transition	195 (0.1)	24 **(-2.9)**	132 (0.9)	36 (1.3)	387
Unfamiliar	91 (1.4)	19 (1.1)	43 **(-2.1)**	13 (-0.1)	166
Total	389	72	253	62	776

[1]Adjusted standardized residuals are considered significant and then highlighted in boldface if they are greater than 2 (absolute value), since this is can be considered as a small table (Agresti, 2007 [69]). The expected frequencies are greater than 5.