Accident Analysis and Prevention 103 (2017) 112–122

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Full length article

Analysis of factors affecting the severity of crashes in urban road intersections

L. Mussone[a,*], M. Bassani[b], P. Masci[b]

[a] Politecnico di Milano, Department of Architecture, Built Environment and Construction Engineering (DABC), 9, Via Bonardi, Milano 20133, Italy
[b] Politecnico di Torino, Department of Environmental, Land and Infrastructures Engineering (DIATI), 24, Corso Duca Degli Abruzzi, Torino 10129, Italy

ARTICLE INFO

Keywords:
Urban roads
Road intersection
Crash severity level
5-min flow
Short-term data
Back-propagation neural network
Generalized linear mixed model

ABSTRACT

Road crashes are events which depend on a variety of factors and which exhibit different magnitudes of outputs when evaluated with respect to the effects on road users. Despite a lot of research into the evaluation of crash likelihood and frequency, only a few works have focused exclusively on crash severity with these limited to sections of freeways and multilane highways. Hence, at present there is a large gap in knowledge on factors affecting the severity of crashes for other road categories, facilities, and scenarios.

The paper deals with the identification of factors affecting crash severity level at urban road intersections. Two official crash records together with a weather database, a traffic data source with data aggregated into 5 min intervals, and further information characterising the investigated urban intersections were used. Analyses were performed by using a back propagation neural network model and a generalized linear mixed model that enable the impact assessment of flow and other variables. Both methods demonstrate that flows play a role in the prediction of severity levels

1. Introduction

In road safety research, analysts are interested in predicting road crashes in terms of location, frequency, pattern, and severity in order to ensure better traffic operations and save lives. A knowledge of the relationship between crash severity and the environmental and traffic conditions is fundamental to achieve this goal.

Regarding weather, the most important factors influencing crashes are those that affect the available friction between wheel and pavement, and/or driver visibility, resulting in crashes when the driver is unable to avoid collisions with moving or fixed obstacles. The conclusions of Caliendo et al. (2007) and Theofilatos et al. (2012) confirm that in the case of rainfall, the frequency of severe crashes increases. Rainfall may change its intensity rapidly, hence average annual rainfall precipitation, and even hourly rainfall may not be sufficient to capture the real-time rainy weather conditions prior to or during crash occurrence. But weather data cannot always be collected in very short intervals and close to each crash location, so Jung et al. (2010) suggested taking into consideration interpolation techniques to derive unmeasured (or even unmeasurable) data.

It is normally difficult to interpret and compare crash data in adverse weather conditions with those in good weather conditions. This is due to factors related to the ability of drivers to adapt their behaviour to weather conditions (Theofilatos and Yannis, 2014), and to possible changes in the composition of the driver population under adverse weather conditions where aggressive drivers (usually young males) tend to persist with their behaviour and speed, while the rest of the population tends to assume a more prudent attitude or avoid driving altogether (Hill and Boyle, 2007).

On the other hand, traffic flow can explain the number of conflicts between vehicles, hence if flow increases then the interferences between vehicles should increase, and their crash risk exposure as well. Theofilatos and Yannis (2014) stated that the influence of traffic flow has been considered more than other traffic related parameters such as speed, density and occupancy, mainly because it is simpler to measure. However, different circulation regimes are possible for the same vehicular volume (daily or even hourly), and this may lead to contrasting or difficult interpretations of crash outcomes. In the case of urban areas, Noland and Quddus (2005) observed that congestion does not significantly affect crash severity in the greater London metropolitan area.

One way to overcome the effect of traffic variables is to avoid the use of aggregate traffic data (i.e., AADT, hourly flow) which are not consistent with the traffic flow levels at the time of crashes. In fact, they hide and smooth out volume peaks, and provide a brief reference to the volume of traffic characterizing a road section or intersection.

* Corresponding author.
E-mail addresses: mussone@polimi.it, lorenzo.mussone@polimi.it (L. Mussone), marco.bassani@polito.it (M. Bassani), pietro.masci@gmail.com (P. Masci).

http://dx.doi.org/10.1016/j.aap.2017.04.007
Received 15 December 2016; Received in revised form 6 April 2017; Accepted 8 April 2017
Available online 19 April 2017
0001-4575/

Furthermore, neither hourly traffic volumes, volume per lane, nor V/C (volume/capacity) ratio necessarily link the crash events to the explanatory traffic variables. This conclusion was also drawn for weather, therefore traffic and weather data available in an aggregated format cannot always depict the real event conditions.

Xu et al. (2013) carried out a crash severity analysis along a 29-mile of a freeway segment using real-time traffic data on flow, occupancy, and speed. They used a sampling frequency of 30 s to calibrate and validate a sequential logit model, linking the likelihood of crash occurrences at different severity levels (SL) to the previously mentioned traffic flow characteristics. Results showed that property-damage only (PDO) crashes tend to take place in congested conditions with a highly variable speed and frequent lane changes. Injury and fatal crashes occur more often in less congested flow conditions, while fatal crashes occur under uncongested conditions as well as when there are large differences in speed between adjacent lanes. Similar conclusions were drawn by Christoforou et al. (2010) with traffic data records of 6 min each. Yu and Abdel-Aty (2014) conducted a SL analysis on a mountainous freeway on the basis of 6-min traffic and weather data, with the SL classified into two levels (severe injury and PDO). Steep grades, standard deviation of speed, temperature, and snow were found to be the most influential variables for this type of facility. Other authors used real-time traffic data for crash and safety analysis in urban arterials and urban expressways (Shi et al., 2016a,b; Theofilatos and Yannis, 2016; Hossain and Muromachi, 2013a,b).

Shankar et al. (1996) observed that road safety studies were historically limited to the localization of fatalities, even though the estimation of consequences in terms of SL (from PDO to fatalities) could help in understanding the benefits accruing from countermeasures. Furthermore, at present there are no studies in literature that focus on the contributing causes to a specific SL in the event of a crash on urban road networks.

To bridge this gap, the paper aims at providing knowledge on factors contributing to crash severity in urban road intersections. Crash, traffic and the weather databases of the Turin road network in the North-West of Italy were collected and used to calibrate and validate predictive models for crash SL. The Artificial Neural Network (ANN), a robust tool used to investigate complex phenomena without assuming any preliminary hypotheses on the model, was used. Since the ANN cannot provide an analytical formulation, a Generalized Linear Mixed Model (GLMM) was also applied. Both models were subjected to sensitivity analysis to comprehend the effects of each variable.

The ANN method is well-known and there are many papers on its use for safety analysis in different scenarios (Karlaftis and Vlahogianni, 2011). There are fewer works using GLMM on the same subject. An example is the paper by Bailey and Hewson (2004) which analyses the incidence of fatal and serious crashes for different types of road users by a multivariate GLMM (Bailey and Hewson, 2004). Gargoum et al. (2016) explored the relationships between some features of road surroundings (geometric, temporal factors, and weather conditions), and driver compliance (a categorical variable) with speed limits by a GLMM (data are modelled using a cumulative logit model with random intercepts). A mixed logit model was also used by Milton et al. (2008) in an exploratory analysis of crash severity in highways. Similarly, Chen and Chen (2011) applied a mixed logit model to the study of the severity of traffic accidents involving trucks in single or multi-vehicle crashes. A comparison of three crash severity models, multinomial logit, ordered probit, and mixed logit with regard to crash data underreporting effects was proposed in Ye and Lord (2011). Qin et al. (2013) also compared the results obtained by three logistic regression models (multinomial logit, partial proportional odds and mixed logit) used to investigate the effects on crash severity of large trucks. With the aim of employing a multivariate approach to the investigation of crash severity, Ma and Kockelman (2006) proposed a Bayesian Poisson regression. Ma et al. (2008) introduced a multivariate Poisson approach to model injury counts by severity whereas Wang et al. (2017) identified the effects of a number of factors on different crashes by using a multivariate Poisson log-normal regression

2. Database formation

2.1. Crash and traffic data

The crash data used in this research were obtained from the *Istituto Nazionale di Statistica* (ISTAT). The database contains details on crash dynamics and location, on vehicles and on the individuals involved, but it does not include PDO events in accordance with current Italian legislation, specifically articles number 582, 583 and 590 of the Italian Penal Code 2015 (Repubblica Italiana, 2015). In fact, Italian law defines road accidents as crashes only when they result in at least one injury, and crash consequences are classified into five severity levels (SL) all of which refer to the most seriously injured road user in any particular crash:

- very slight injuries (VSI), when the most seriously injured person has a prognosis of fewer than 20 days;
- slight injuries (SLI), when the prognosis is between 21 and 40 days;
- severe injuries (SEI), if the event causes an illness that endangers the life of the injured party, and/or if the event results in the permanent weakening of brain function or a body organ;
- guarded prognosis (GPR), if the doctor cannot determine the disability, and he issues a report of “guarded prognosis” (until his reservations can be resolved, the road crash must be considered and treated as a determining factor); and
- fatalities (FAT), which include any injured persons who die within 30 days of the crash.

The ISTAT database was matched up with information from Turin’s Municipal Police, to include: (a) historical data, in particular the time, to the nearest minute, day, month and year of the crash event; (b) locality data with the name of the street and house number where the crash took place along a road segment, or the denomination of two streets when it occurred at an intersection; and (c) generic information concerning crash SL.

Traffic data were provided by the 5T Company which uses induction-loop traffic sensors, located along the exiting lanes of the monitored intersections, with flow data collected every 5 min. Table 1 reports the crash data counts that were associated with 5-min flow data, while Fig. 1 shows the portion of the road network monitored by 5T in 2006, and includes the time scale used to estimate the seven 5-min flows of the 35 min before, during and after the crash event.

They are divided into function of SL, road typology, pavement conditions, vehicle type, gender and age of drivers (no more than two, and indicated as A and B). Variables relating to driver B have a relatively high percentage of unknowns since some crashes are single--vehicle collisions (e.g. rollover, roadway departure, collision with animals) thus involving only vehicle A. Driver B may be a pedestrian or even a child riding a bicycle.

Table 2 reports the traffic flow (TF) as the number of vehicles counted in 5-min intervals over a period of 35 min. Only crashes occurring along intersections yielding valid and reliable traffic data were extracted from the main database for further use. As a result, the database used for model calibration was obtained as a subset of the total number of crashes recorded in the official database, since only crashes that occurred on monitored roads were included (therefore it can be considered a random sample of crashes on monitored roads).

2.2. Weather data

The Environmental Protection Agency of the Piedmont Region (ARPA Piedmont) provided data on weather conditions from the Turin weather station. It is located in the city centre at 238 m a.s.l.,

Table 1
Crash data characteristics (distinguished in terms of severity, day, hour, brightness, rainfall, road typology, pavement conditions, type of vehicle, gender and age of drivers).

Data			Frequency
Crashes	Severity Level (SL)	VSI (SL = 2)	1531 (83.3%)
		SLI (SL = 3)	207 (11.3%)
		SEI (SL = 4)	49 (2.7%)
		GPR (SL = 5)	33 (1.8%)
		FAT (SL = 6)	18 (1.0%)
	Day	Weekdays	1307 (71.1%)
		Weekend	531 (28.9%)
	Hours	12.00 am–6.59 am	239 (13.0%)
		7.00 am–8.59 am	192 (10.4%)
		9.00 am–10.59 am	182 (9.9%)
		11.00 am–12:59 pm	209 (11.4%)
		1.00 pm–3.59 pm	309 (16.8%)
		4.00 pm–5.59 pm	208 (11.3%)
		6.00 pm–11:59 pm	499 (27.1%)
	Brightness	dark	963 (52.4%)
		bright	875 (47.6%)
	Rainfall	not event	1709 (93.0%)
		event	129 (7.0%)
Road typology		Unknown	497 (27.0%)
		One carriageway	208 (11.3%)
		Two or more carriageways	1133 (61.6%)
Pavement conditions		Unknown	497 (27.0%)
		Dry	1132 (61.6%)
		Wet (other)	209 (11.4%)
Type of vehicle A		Unknown	497 (27.0%)
		Passenger car	1111 (60.4%)
		Other	230 (12.5%)
Type of vehicle B		Unknown	704 (38.3%)
		Passenger car	869 (47.3%)
		Other	265 (14.4%)
Gender of driver A		Unknown	500 (27.2%)
		Male	1021 (55.5%)
		Female	317 (17.2%)
Age of driver A		Unknown	540 (29.4%)
		18–24	233 (12.7%)
		25–64	973 (52.9%)
		> 64	92 (5.0%)
Gender of driver B		Unknown	704 (38.3%)
		Male	828 (45.0%)
		Female	306 (16.6%)
Age of driver B		Unknown	739 (40.2%)
		18–24	225 (12.2%)
		25–64	804 (43.7%)
		> 64	70 (3.8%)

1.5 m off the ground, and at a latitude of 45°.066667 and longitude of 7°.683333. The station collects hourly data on several variables, but the data used in this investigation and summarized in Table 3 were limited to air temperature in °C, total light radiation in W/m^2, and rainfall intensity in mm/h. Each crash record was associated with the three weather data recorded in the hour of the event.

3. Variables and preliminary operations

Table 4 lists the independent variables, their numbering and labels, referring to injury crashes included in the database and used for model calibration and validation. The variables referring to the road conditions were:

- road type (C1), indicating the organization of the carriageways and the directions served (0 = unknown; 1 = one carriageway, one way; 2 = one carriageway, two ways; 3 = two carriageways, two ways; 4 = more than two carriageways, two ways);
- pavement conditions (C2), distinguished according to the presence of water, snow or ice (0 = unknown, 1 = dry, 2 = wet, 3 = slippery, 4 = icy/frozen, 5 = snowy); and
- road signage (C3), indicating if it was absent (0), composed of vertical signs only (1), horizontal markings only (2), if both were present (3), or if a temporary construction signage was present (4).

The vehicle characteristic variables were:

- vehicle type (C4 for vehicle A, and C6 for vehicle B) ranging from 0 (passenger cars) to 20 (quad), including the case of vehicles that fled the crash scene (19); and
- vehicle category (C5 for vehicle A, and C7 for vehicle B) from 0 to 8, in which 1 represents cars, 2 buses, 3 trams, 4 heavy vehicles, 5 industrial vehicles, 6 bikes, 7 motorcycles, 8 vehicles that fled the crash scene and 0 unclassified vehicle s.

The driver description variables were:

- age (C8 for driver of vehicle A, and C11 for driver of vehicle B) ranging from a minimum of 10 (for driver B) to a maximum of 89 (for driver A); this variable also assumed the null value in cases of unknown/unrecorded age;
- age class (C9 for driver of vehicle A, and C12 for driver of vehicle B) which groups the ages into 6 intervals ranging from 0 to 5: 0 in the case of unknown/unrecorded data, 1 for very young drivers (15–19 years old), 2 for young drivers (20-24 years old), 3 for adults (25–64 years old), 4 for elderly drivers (from 65 to 79), and finally 5 for very old drivers (over 80); and
- gender of drivers (C10 for driver of vehicle A, and C13 for driver of vehicle B) which assumes the value 0 in cases of unknown/unrecorded data, 1 for males, and 2 for females.

The lowest values of 'age of driver A' could refer to scooter drivers, while those of driver B may refer to pedestrians or cyclists. The measured wind speed, solar radiation, and rainfall precipitation values were numerical.

The flow data are numerical and represent the volume of vehicles per hour (veh/h), while the standard deviation of the seven flow values is calculated and added to the list, in order to take flow fluctuations directly into account. Finally, the output variable indicating the severity is also reported in Table 4 as a number ranging from 2 (VSI) to 6 (FAT).

3.1. Introductory data analyses

A correlation analysis was performed to assess the level of collinearity in input data for injury crashes. From variable C1 to C13, the set-up of road, vehicle, and driver variables showed high correlation values for each subset of variables (grouped by 3, 2, 2, 3, and 3 variables, respectively, referring to road, vehicle A, vehicle B, driver A, and driver B). From C14 to C18, the set of weather variables were only slightly correlated, while variable C18 (rainfall) showed no correlation with any other variable. The traffic flow variables from C19 to C25 and their standard deviation (variable C26) were highly correlated. The high correlation between TF variables would seem to preclude their contextual use. Since the aim was to investigate the role of flow values along the intervals around the crash event, and because the exact combinations of the seven flow values are generally not trivial (in the sense that their combination is not easily predictable), the authors decided to consider all of them in the calibration of models.

A cluster analysis using the SOM technique (Kohonen, 2001) was also performed to understand possible relationships between data, following the results of a previous research (Mussone and Kim, 2010). However, no specific relationship was discovered between these

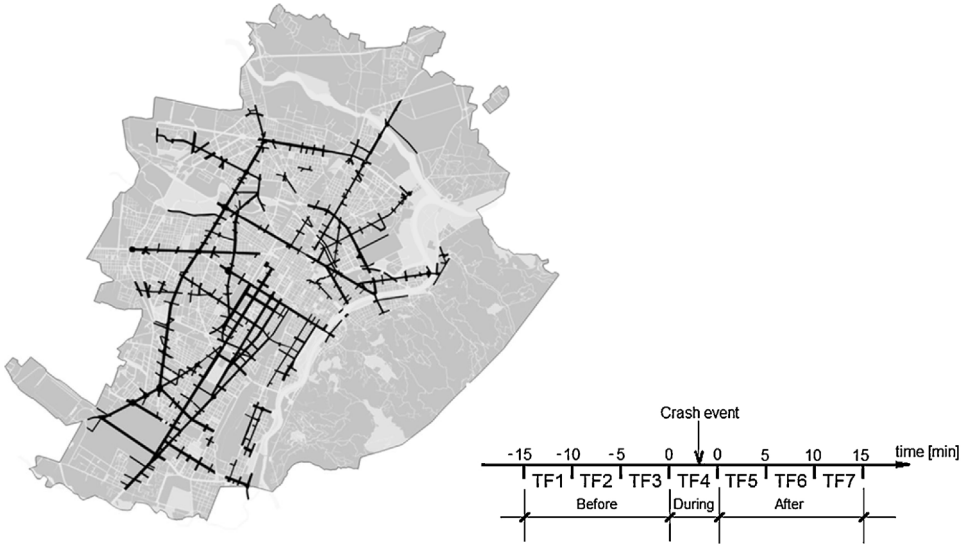

Fig. 1. Turin's traffic monitoring network operated by 5T in 2006 (highlighted in black), and time scale used to aggregate traffic flows (TF) in the seven 5-min periods across the crash event.

Table 2
Descriptive statistics for traffic flow (TF) data included in the injury crash database.

TF	Description		Values
1	Traffic flow from −10 to −15 min before the 5 min that include the event (veh/5 min)	Mean	201.0
		St. dev.	140.9
		Min	0
		Max	1019
2	Traffic flow from −5 to −10 min before the 5 min that include the event (veh/5 min)	Mean	200.3
		St. dev.	141.6
		Min	0
		Max	1065
3	Traffic flow from −0 to 5 min before the 5 min that include the event (veh/5 min)	Mean	198.6
		St. dev.	142.8
		Min	0
		Max	1218
4	Traffic flow in the 5 min that include the event (veh/5 min)	Mean	196.6
		St. dev.	140.0
		Min	0
		Max	1175
5	Traffic flow from 0 to +5 min after the 5 min that include the event (veh/5 min)	Mean	196.4
		St. dev.	139.6
		Min	0
		Max	1095
6	Traffic flow from +5 to +10 min after the 5 min that include the event (veh/5 min)	Mean	196.3
		St. dev.	140.6
		Min	0
		Max	1009
7	Traffic flow from +10 to +15 min after the 5 min that include the event (veh/5 min)	Mean	196.2
		St. dev.	140.2
		Min	0
		Max	990

Table 3
Descriptive statistics for weather conditions in the database.

Weather variable		Values
Air temperature (°C)	Mean	+14.9
	St. dev.	+9.2
	Min	−5.6
	Max	+36.4
Total light radiation (W/m^2)	Mean	189.6
	St. dev.	249.0
	Min	0.0
	Max	966.0
Rainfall intensity (mm/h)	Mean	0.10
	St. dev.	0.7
	Min	0.0
	Max	24.2
Wind speed (km/h)	Mean	0.10
	St. dev.	0.7
	Min	0.0
	Max	24.2

clusters and crash characteristics.

Finally, the Principal Component analysis (PCA) (Lebart et al., 1977; Jolliffe, 1986) was conducted to investigate the information content of the database. Table 5 reports the variance explained by the first eight components that account for about 93% of the total variance for both databases, while the first two components account for about 62%. The variables most closely linked to the first component are signage, road type, and gender of drivers A and B (variables C3, C1, C10, C13, respectively); whereas those linked to the second component are light/dark, light radiation, and air temperature (variables numbers 17, 16, 14, respectively). Flow variables (TF1-TF7) are all linked to the third component. As a result, the set of variables relating to road and driver can together account for about 46% of the variance; those related to meteorological conditions about 16%, and those related to flow about 7%.

3.2. Data treatment

According to Table 1, the categories of crash SL are not equally represented with a high number of VSI and SLI injury crashes and a very low number of FAT crashes, so the dataset is unbalanced. This should not generally present a problem for logistic regression, but it certainly presents one for machine learning tools, and especially with artificial neural networks (ANN). With unbalanced datasets, ANN could not find the correct relationships between input and output for all categories present in the dataset. Therefore, the natural distribution of the dataset is not the best distribution for training a classifier.

Since the focus of the paper is on the effect of weather and flow on crash severity, and considering that overfitting does not distort the relationships between input and output variables, the authors chose to oversample the data to run both the BPNN and GLMM codes (described in the next Section 4), in order to have only one reference database. The oversampling method proposed by Japkowicz (2000), which consists of the duplication of those data in the minority classes (in this case FAT, GPI, and SEI), was adopted. According to this approach, duplication

Table 4
Number and labels of variables.

Variable label	Description	Type	u.m.	min	max
C1	Road type	C	–	0	4
C2	Pavement conditions	C	–	0	5
C3	Road signage	C	–	0	4
C4	Vehicle type A	C	–	0	20
C5	Vehicle category A	C	–	0	8
C6	Vehicle type B	C	–	0	18
C7	Vehicle category _B	C	–	0	7
C8	Age of driver vehicle A	N	–	16	86
C9	Age classes A	N	–	0	5
C10	Gender of driver vehicle A	C	–	0	2
C11	Age of driver vehicle B	N	–	10	80
C12	Age classes B	N	–	0	5
C13	Gender of driver vehicle B	C	–	0	2
C14	Air temperature	N	°C	−5.6	+36.4
C15	Wind speed	N	m/s	0	9.10
C16	Light radiation	N	W/m^2	0	966
C17	Light/dark (day/night)	B	–	0	1
C18	Rainfall	N	mm/h	0	24.2
C19	Traffic Flow #1-TF1 (*)	N	veh/h	0	1129
C20	Traffic Flow #2 – TF2 (*)	N	veh/h	0	1124
C21	Traffic Flow #3 – TF3 (*)	N	veh/h	0	1118
C22	Traffic Flow #4 – TF4 (*)	N	veh/h	0	1175
C23	Traffic Flow #5 – TF5 (*)	N	veh/h	0	1095
C24	Traffic Flow #6 – TF6 (*)	N	veh/h	0	1110
C25	Traffic Flow #7 – TF7 (*)	N	veh/h	0	1075
C26	Flow standard. deviation	N	veh/h	0	337
C27	Severity level – SL	N	–	2	6

Notes: N indicates “numerical”, B indicates “Boolean”, C indicates “categorical”; u.m. stands for unit of measurement; (*) Refer to Table 2 for a complete description of the variables.

Table 5
Percentage of variance accounted for by the first eight components in PCA.

Component	1	2	3	4	5	6	7	8
Simple value	46.16	15.76	7.46	6.45	5.42	4.66	3.87	2.28
Cumulative value	46.16	61.92	69.37	75.83	81.25	85.91	89.78	92.97

was carried out on the subset of data until their count was of the same dimension as that of the most populated class.

Data was also normalized since this helps the learning machines working. Simple rescaling (or feature scaling, or unity-based normalization) was employed due to its simplicity. Assuming that X_{min} and X_{max} are the two extreme values (minimum and maximum) of a variable X, the normalized variable X' will be:

$$X' = \frac{X - X\min}{X\max - X\min} \quad (1)$$

Only the input variables were normalized according to Eq. (1), while the output variable (SL) remained numerical ranging from 2 to 6.

4. Modelling and results

In this paper, the authors have used two different modelling approaches: the Artificial Neural Network (ANN) model, and the Generalized Linear Mixed Model (GLMM). The ANN model provides a relatively good approximation of the relationship between input and output irrespective of linearity. The GLMM method provides the authors with a direct insight into the equations describing that relationship. Compared to other linear regression models, GLMM has certain advantages: it is an extension of logistic regression but with random effects on some grouping variables (overcoming some difficulties of GLM) and, apart from the Gaussian one, may have different distributions for the output variable. Their use has to be considered complementary.

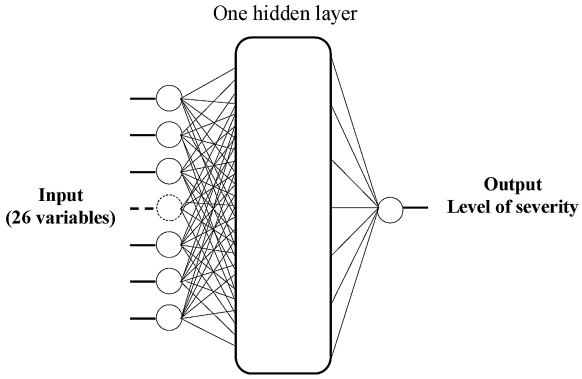

Fig. 2. Back-propagation Neural Network structure for SL modelling.

4.1. Back-propagation neural network (BPNN)

The BPNN is one of the ANN models, and has a classical multilayer topology with feed-forward connections. ANN models have already been used in contributions dealing with the problem of crash prediction or severity (Abdelwahab and Abdel-Aty, 2001; Chong et al., 2004; Delen et al., 2006; Baluni and Raiwani, 2014).

A BPNN does not need any a priori assumptions on relationships between linear or non-linear variables, and offers the opportunity to investigate and create the first discriminant analysis in problems where the phenomena (the relationships between input and output) are not well known and an analytical approach could be time consuming. A BPNN does not provide an analytical formulation between input and output, so the only way to understand the effects of input variables is to carry out a sensitivity analysis of the model.

In this investigation, the BPNN model was calibrated and validated with the Levenberg-Marquardt training algorithm, and performances were evaluated according to Mean Squared Errors (MSE) through the three phases of train, test, and validation. According to Fig. 2, the model has an input layer of 26 independent variables listed in Table 5, a hidden layer, and an output layer corresponding to the SL (with one neuron according to the numerical output previously mentioned). Finally, the best model was found to be made up of 35 neurons in the hidden layer, with a MSE close to 0.08, which means, effectively, that there are at most 8 errors for every 100 classifications.

The variables used in the model were selected through the backward procedure that involves the elimination of one variable at a time, while assessing whether or not it affects the overall performance of the neural network in predicting the output. However, it should be noted that this approach does not necessarily guarantee the best performance since the elimination of correlated variables depends on the order in which they are evaluated.

4.2. Generalized linear mixed model

For comparison purposes and in order to obtain an analytical relationship between input and output, a generalized linear mixed model (GLMM) was calibrated and validated. Mixed models for continuous normal outcomes have been developed since the seminal paper of Laird and Ware (1982). Many developments were also proposed for non-normal data (Booth and Hobert, 1998) and generically classified as generalized linear mixed models, an extension of generalized linear models (GLMs) that include random effects. Their inclusion leads to valid results and limits the extent to which variation can be attributed to variables (a normal distribution on random effects is generally assumed). GLMM is a regression model of a response variable that comprises data, a model description, fitted coefficients, covariance parameters, design matrices, residuals, residual plots, and other diagnostic information. Fixed-effects terms usually refer to the conventional linear regression part of the model, while random effects

Table 6
Fixed effects coefficients estimates and Random effects covariance parameters at 95% CIs for the GLMM intersection model.

Variable	Description	Estimate	SE	p-value	Lower	Upper
Intercept	–	0.90914	0.14495	3.76×10^{-10}	0.625	1.1933
C11	Age driver veh. B	−0.00236	0.001023	0.020894	−0.00437	−0.00036
C12	Age classes B	0.070388	0.019014	0.000216	0.033115	0.10766
C14	Air temp.	−0.00475	0.0007	1.24×10^{-11}	−0.00612	−0.00338
C19	TF1	0.000625	0.000177	0.000428	0.000277	0.000972
C20	TF2	0.000684	0.000161	2.19×10^{-5}	0.000368	0.001
C21	TF3	0.001861	0.000214	4.53×10^{-18}	0.001441	0.002281
C22	TF4	−0.00106	0.000179	4.03×10^{-9}	−0.00141	−0.0007
C23	TF5	−0.00131	0.0002	6.49×10^{-11}	−0.0017	−0.00091
C24	TF6	−0.00078	0.000166	2.77×10^{-6}	−0.0011	−0.00045

Group variable		Estimate
C3 (Intercept)	Road signage	0.094612
C4 (Intercept)	Vehicle type A	0.21884
C6 (Intercept)	Vehicle type B	0.26757
C10 (Intercept)	Gender driver veh. A	0.069366
C13 (Intercept)	Gender driver veh. B	0.10736

Indexes	
LogLikelihood	−13538
AIC	27107
BIC	27211
Deviance	27077
R^2 adjusted	0.2684

terms are associated with individual experimental units taken at random from a population, and account for variations between groups that might affect the response.

The GLMM structure used for this investigation is made up of the following equations:

$$y_i \Big| b \sim \text{Distr}\left(\mu_i, \frac{\sigma^2}{w_i}\right) \tag{2}$$

$$g(\mu) = \beta X + bZ + \delta \tag{3}$$

where y_i is the i-th element (dependent variable) of the y response vector, b is the random-effects vector (complement to the fixed β), Distr is a specified conditional distribution of y given b, μ is the conditional mean of y given b, and μ_i is its i-th element, σ^2 is the variance or dispersion parameter, w is the effective observation weight vector, and w_i is the weight for observation i, g(μ) is a link function that defines the relationship between the mean response μ and the linear combination of the predictors, X is a fixed-effects design matrix (of independent variables), β is a fixed-effects vector, Z is a random-effects design matrix (of independent variables), and δ is a model offset vector (residuals).

The model for the mean response μ is:

$$\mu = g^{-1}(\hat{\eta}) \tag{4}$$

where g^{-1} is the inverse of the link function g(μ), and $\hat{\eta}$ is the linear predictor of the fixed and random effects of the generalized linear mixed-effects model:

$$\eta = \beta X + bZ + \delta \tag{5}$$

The significant variables in the GLMM were selected through the maximization of the log-likelihood function. In the selection process, the Aikake Index Criterion (AIC), the Bayesian Information Criterion (BIC) and the Deviance (a combination of AIC and BIC) were also estimated. For the SL output, a log link function and the probability mass function (PMF) for the Poisson distribution were used. The fit method was the 'Laplace' one.

Model building is a difficult task even with GLMM and, generally speaking, an investigation of the entire set of variable permutations and their reciprocal interactions is not feasible. In order to render the model complex enough to accommodate the data without over-fitting, and simple enough to interpret by smoothing out the data, we only investigated models with linear predictors without mutual interactions, and random effects were investigated only for intercepts, leaving possible improvements for future research.

Table 6 reports the fixed effect coefficients that were calculated with a 95% confidence interval and an estimate for the random parameter. All p-values for significant variables are lower than 0.021, the standard error of estimates is generally much lower than the estimates, and lower and upper bonds of CI never include zero.

The effect of flows (from C19 to C24) on the SL has a different sign, positive for C19, C20 and C21 (before the crash), negative for C22, C23, and C24 (during and after the crash). This result indicates that traffic flow after a crash depends on crash severity: the higher the crash severity the lower the subsequent traffic flow. Other significant variables are age and age class of vehicle B driver (C11 and C12) and air temperature (C14). The variables relating to driver B have opposite sign and this is a little surprising but since the significance and coefficients are very different (in favour of C12) it may be that this result is due to a non-homogenous distribution of samples by age.

Therefore, the most likely conclusion is that severity increases with age of driver B. An increase in air temperature is related to a decrease in crash severity. It is important to highlight the relevance of grouping variables, which explain random effects and demonstrate the complexity and variability of crash data: C3 (signage), C4 (vehicle type A), C6 (vehicle type B), C10 (sex of driver A), and C13 (sex of driver B).

5. Model outputs assessment

According to Powers (2011), the results provided by classifiers can be evaluated by confusion matrix (or "contingency table" or "error

Table 7
BPNN model confusion matrix (percentage values in brackets), and "a priori" (PR) and "a posteriori" (PO) rates.

Real SL	Predicted SL							Number of crash data in the re-sampled database	PR
	< 2	2	3	4	5	6	> 6		
2 (VSI)	33 *(2%)*	1098 *(72%)*	247 *(16%)*	48 *(3%)*	29 *(2%)*	64 *(4%)*	12 *(1%)*	1531 *(100%)*	28%
3 (SLI)	0	63 *(4%)*	1372 *(95%)*	7 *(0.5%)*	0	7 *(0.5%)*	0	1449 *(100%)*	5%
4 (SEI)	0	0	0	1519 *(100%)*	0	0	0	1519 *(100%)*	0%
5 (GPR)	0	0	0	0	1518 *(100%)*	0	0	1518 *(100%)*	0%
6 (FAT)	0	0	0	0	0	1512 *(100%)*	0	1512 *(100%)*	0%
PO		5%	15%	3%	2%	4%			

matrix"), which represents, for each output, the number of predicted cases (a_{ij}) in the reference databases. In the two modelling techniques used in this investigation, the output is the SL and the number of predicted cases (a_{ij}) is calculated from the resampled databases (as discussed in Section 3.4).

It is also interesting to measure the ability to predict the percentage of correct data (i.e., precision), and the percentage of corrected data in respect of the total to be predicted (i.e., recall), with the goal of improving the recall measurement without weakening the precision one. Tables 7 and 8 include the "a priori" rate (PR) corresponding to the percentage of the predicted crashes to the total to be predicted for each SL (which also indicates the complement of the recall rate), and the "a posteriori" rate (PO) which is the complement of the precision rate, according to the following equations:

$$PR_i = 1 - a_{ii}/(a_{i1} + \ldots + a_{in}) \quad (9)$$

$$PO_i = 1 - a_{ii}/(a_{1i} + \ldots + a_{ni}) \quad (10)$$

where n is the matrix dimension. Furthermore, comments to the results are also supported by the estimate of the accuracy (A):

$$A = (a_{11} + a_{22} + \ldots + a_{nn})/\Sigma a_{ij} \quad (11)$$

Table 7 reports the confusion matrix for the BPNN model. SL values lower than 2 (corresponding to the PDO crash type) and greater than 6 (which are unrealistic values) have also been included in the table considering that the model output can fall outside the range of numerical values associated with each SL. The 93% accuracy rate for the BPNN model is certainly very high. PR and PO rates are low with the exception of SL 2 and 3, thus suggesting that the results pertaining to either of these two SLs can overlap. SL 2 is the more difficult to predict while SL 3 has the largest number of wrong cases assigned to it.

Table 8 contains the confusion matrix for the model calibrated with the GLMM. With this model, the SL prediction capacity is significantly lower than that for the BPNN model as indicated by the accuracy rate of 30%. The GLMM is better at generating results within the SL limits of 2 and 6, as confirmed by the fact that only eight values fall outside these limits. PR and PO rates are constantly low, showing how difficult it is to predict every SL.

6. Discussion

A sensitivity analysis was carried out on the BPNN model to assess how output changes by varying input normalized variable values in the range [0,1] one by one. A first set of scenarios referring to a particular set of input variables was prepared. In scenario S1 all variables were set to zero, in scenario S2 all variables were set to 0.5; while in scenario S3 all variables were set to 1. In addition, six other scenarios (4a, 4b, 4c, 4d, 4e, and 4f) were considered to study particular combinations of variable values, as reported in Table 9. These scenarios aim to consider some possible and typical crash situations involving male (4a, 4c, 4e, 4f), or female drivers (4b, 4d), during daytime (all except 4e) or at night-time (4e), with rainy weather (4c, 4d) or dry road surface (4a, 4b, 4e, 4f), or with elder drivers (4f).

Table 10 reports the results of the sensitivity analysis of BPNN, with numbers at the top of columns referring to the scenario, whereas numbers in the table refer to the maximum range observed in the SL fluctuation when varying the considered variable from 0 to 1 (by steps of 0.1). Outputs do not generally change monotonically when varying the input variable from 0 to 1, as a result of the complex interactions between variables represented by BPNN: the output values turn out to be concave, convex, or exhibit even more complex trends. The maximum SL fluctuation is 4 and the minimum is 0, hence a value of 4 in the table indicates the maximum possible effect produced by that

Table 8
GLMM model confusion matrix (percentage values in brackets), and "a priori" (PR) and "a posteriori" (PO) rates.

Real SL	Predicted SL							Number of crash data in the re-sampled database	PR
	< 2	2	3	4	5	6	> 6		
2 (VSI)	0	121 *(8%)*	673 *(44%)*	645 *(42%)*	80 *(5%)*	11 *(0.7%)*	1 *(0.1%)*	1531 *(100%)*	92%
3 (SLI)	0	56 *(4%)*	483 *(33%)*	721 *(50%)*	175 *(12%)*	7 *(0.5%)*	7 *(0.5%)*	1449 *(100%)*	67%
4 (SEI)	0	0	310 *(20%)*	1054 *(69%)*	155 *(10%)*	0	0	1519 *(100%)*	31%
5 (GPR)	0	0	46 *(3%)*	828 *(55%)*	506 *(33%)*	138 *(9%)*	0	1518 *(100%)*	67%
6 (FAT)	0	0	0	1092 *(72%)*	336 *(22%)*	84 *(6%)*	0	1512 *(100%)*	94%
PO		32%	68%	76%	60%	65%			

Table 9
Variable values for some of the investigated scenarios (4a-4f).

Variable no.	Description	Scenario					
		4a	4b	4c	4d	4e	4f
1	Road type	Two way					
2	Pavement conditions	Dry	Dry	Wet	Wet	Dry	Dry
3	Road signage	Present					
4	Vehicle type A	Light vehicle					
5	Vehicle category A	1					
6	Vehicle type B	Light vehicle					
7	Vehicle category _B	1					
8	Age of driver vehicle A	20	20	20	20	20	70
9	Age classes A	2	2	2	2	2	5
10	Gender of driver vehicle A	Male	Female	Male	Female	Male	Male
11	Age of driver vehicle B	20	20	20	20	20	70
12	Age classes B	2	2	2	2	2	5
13	Gender of driver vehicle B	Male	Female	Male	Female	Male	Male
14	Air temperature	20 °C					
15	Wind speed	Weak					
16	Light radiation [W/m^2]	740	740	480	480	0	740
17	Light/dark (day/night)	1	1	1	1	0	1
18	Rainfall [mm]	0	0	3	3	0	0
19	Traffic Flow 1–TF1	60%*	60%*	60%*	60%*	30%*	60%*
20	Traffic Flow 2 – TF2	60%*	60%*	60%*	60%*	30%*	60%*
21	Traffic Flow 3 – TF3	60%*	60%*	60%*	60%*	30%*	60%*
22	Traffic Flow 4 – TF4	60%*	60%*	60%*	60%*	30%*	60%*
23	Traffic Flow 5 – TF5	60%*	60%*	60%*	60%*	30%*	60%*
24	Traffic Flow 6 – TF6	60%*	60%*	60%*	60%*	30%*	60%*
25	Traffic Flow 7 – TF7	60%*	60%*	60%*	60%*	30%*	60%*
26	Flow std. dev.	20%*	20%*	20%*	20%*	40%*	20%*

Note: (*) Referred to the maximum observed.

variable in that scenario; a value of zero means that the input variable does not affect SL. Table 10 cannot show the elasticity of variable sensitivity which is graphically illustrated later.

The average SL (ASL) is calculated for both the columns and rows in Table 10 that is for each scenario and variable. This index measures the (average) impact of a variable or of a scenario on SL. It is difficult to find a meaning for ASL in itself (as an absolute value) but, certainly, it can be used for comparison purposes, since the ASL is determined by

Table 10
Sensitivity analysis of BPNN model (the SL fluctuation values ranging from 0 to 4 are observed by varying the reference variable from 0 to 1 by step of 0.1; ASL is the average SL calculated on all values of the scenario).

Variable	Scenarios									
	1	2	3	4a	4b	4c	4d	4e	4f	ASL
Road type	3	4	0	4	4	0	4	0	4	3.31
Pavement conditions	1	4	1	2	4	0	4	0	0	2.97
Road signage	4	4	4	2	0	2	0	4	1	3.17
Vehicle type A	1	4	0	2	2	0	4	0	2	3.24
Vehicle category A	3	4	0	4	4	2	2	4	3	2.97
Vehicle type B	4	4	0	0	2	0	4	0	0	3.42
Vehicle category B	0	2	0	4	4	2	2	2	2	2.97
Age of vehicle A	1	4	0	2	4	2	4	2	0	3.42
Age classes A driver	3	4	0	4	4	0	4	0	4	3.06
Gender of vehicle A driver	4	4	0	4	4	2	4	2	0	2.95
Age of vehicle B driver	0	4	0	4	4	2	4	2	0	3.61
Age classes B	3	4	0	2	4	0	4	2	4	3.34
Gender of vehicle B driver	0	2	0	2	0	2	4	0	0	3.78
Air temperature	2	4	1	4	0	4	4	0	1	3.46
Wind speed	2	4	1	2	0	2	4	0	0	3.11
Light radiation	4	4	1	2	4	2	4	2	1	3.53
Light/dark (day/night)	1	4	0	0	4	0	4	0	4	3.44
Rainfall	4	4	1	0	2	0	4	0	3	3.64
TF1	1	4	1	4	4	0	2	2	2	3.08
TF2	1	4	0	4	4	2	2	2	4	3.81
TF3	4	4	1	2	1	4	2	4	2	3.25
TF4	3	4	0	4	4	0	4	0	2	3.43
TF5	4	4	3	2	4	2	4	0	2	3.17
TF6	3	4	1	2	2	2	4	4	1	2.96
TF7	3	4	0	4	4	4	2	2	4	3.29
Flow std. dev.	0	4	1	4	4	4	4	2	3	2.95
ASL	3.22	4.25	2.91	3.52	2.98	3.79	3.10	3.86	2.39	-

Note: ASL = Average Severity Level; (*) concerning driver A or B, accordingly.

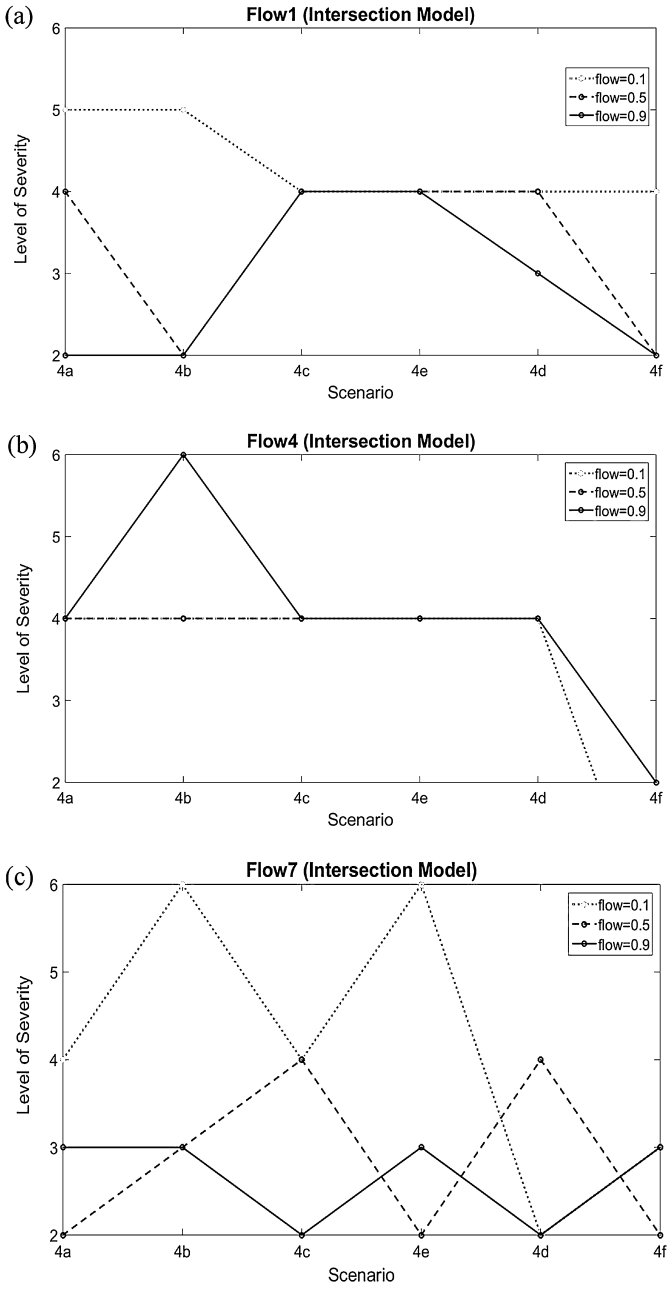

Fig. 3. BPNN sensitivity analysis of TF1 (a), TF4 (b), and TF7 (c) on SL for different flow values.

the overall values used for the input variables.

The BPNN model is very sensitive to flow variables and less sensitive to meteorological ones. Some other detailed results, such as examples of possible results achievable by ANN models, are illustrated in Fig. 3, where the sensitivity to some flow changes is considered for the scenarios 4a–4e.

In Fig. 3, the three different flow values for TF1, TF4 (when the crash occurred), and TF7, are considered in order to evaluate the effect of light, medium, and heavy traffic conditions. It must be emphasized that only one variable at a time changes while the others are fixed (in these cases other flows are set to 0.6).

The effect of flow varies a lot by scenario. When TF4 is high (continuous line curves in Fig. 3b), the SL is generally high for most of the scenarios, except for scenario 4f (elderly driver). Other results are difficult to interpret but what appears evident from the intersection model for TF7 (10–15 min after the crash) is that the higher the flow the lower the SL. This is reasonable since in the case of crashes with a low SL, vehicles involved in the crash do not need the intervention of emergency vehicles and so no significant disturbance is caused to the traffic.

The BPNN model shows a prevailing relationship between high flow and high SL, especially for TF4 (when the crash occurred). TF7 clearly indicates an inverse relationship between SL and flow: when a crash occurs, incoming flow is impeded and, therefore, is low.

One criticism arising from the use of flow data after the crash event (TF5–TF7) regards the reversal of the causality of variables: in fact, the crash event itself becomes the cause of alteration of flow. A possible justification is that there is an interest not only in the causality relationship but in the whole time series pertaining to the same process (such as occurs in incident detection).

The effect of driver A age (C8) is proposed in Fig. 4. The highest SL is attained by young female drivers for both dry pavement and rainy conditions (scenarios 4b and 4d). In general, for ages up to 30 years the SL is higher. Finally, light radiation has a limited impact on SL: the most relevant effect is with a quite low light radiation on female drivers with both dry pavement and rainy conditions. Generally, low radiation is more related to high SL than high radiation.

Flow plays a relevant role in the GLMM model, with only variable TF7 (flow from 15 to 20 min after crash) resulting as insignificant. The analysis of coefficient signs shows that it is negative for TF4, TF5, and TF6 which indicates that when the flow increases after a crash the SL is low. This result is comparable with that obtained by BPNN and shown in Fig. 3. In this figure three different flow values of TF1, TF4 and TF7 (0.1, 0.5, 0.9) are used to calculate SL. From Fig. 3c we can see that an increase in TF7 leads to a decrease in SL for most of the considered scenarios. TF1 and TF4 affect SL only for some scenarios and show more complex trends with respect to driver gender and age.

In order to complete an investigation of the models, a sensitivity analysis of GLMM, similar to that performed for BPNN, was also carried out, and reported in Table 11. It is based on the same scenarios (and predictor values) used for BPNN (Table 10) and, instead of marginal effects, the conditional means of response is considered here. This table shows the highest sensitivity of flow variables, particularly TF3, TF4, TF5 which are centred around the crash time. Some scenarios show a higher sensitivity to variable changes and a higher ASL than others, especially 4a and 4c which are focused on the role of male drivers in dry and wet pavement conditions.

Comparisons between road, driver, and environmental variables in the two modelling approaches is more difficult. The BPNN model shows the importance of all considered variables, though limited to some scenarios, while GLMMs consider a relatively limited set of significant variables for fixed effects but some others for random effects (in particular for vehicle type B) as already mentioned in Section 4.2.

7. Conclusions

The paper aims were the evaluation of crash severity level (SL) at intersections using environmental and traffic variables (some of which, like short-term flow, are new in this research area), through a back-propagation neural network model (BPNN) which uses a computational approach, and the generalized linear mixed model (GLMM) which uses an analytical approach.

Flow measurements can be used to quantify the number of potential conflicts occurring on the monitored road intersections. In this investigation, hourly weather parameters were also considered. These data made it possible to investigate how both flow and weather conditions are related to each other and to the other driver characteristic variables (and those of road users in general) involved in the crash. The results presented here, address a gap in the knowledge acquired from the number of studies on rural freeways and expressways repeatedly reported in literature.

BPNN models evidenced better performance in the prediction of the

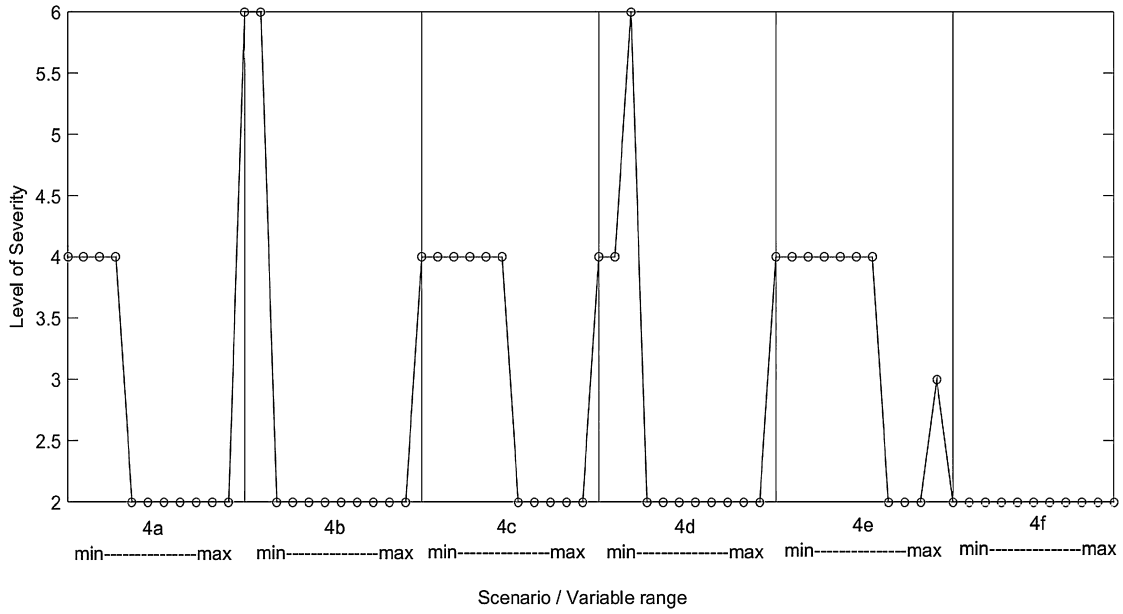

Fig. 4. Effect of driver A age on SL.

Table 11
Sensitivity analysis of GLMM model (the SL fluctuation values ranging from 0 to 4 are observed by varying the reference variable from 0 to 1 by step of 0.1; ASL is the average SL calculated on all values of the scenario).

Variable	Scenarios									
	1	2	3	4a	4b	4c	4d	4e	4f	ASL
Age of vehicle B driver	1	0	0	1	0	1	0	1	2	2.65
Age classes B	2	1	1	1	1	1	1	1	1	2.94
Air temperature	1	0	0	1	0	1	0	1	1	2.71
TF1	4	1	1	2	1	2	1	2	2	2.78
TF2	4	2	1	2	2	2	2	2	2	2.78
TF3	4	4	2	4	4	4	4	4	4	2.91
TF4	3	3	4	4	4	4	3	3	4	3.05
TF5	3	4	4	4	4	4	4	3	4	3.12
TF6	2	2	3	3	2	3	2	2	2	2.85
ASL	4.12	2.60	2.90	3.37	2.70	3.37	2.70	2.92	3.13	

Note: ASL = Average Severity Level; (*).

SL than those obtained by the GLMMs, according to the results obtained from the confusion matrixes. In fact, it is worth noting that BPNN models are able to accurately estimate any continuous and non-linear relationship between variables. However, BPNN does not allow a readier interpretation of model results, which instead is possible using GLMM. To have a clear idea of the pros and cons of statistical and neural network methods, the readers could refer to the already mentioned work of Karlaftis and Vlahogianni (2011). The authors are convinced that the main limit of GLMMs for these applications is represented by the linearity of the function. Kashani and Mohaymany (2011), as well as Yu and Abdel-Aty (2014), came to the same conclusion. In addition, missing data may have contributed to facilitate BPNN.

The BPNN and GLMM approaches show how underlying processes are likely to have different prevailing and concurrent causes. Both methods demonstrate that flows have a relevant role in predicting severity: this role is not limited to the flow when the crash occurred (TF4), but also extends to the other crash flow data (TF1-TF3 before crash occurrence, TF5-TF7 after crash occurrence). This finding may be controversial since it involves data gathered after the crash event but it merits greater attention in on-going research, for example working out an algorithm capable of confirming or identifying the real time interval when the crash occurred.

Unfortunately, the set of selected variables do not facilitate the assumption of specific countermeasures to reduce crash severity, since a high SL may be obtained by a combination of variables that cannot be avoided on the road, such as flow and light radiation. This suggests that information on some variables could be the basis for a future work investigating possible interventions.

Future research will include the application of generalized non-linear models and the study of higher order effects and the interaction between variables; then the use of GNLMM (Generalized Non Linear Mixed Model) will be considered. What is more, a mixed approach, using both short-term flow and AADT values, could be of some interest in order to extend the creation of models over a mid-long term period and investigate the relationship between them. Finally, future investigation should consider the possibility of calibrating and validating new models distinguishing the different types of intersection, corresponding to different types of manoeuvres and different effects of the variables already considered in the present investigation.

Acknowledgments

The authors thank the Polizia Municipale di Torino and the Città Metropolitana di Torino/Regione Piemonte for having provided crash data. Thanks are also due to Consorzio 5T s.r.l. for providing short-term

flow data and to the Environmental Protection Agency of the Regione Piemonte (ARPA Piemonte) for providing data on weather conditions.

References

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec. 1746, 6–13.

Bailey, T.C., Hewson, P.J., 2004. Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model. J. R. Stat. Soc. A 167 (Part 3), 501–517.

Baluni, P., Raiwani, Y.P., 2014. Vehicular accident analysis using neural networks. Int. J. Emerg. Technol. Adv. Eng. 4 (9), 161–164.

Booth, J.G., Hobert, J.P., 1998. Standard errors of prediction in generalized linear mixed models. J. Am. Stat. Assoc. 93, 262–272.

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accid. Anal. Prev. 39, 657–670.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single-and multi-vehicle accidents on rural highways. Accid. Anal. Prev. 43, 1677–1688.

Chong, M.M., Abraham, A., Paprzycki, M., 2004. Traffic Accident analysis using decision trees and neural network. arXiv preprint cs/0405050

Christoforou, Z., Cohen, S., Karlaftis, M., 2010. Vehicle occupant injury severity on highways: an empirical investigation. Accid. Anal. Prev. 42, 1606–1620.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38, 434–444.

Gargoum, S.A., El-Basyouny, K., Kim, A., 2016. Towards setting credible speed limits: identifying factors that affect driver compliance on urban roads. Accid. Anal. Prev. 95, 138–148.

Hill, J.D., Boyle, L.N., 2007. Driver stress as influenced by driving maneuvers and roadway conditions. Transp. Res. Part F: Traffic Psychol. Behav. 10 (3), 177–186.

Hossain, M., Muromachi, Y., 2013a. A real-time crash prediction model for the ramp vicinities of urban expressways. IATSS Res. 37 (1), 68–79.

Hossain, M., Muromachi, Y., 2013b. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.

Japkowicz, N., 2000. The class imbalance problem: significance and strategies. In: Proc. of the 2000 Intern. Conf. on Art. Intel. (IC-AI'2000). Las Vegas, Nevada. .

Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlaghttp://dx.doi.org/10.1007/b98835.978-0-387-95442-4.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accid. Anal. Prev. 42, 213–224.

Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. Transp. Res. Part C: Emerg. Technol. 19 (3), 387–399.

Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49 (10), 1314–1320.

Kohonen, T., 2001. Self-Organizing Maps, 3rd edition. Springer-Verlag, New York.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

Lebart, L., Morineau, A., Tabard, N., 1977. Techniques de la description statistique: méthodes et logiciels pour l'analyse des grands tableaux. Dunod, Paris.

Ma, J., Kockelman, K., 2006. Bayesian multivariate poisson regression for models of injury count, by severity. Transp. Res. Rec. 1950, 24–34.

Ma, J., Kockelman, K., Damien, P., 2008. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. Accid. Anal. Prev. 40, 964–975.

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40, 260–266.

Mussone, L., Kim, K., 2010. The analysis of motor vehicle crash clusters using the vector quantization technique. J. Adv. Transp. 44, 162–175.

Noland, R.B., Quddus, M.A., 2005. Congestion and safety: a spatial analysis of London. Transp. Res. Part A: Pol. Pract. 39, 737–754.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. 2, 37–63.

Qin, X., Wang, K., Cutler, C., 2013. Logistic regression models of the safety of large trucks. Transp. Res. Rec. 2392, 1–10.

Repubblica Italiana, 2015. Codice Penale (in *Italian*). Testo coordinato del Regio Decreto 19 ottobre 1930, n. 1398, aggiornato con le modifiche apportate dalla L. 28 aprile 2015, n. 58, dalla L. 22 maggio, 2015, n. 68 e dalla L. 27 maggio 2015, n. 69.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accid. Anal. Prev. 28 (3), 391–401.

Shi, Q., Abdel-Aty, M., Lee, J., 2016a. A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accid. Anal. Prev. 88, 124–137.

Shi, Q., Abdel-Aty, M., Yu, R., 2016b. Multi-level Bayesian safety analysis with unprocessed automatic vehicle Identification data for an urban expressway. Acc. Anal. Prev. 88, 68–76.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256.

Theofilatos, A., Yannis, G., 2016. Investigation of powered-two-wheeler accident involvement in urban arterials by considering real-time traffic and weather data. Traffic Inj. Prev. 18 (3), 293–298.

Theofilatos, A., Graham, D.J., Yannis, G., 2012. Factors affecting accident severity inside and outside urban areas in Greece. Traffic Inj. Prev. 13 (5), 458–467.

Wang, K., Ivan, J.N., Ravishanker, N., Jackson, E., 2017. Multivariate Poisson lognormal modeling of crashes by type and severity on rural two lane highways. Acc. Anal. Prev. 99, 6–19.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.

Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models multinomial logit, ordered probit, and mixed logit. Transp. Res. Rec. 2241, 51–58.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.