Transport Infrastructure and Systems – Dell'Acqua & Wegman (Eds)
© 2017 Taylor & Francis Group, London, ISBN 978-1-138-03009-1

Back-propagation neural networks and generalized linear mixed models to investigate vehicular flow and weather data relationships with crash severity in urban road segments

L. Mussone
Politecnico di Milano, Milano, Italy

M. Bassani & P. Masci
Politecnico di Torino, Torino, Italy

ABSTRACT: The paper deals with the identification of variables and models that can explain why a certain Severity Level (SL) may be expected in the event of a certain type of crash at a specific point of an urban road network. Two official crash records, a weather database, a traffic data source, and information on the characteristics of the investigated urban road segments of Turin (Italy) for the seven years from 2006 to 2012 were used. Examination of the full database of 47,592 crash events, including property damage only crashes, reveals 9,785 injury crashes occurring along road segments only. Of these, 1,621 were found to be associated with a dataset of traffic flows aggregated in 5 minutes for the 35 minutes across each crash event, and to weather data recorded by the official weather station of Turin. Two different approaches, a back-propagation neural network model and a generalized linear mixed model were used. Results show the impact of flow and other variables on the SL that may characterize a crash; differences in the significant variables and performance of the two modelling approaches are also commented on in the manuscript.

1 INTRODUCTION

Road crashes are events that depend on a variety of factors characterising human behaviour, weather, road pavement, vehicle stability and performance. Crash events show different magnitudes when evaluated with respect to the effects on road users (crash severity), and the knowledge of the contributing factors that affect the severity should be used to improve road safety through the action of transport policy makers, designers and road agencies.

Traffic volume, weather conditions, and road characteristics affect crash severity in a multifaceted way (Wang et al. 2013). Specifically, Theofilatos & Yannis (2014) pointed out that the few papers available mainly deal with roads operating under uninterrupted flow conditions, and recur mainly to logit modelling (Al-Ghamdi 2002, Golob et al. 2008, Christoforou et al. 2010, Jung et al. 2010, Xu et al. 2013, Yu & Abdel-Aty 2013). Earlier, Shankar et al. (1996) stated that crash severity investigations had been historically limited to the localization of fatalities, even though the estimation of the other severity levels (i.e., property damage only—PDO –, possible injury, non-incapacitating injury) could help in understanding the benefits of safety-improvement projects. Seventeen years later, Xu et al. (2013) again underlined that most of the research studies have been focused on the likelihood of a crash without considering the crash outcome severity.

One of the main obstacles to investigations of this type is the limited availability of comprehensive crash databases, and associated robust weather and traffic databases. Nowadays, however, a continuous flow of environmental and traffic data is collected by local road agencies with sensors of increasing quality and performance (Chong & Kumar 2003, Nekovee 2005). Contrary to what happened in the past, data are now frequently collected in intervals of shorter duration. Hence, available databases can be associated and merged with others containing data collected over several years of observations, thus supporting new robust inferences (El Faouzi et al. 2011).

To obtain reliable models and convincing results, the availability of high quality data representing the characteristics of drivers, together with traffic, weather, and pavement conditions is fundamental. Unfortunately, data describing every significant factor affecting crashes needs work to associate the available contrasting information that could be long, arduous, and sometimes unproductive.

The paper aims to bridge these gaps by providing knowledge on factors contributing to crash severity in an urban road network, considering only those

influencing crashes along road segments. Data on crashes, traffic and the weather database of Turin's road network (Italy) were collected and used to calibrate and validate predictive models of crash severity. The Back Propagation Neural Networks (BPNN), a robust tool used to investigate complex phenomena without assuming any preliminary hypotheses on the model, was used. But BPNN cannot give an analytical formulation of the mathematical functions linking the variables that significantly affect a certain phenomenon, thus only a sensitivity analysis of the model can be performed. A Generalized Linear Mixed Model (GLMM) was also used (with its analytical formulation) to compare and assess results with those obtained by BPNN.

2 DATABASE FORMATION

2.1 *Crash classification*

Crash data were provided by the *Istituto Nazionale di Statistica* (National Statistics Institute, ISTAT). The ISTAT database contains details on crash dynamics and location, on the vehicles, and on gender and age of people involved, in accordance with current Italian legislation, specifically articles number 582, 583 and 590 of the Italian Penal Code 2015 (Repubblica Italiana 2015). The Italian law considers a road accident to be a crash when it results in at least one injury, and crash consequences are classified into the following five Severity Levels (SL):

- Very Slight Injuries (VSI), when the most seriously injured person has a prognosis of less 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 the event results in permanent damage to the brain or any body organ;
- Guarded Prognosis (GPR), if the doctor cannot determine the disability, and issues a report of "guarded prognosis" (pending resolution of prognosis, the road crash must be considered and treated as a determining factor); and
- fatalities (FAT), including injured persons who die within 30 days of the crash.

The dearth of information in the ISTAT database was overcome by including crash data collected by Turin's Municipal Police (TMP). All the records of the ISTAT database were matched up to the TMP database and implemented with the following information: (a) historical data (time, nearest minute, day, month and year of the crash event); (b) locality data (street name, house number); and (c) generic information concerning crash SL.

Table 1 shows the number of crashes per year and SL, and evidences the decrease in all the SL classes between 2006 and 2012. Assuming the year 2006 as a reference point, the following years witnessed a decrease in crash occurrence across all severity classes.

2.2 *Traffic data*

Traffic data were provided by the 5T Company (Telematics, Technologies for Traffic and Transport in Turin), which monitors and controls over 300 urban traffic lights in Turin, and collects traffic data. 5T uses induction-loop traffic sensors located along the exiting lanes of the monitored intersections to collect vehicle flow data at 5 minute intervals. It is worth noting that from 2006 to 2012, the number of traffic sensors available varied from 662 to 1051 due to the installation of new ones and the elimination of some of the damaged ones. Figure 1a shows the portion of the road network monitored by 5T in 2006.

2.3 *Weather data*

The Environmental Protection Agency of the Piedmont Region (ARPA Piedmont) provided data on weather conditions. The Turin weather station considered in the paper is located in the city centre (238 m a.s.l., 1.5 m off the ground, latitude 45°.066667, longitude 7°.683333), and collects temperature, atmospheric pressure, wind speed and direction, solar radiation, and rainfall intensity data on an hourly basis. The maximum distance between the weather station and the farthest crash location included in the database was 9.6 km. Each crash record was associated with the weather data recorded at the time of the crash.

2.4 *Database formation*

Only crashes that occurred along segments provided with valid and reliable traffic data were extracted from the main database and used. The database adopted for modelling is a subset (and

Table 1. Injury crash records along segments of the urban road network of Turin (Italy, 2006–2012).

Year	VSI	SLI	SEI	GPR	FAT	Total
2006	1317	173	36	33	22	1581
2007	1303	205	69	42	29	1648
2008	1128	152	41	33	12	1366
2009	1061	160	34	27	21	1303
2010	1198	169	46	31	19	1463
2011	1037	171	51	18	18	1295
2012	917	141	44	13	14	1129
2006-'12	7961	1171	321	197	135	9785

Figure 1. (a) Turin's traffic monitoring network operated by 5T in 2006 (highlighted in black); (b) spatial distribution of road crashes that occurred in 2006; and (c) crashes which were associable to 5 min traffic flow (487 in total).

a random sample) of the total number of crashes that occurred and were recorded in the official database. Figure 1b shows all the crash records in 2006, while Figure 1c shows only crashes associated with traffic flow data.

3 DATA ANALYSIS AND TREATMENT

3.1 *Variables*

Table 2 lists the independent variables, their numbering and labels, the type of variable, the unit of measurement, and the range. The variables referring to the road are:

– road type (C1), which indicates 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), which have been distinguished with a numerical variable indicating the presence of water, snow or ice (0 = unknown, 1 = dry, 2 = wet, 3 = slippery, 4 = icy/frozen, 5 = snowy); and
– the road signage (C3), which indicates if it was absent (0), if it was 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 variables that reflect the characteristics of vehicles A and B involved in the crash are:

– vehicle type (C4 and C6), ranging from 0 (passenger cars) to 20 (quad), also including the case of vehicles that fled the crash scene (19); and
– vehicle category (C5 and C7), 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 vehicles.

In modelling, variables describing roads and vehicles were assumed as categorical. The variables describing drivers involved in the crashes were assumed as numerical. They are:

– age (C8 and C11), which ranges from a minimum of 10 (driver B) to a maximum of 89 (driver A); this variable also assumed the null value in cases of unknown/unrecorded age;
– age class (C9 and C12), 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
– sex of drivers (C10 and C13), which assumes the value 0 in cases of unknown/unrecorded data, 1 for males, and 2 for females.

In Table 2, the lowest values for 'age of driver A' refer to scooter drivers, while those for driver B refer to pedestrians or cyclists. Air temperature (C14), wind speed (C15), solar radiation (C16), and rainfall precipitation (C18) were assumed as numerical with values that correspond to the measured values. The lighting condition (C17) was assumed as a Boolean variable (0 = dark, 1 = light). The Traffic Flow (TF) variables (C20 ÷ C25) are numerical and represent the volume of vehicles per hour (veh/h) measured every 5 min across the crash event, according to the time scale reported in Figure 2. Finally, the standard deviation (C26) for the seven flow values was added to the list to take into account flow fluctuations for the 35 min period before and after the crash. Finally, the output variable indicating the SL (C27) was assumed numerical and ranging from 2 (VSI) to 6 (FAT).

Table 2. Number and labels of variables.

#	Code	Description	Type	u.m.	Range min	Range max
1	C1	Road type	C	_	0	4
2	C2	PC	C	_	0	5
3	C3	Road signage	C	_	0	4
4	C4	Veh. A type	C	_	0	20
5	C5	Veh. A cat.	C	_	0	8
6	C6	Veh. B type	C	_	0	20
7	C7	Veh. B cat.	C	_	0	8
8	C8	Dr. veh. A age	N	_	16	89
9	C9	Dr. veh. A cl. age	N	_	0	5
10	C10	Dr. veh. A sex	N	_	0	2
11	C11	Dr. veh. B age	N	_	10	86
12	C12	Dr. veh. B cl. age	N	_	0	5
13	C13	Dr. veh. B sex	N	_	0	2
14	C14	Air temp.	N	°C	–7.5	+35.3
15	C15	Wind speed	N	m/s	0	9.95
16	C16	Light radiation	N	W/m²	0	996
17	C17	Light/dark	B	_	0	1
18	C18	Rainfall	N	mm/h	0	12.8
19	C19	TF1 (*)	N	veh/h	0	730
20	C20	TF2 (*)	N	veh/h	0	750
21	C21	TF3 (*)	N	veh/h	0	765
22	C22	TF4 (*)	N	veh/h	0	775
23	C23	TF5 (*)	N	veh/h	0	570
24	C24	TF6 (*)	N	veh/h	0	565
25	C25	TF7 (*)	N	veh/h	0	494
26	C26	Flow st. dev.	N	veh/h	0	193
27	C27	SL	N	–	2	6

Notes: PC = pavement conditions, TF = traffic flow, Dr. = driver, veh. = vehicle, cl. = class, N = numerical, B = Boolean, C = categorical.

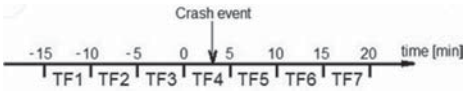

Figure 2. Time scale used to aggregate Traffic Flows (TF) across the crash event.

A criticism may be made of the use of all seven flow values (TF1-TF7). In fact, some of these values (in particular TF5-TF7) refer to intervals after the crash and therefore were caused by the crash itself. Nevertheless, the reason for using them is that they belong to the same time series and the interrupted nature of flow in urban roads makes each interval (though not as long as 5 minutes) a story apart, even when there is no crash. In addition, results will show that they play a different role in the models.

3.2 *Database information content*

The Principal Component Analysis (PCA) (Lebart et al. 1977) was used to investigate the information content of the database. Table 3 reports the variance explained by the first eight components. They account for about 81% of the total variance for both databases, while the first two components account for about 61%. The variables most linked to the first component are road type, road signage, and age of driver A, whereas those linked to the second component are light/dark, light radiation, and air temperature. Finally, traffic flows (TF1-TF7) are mainly linked to the third component. This means that the set of variables relating to road and driver can explain about 45% of the variance; those related to meteorological conditions about 16%, and those related to flow about 8%.

3.3 *Data treatment*

According to Table 1, the five SLs contained in the database were not equally represented and the dataset resulted imbalanced. This is not a problem for modelling approaches such as logistic regression, but it is for machine learning tools, and especially Artificial Neural Networks (ANN). With imbalanced datasets, ANN could not find the correct relationships between input and output for all categories present in the dataset. Over-sampling (with data duplication) or under-sampling (with data cancellation) techniques present advantages and disadvantages: under-sampling can remove important data and over-sampling can lead to over-fitting problems. Studies on imbalanced datasets have shown over-sampling to be more advantageous and useful than under-sampling (Chawla 2010).

The "focused re-sampling" method proposed by Japkowicz (2000), which consisted of an oversampling of those examples that occurred in the minority classes (specifically FAT, GPI, and SEI), was used. This approach implies duplication of the entire subset of data until their count is of the same magnitude as the most populated class. This approach avoids other possible biases in re-sampling data.

Another task performed was data normalization. Feature scaling, also called unity based normalization, was used for its simplicity. Let X_{min} and X_{max} be the two extreme values (minimum and maximum) of a variable X, the normalized variable X' (according to feature scaling) is:

Table 3. Percentage of variance explained by the first eight components in PCA.

Component	Simple value	Cumulative value
1	44.72	44.72
2	16.34	61.06
3	8.32	69.38
4	6.72	76.10
5	4.91	81.01
6	4.38	85.38
7	3.49	88.88
8	2.28	91.16

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

All the input variables were then normalized according to eq. 1, hence with values falling within the range [0,1]. The output variable (the output) is numerical, ranging from 2 to 6.

4 DATA MODELLING AND RESULTS

4.1 *Back-Propagation Neural Networks (BPNN)*

The BPNN used in this work is an example of an Artificial Neural Network (ANN) model, ANN models have a classical multilayer topology with feed-forward connections. Cybenko (1989), and Hornik (1991), described the capability of ANN in approximating any function belonging to the Lebesgue two space (L^2 space) with minimum error. Applications regarding transport, planning, control fields, and crash analysis are numerous starting from the 90's (Dougherty 1995, Mussone 1999, Mussone et al. 1999). Other contributions have faced the problem of crash prediction or severity (Abdelwahab & Abdel-Aty 2001, Chong et. al. 2004, Delen et al. 2006, Baluni & Raiwani 2014).

The downside in using the BPNN approach is that the relationships between variables are in a black box (the hidden layer of Figure 3), and no analytical formulation between input and output can be directly obtained. The effects of independent (input) variables can be interpreted only through a sensitivity analysis of the model.

The BPNN models were calibrated and validated with the Levenberg-Marquardt training algorithm. Performances were evaluated according to Mean Squared Errors (MSE) through the three phases of train, test, and validation. The model was constructed with an input layer including the 26 independent variables listed in Table 2, the hidden layer, and the output layer corresponding to the SL, tested with one neuron. All categorical variables are coded in binary format to reduce connection between their values. Finally, the best model was found to be made up of 25 neurons in the hidden layer for the model. It has a MSE lower than 0.08, which means that there are only 8 errors to each 100 classifications.

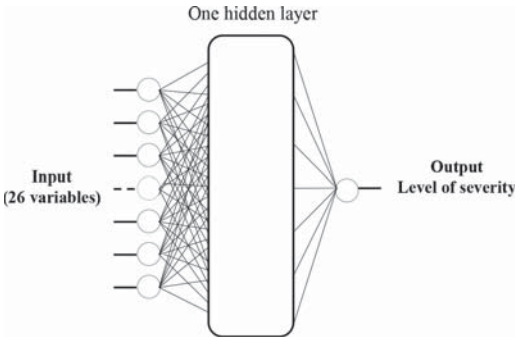

Figure 3. Back-propagation Neural Network structure for SL modelling adopted in this investigation.

4.2 *Generalized Linear Mixed Model (GLMM)*

For the analysis of multilevel data, random clusters and/or subject effects should be included in the regression model to account for the correlation of data. The resulting model is a mixed model including fixed and random effects. Mixed models for continuous normal outcomes have been proposed for non-normal data and are generically classified as Generalized Linear Mixed Models (GLMMs). The extension of the methods from dichotomous responses to ordinal response data was actively pursued in the reviews of Agresti & Natarajan (2001).

The GLMM model is a regression model of a response variable that contains both fixed and random effects and comprises data, a model description, fitted coefficients, co-variance parameters, design matrices, residuals, residual plots, and other diagnostic information. Fixed-effects terms usually refer to the conventional linear regression part of the model. Random effects 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 random effects have prior distributions, whereas the fixed effects do not.

The GLMM model structure is:

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

$$g(\mu) = \beta X + bZ + \delta \quad (3)$$

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

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

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

$$\eta = \beta X + bZ + \delta \quad (5)$$

According to the Wilkinson notation, the GLMM model has the following structure:

$$y \sim \text{fixed} + (\text{random1}|\text{group1}) + \ldots + (\text{random N}|\text{group N}) \quad (6)$$

where, the terms "fixed" and "random" are associated with independent variables and contain fixed and random effects, N = number of grouping variables in the model. Grouping variables are utility variables used to group, or categorize, observations, and are useful for summarizing or visualizing data by group.

For the SL output, a log link function and the Probability Mass Function (PMF) for the Poisson distribution was used. The fit method was the 'Laplace' one. Finally, the best performance was calculated through the minimization of the log-likelihood index; other indexes (i.e., Akaike's information criterion—AIC, Bayesian information criterions—BIC, and the Deviance parameter) were also estimated to control the minimization process.

According to eq. 6 notation, the GLMM model that gives the best performance was:

$$C27 \sim 1 + C8 + C9 + C11 + C14 + C16 + C18 + C21 + C24 + C25 + (1 \mid C1) + (1 \mid C2) + (1 \mid C3) + (1 \mid C4) \quad (7)$$

where Cx identifies the x-th variable (Table 2). In Table 4 the fixed effect coefficients are drawn with a 95% confidence interval. The p values are lower than 0.001, with the exception of C9 (driver vehicle A age) and C18 (rainfall intensity) which are lower than 0.05. The Standard Error of estimates (SE) is generally much lower than the estimates, and lower and upper bonds of CI never include zero. Table 4 also reports the estimates for random parameters.

The effect of flows (C21, C24, C25) on SL has a different sign, positive for C21 (TF4), which anticipates the crash event, and for C24 (TF6), and negative for C25 (TF7). When flow after the crash (C25) increases, it is more likely that the SL decreases. The grouping variables are road type (C1), pavement condition (C2), road signage condition (C3) and vehicle A type (C4). The driver vehicle A age (C9) as well as the class age of vehicle B driver (C11) are negatively related to the SL. Furthermore, also light radiation (C16) and rainfall intensity (C18) are inversely related to the SL.

5 DISCUSSION

5.1 *Sensitivity analysis of BPNN*

In the case of the BPNN model, a sensitivity analysis was carried out to assess how output changes by varying input normalized variable values in the range [0,1] one by one. With this aim in mind, a first set of scenarios, referring to a particular set of input variables, was prepared. In addition to basic scenarios where variables are all zero or 1, another six scenarios were considered to study particular combinations of variable values (in Figure 4, from 4a to 4f). These scenarios aim to consider some possible and typical crash situations involving male or female drivers, during daytime or at nighttime, with rainy weather or dry road surface, or with elder drivers.

The effect of flow varies a lot for scenario and flow itself. When TF4 is high (Figure 4a), the SL is generally high for most of the scenarios, except for elderly drivers. On the other hand, the effect of TF7 (Figure 4b) depends very much on the particular scenario, though, generally, a higher SL is related to a higher flow. Low light radiation (nighttime to dawn) has a strong effect on young male drivers with dry pavement, while middle radiation has a strong effect on young female drivers in rainy conditions. Generally, a low radiation is more related to high SL than a high radiation.

5.2 *Model output comparison*

According to Powers (2011), the two model outputs were evaluated by confusion matrixes representing, for each output, the number of predicted cases (a_{ij}) on the reference databases. In this case, the output coincides to the SL, and a_{ij} are calculated on the resampled databases obtained according to what reported in Section 3.3.

There is s also interest in the measurement of its precision (the percentage of correct data predicted in respect of the total predicted) and its recall (the percentage of corrected data predicted in respect of the total to be predicted) capability. The main goal of learning is to improve the recall measurement without hurting the precision one. Tables 5 and 6 include the percentage of the predicted crashes to the total predicted for each SL, the "a priori" rate (PR), which expresses 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 (n is the matrix dimension):

$$PR_i = 1 - a_{ii}/(a_{i1} + \ldots + a_{in}) \quad (8)$$

$$PO_i = 1 - a_{ii}/(a_{1i} + \ldots + a_{ni}) \quad (9)$$

Furthermore, comments on results are supported by the estimation of their accuracy (A):

Table 4. Fixed effects coefficients estimates and Random effects covariance parameters at 95% CIs) for the GLMM segment model.

Variable	Estimate	SE	p-value	Lower	Upper
Intercept	1.24210	0.2019	$<10^{-3}$	0.84629	1.63790
C8	0.00594	$<10^{-3}$	$<10^{-3}$	0.00428	0.00759
C9	–0.06953	0.018	$<10^{-3}$	–0.10540	–0.03367
C11	–0.00620	$<10^{-3}$	$<10^{-3}$	–0.00707	–0.00534
C14	0.00497	$<10^{-3}$	$<10^{-3}$	0.00318	0.00675
C16	–0.00016	$<10^{-4}$	$<10^{-3}$	–0.00023	–0.00010
C18	–0.07152	0.0203	$<10^{-3}$	–0.11142	–0.03163
C21	0.00056	$<2{\cdot}10^{-4}$	0.002	0.00020	0.00092
C24	0.00111	$<3{\cdot}10^{-4}$	$<10^{-3}$	0.00062	0.00161
C25	–0.00242	$<3{\cdot}10^{-4}$	$<10^{-3}$	–0.00284	–0.00199

Group variable	Estimate
C1 (Intercept)	0.13646
C2 (Intercept)	0.38224
C3 (Intercept)	0.19893
C4 (Intercept)	0.11375
Indexes	
LogLikelihood	–11465
AIC	22959
BIC	23053
Deviance	22931
R^2 adjusted	0.3107

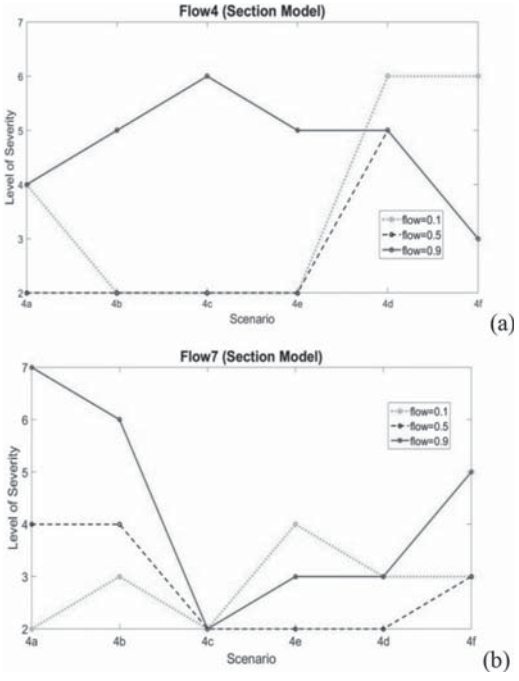

Figure 4. Effect of TF4 (a) and TF7 (b) (variables 22 and 26) on SL (BPNN model) for different flow values (0.1, 0.5, 0.9).

$$A = (a_{11} + a_{22} + \ldots + a_{nn})/\Sigma a_{ij} \quad (10)$$

Table 5 reports the confusion matrices for the model calibrated through the BPNN. 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 tables considering that the model output can fall outside the range of numerical values associated with each SL. The accuracy of 90% is certainly very high for BPNN model. PR and PO rates are low with the exception of SL 2 and 3. SL 2 is the more difficult to predict while SL 3 has the largest number of wrong cases assigned to it.

Table 6 contains the confusion matrices for the GLMM model. In this case, the capacity of SL prediction is significantly lower than the one for BPNN as indicated by the accuracy of 33%. GLMM has a superior capacity to provide results within the SL limits of 2 and 6, as confirmed by the absence of values that fall outside of the two limits. PR and PO rates are lower than the corresponding values for BPNN, showing a greater difficulty in predicting SL than the neural network modelling approach.

Comparisons with GLMM show a marked superiority of BPNN modelling as regards performance measured through confusion matrices. GLMMs, on the other hand, clearly show what variables are significant and their effect (sign and value of coefficients) though this is limited to the

Table 5. BPNN model confusion matrix for segments (row percentage values in brackets), and "a Priori" (PR) and "a Posteriori" (PO) rates.

	Predicted SL								
Real SL	<2	2	3	4	5	6	> 6	CD	PR (%)
2	21 *2%*	859 *65%*	257 20%	71 *5%*	39 *3%*	60 *5%*	7 *0%*	1314	35%
3	18 *1%*	150 *12%*	1056 *86%*	6 *1%*	0	0	0	1230	14%
4	0	0	0	1250 *94%*	0	0	0	1250	0%
5	0	0	0	0	1287 *100%*	0	0	1287	0%
6	0	0	0	0	0	1311 *100%*	0	1311	0%
PO	–	15%	20%	6%	3%	4%	–	–	–

Notes: CD = crash data in the resampled database.

Table 6. GLMM model confusion matrix (row percentage values in brackets), and "a Priori" (PR) and "a Posteriori" (PO) rates.

	Predicted PSL								
Real RSL	<2	2	3	4	5	6	>6	CD	PR (%)
2	0	79 *52%*	534 *41%*	593 *45%*	100 *8%*	8 *1%*	0	1314	94%
3	0	72 *48%*	444 *36%*	588 *48%*	102 *8%*	24 *2%*	0	1230	64%
4	0	0	300 *24%*	850 *68%*	100 *8%*	0	0	1250	32%
5	0	0	393%	702 *55%*	507 *39%*	39 *3%*	0	1287	61%
6	0	0	0	483 *37%*	621 *47%*	207 *16%*	0	1311	84%
PO	–	48%	66%	74%	65%	26%	–	–	–

linear effect of variables without considering their possible reciprocal interaction.

6 CONCLUSIONS

The paper aims to achieve two goals: the evaluation of the crash Severity Level (SL) on urban road segments using environmental variables (some of which, like short-term flow, are innovative for this type of research), and the comparison of two different techniques for calculating SL, the Back-Propagation Neural Network model (BPNN) and the Generalized Linear Mixed Model (GLMM).

The results presented here provide new insights into urban roads and fill a gap in the knowledge acquired from the number of studies on rural freeways and expressways reported in literature.

From the use of the confusion matrixes technique, BPNN models evidenced their superiority in the prediction of the SL when compared to the GLMMs. This is attributable to their greater capability of accurately approximating any continuous and non-linear function. On the other hand, GLMMs (like any analytical model) allow a readier interpretation of model results. Other pros and cons in their use derive from the intrinsic characteristics of statistical and neural network methods, as clearly underlined by Karlaftis and Vlahogianni (2011). The authors suspect that the most significant limit of GLMMs for these applications is related to the constrained linearity of their functions. In addition, missing data may have contributed to the fact that the BPNNs, which are known to be capable of overcoming this problem, achieved better results.

However, both approaches (BPNN and GLMM), though with significant differences, indicate that flows have a relevant role in predicting severity: this role is not limited to the flow when the crash occurred (TF4), but also involves other flow

data recorded before (TF1-TF3) and after (TF5-TF7) the crash. GLMM model shows the relevance of TF3, TF6, and TF7 only, but the BPNN model evinces more complex relationships for all seven variables. Weather variables also (i.e., rainy condition and light radiation) show a strong relation in some scenarios.

In future research, generalized non-linear models will be used to consider the higher order effects and the interaction between variables. Moreover, a mixed approach using both short-term flow and AADT values will be investigated to derive models over a mid-to-long term period and investigate the relationship between them.

ACKNOWLEDGEMENTS

The authors thank *Polizia Municipale di Torino*, and the *Città Metropolitana di Torino* 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 weather data.

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.

Agresti, A. & Natarajan, R. 2001. Modelling clustered ordered categorical data: a survey. *Int. Stat. Rev.* 69, 345–371.

Al-Ghamdi, A.S. 2002. Using logistic regression to estimate the influence of accident factors on accident severity. *Accid. Anal. Prev.* 34, 729–741.

Baluni, P. & Raiwani, Y.P. 2014. Vehicular accident analysis using neural networks. *Int. J. of Emerg. Tech. and Adv. Engin.* 4 (9), 161–164.

Chawla, N.V. 2010. Data mining for imbalanced datasets: an overview. *Data Mining and Knowledge Discovery Handbook*, Springer US, 875–886.

Chong C-Y. & Kumar, S.P. 2003. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE* 91(8), 1247–1256.

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.

Cybenko, G. 1989. Approximation by superpositions of sigmoidal functions. *Math. Control Signals Syst.* 2(4), 303–314.

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.

Dougherty, M. 1995. A review of neural networks applied to transport. *Transp. Res. Part C: Emerg. Tech. 3*(4), 247–260.

El Faouzi, N-E., Leung, H. & Kurian, A. 2011. Data fusion in intelligent transportation systems: progress and challenges—a survey. *Inform. Fus.* 12(1), 4–10.

Golob, T.F., Recker, W.W. & Pavlis, Y. 2008. Probabilistic models of freeway safety performance using traffic flow data as predictors. *Saf. Sci.* 46, 1306–1333.

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

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: Em. Tech.*, 19(3), 387–399.

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

Mussone, L. 1999. A review of feedforward neural networks in transportation research, *e&i Elektrotechnik und Informationstechnik* 116(6), 360–365.

Mussone, L., Ferrari, A. & Oneta, M. 1999. An analysis of urban collisions using an artificial intelligence model. *Accid. Anal. Prev.* 31, 705–718.

Nekovee, M. 2005. Sensor networks on the road: the promises and challenges of vehicular ad hoc networks and vehicular grids. *Proc. Work. on Ubiq. Comp. e-Res.*, Edinburgh, Scotland, UK.

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

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.

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

Wang, C., Quddus, M.A. & Ison, S.G. 2013. The effect of traffic and road characteristics on road safety: a review and future research direction. *Saf. Sci.* 57, 264–275.

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.

Yu, R. & Abdel-Aty, M. 2013. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accid. Anal. Prev.* 62, 161–167.