Re-visiting crash–speed relationships: A new perspective in crash modelling

Maria-Ioanna M. Imprialou, Mohammed Quddus, David E. Pitfield, Dominique Lord

Article history:
Received 21 April 2015
Received in revised form 20 August 2015
Accepted 1 October 2015

Keywords:
Traffic speed
Crash frequency
Crash severity
Pre-crash conditions
Multivariate Poisson lognormal regression
Multivariate spatial correlation

Abstract

Although speed is considered to be one of the main crash contributory factors, research findings are inconsistent. Independent of the robustness of their statistical approaches, crash frequency models typically employ crash data that are aggregated using spatial criteria (e.g., crash counts by link termed as a link-based approach). In this approach, the variability in crashes between links is explained by highly aggregated average measures that may be inappropriate, especially for time-varying variables such as speed and volume. This paper re-examines crash–speed relationships by creating a new crash data aggregation approach that enables improved representation of the road conditions just before crash occurrences. Crashes are aggregated according to the similarity of their pre-crash traffic and geometric conditions, forming an alternative crash count dataset termed as a condition-based approach. Crash–speed relationships are separately developed and compared for both approaches by employing the annual crashes that occurred on the Strategic Road Network of England in 2012. The datasets are modelled by injury severity using multivariate Poisson lognormal regression, with multivariate spatial effects for the link-based model, using a full Bayesian inference approach. The results of the condition-based approach show that high speeds trigger crash frequency. The outcome of the link-based model is the opposite; suggesting that the speed–crash relationship is negative regardless of crash severity. The differences between the results imply that data aggregation is a crucial, yet so far overlooked, methodological element of crash data analyses that may have direct impact on the modelling outcomes.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The primary objective of developing a traffic crash model is to elucidate the association between crashes and their potential contributory factors so as to formulate efficient and targeted crash mitigating measures. The accuracy of the modelling outcomes is therefore critical for inappropriate decisions to be avoided. Motorway crashes appear to have a decreasing trend, especially in western countries; however, the number of casualties is still anything but negligible (IRTAD, 2014; WHO, 2013). The question then arises: are the crash models we currently use accurate enough to develop appropriate preventive measures?

Each crash is the outcome of a unique sequence of events related to the involved driver(s), vehicle(s) and the road environment. The in-depth examination of individual crashes one-by-one though, is rarely possible due to the limited data availability. As a consequence, the crashes of a road network are usually analysed in a way that their volume is reduced while they remain informative (Lord and Mannering, 2010). The main crash aggregation method is based on topological and temporal criteria. In the so-called link-based (or segment-based) approach the counts of crashes that occurred on pre-defined road links during a certain time period are modelled against explanatory variables that represent the average conditions on each link (e.g. speed, traffic flow, road geometry). The explanatory power of these approaches, in terms of statistical methodology, has evolved over the years, reaching high levels of sophistication and offering better understanding of traffic crashes (e.g. Abdel-Aty and Radwan, 2000; Lord and Mannering, 2010; Ma et al., 2008; Mannering and Bhat, 2014). Despite the fact that the link-based approach is straightforward and simple, it is also by default linked with aggregation problems or else with the information loss that is aroused when multiple values are represented by a single measure (Black et al., 2009; Clark and Avery, 1976; Davis, 2004). This limitation may affect the models’ explanatory potential, especially...
for time-varying independent variables (e.g., speed, traffic volume) as their spatial and temporal variations within a link cannot be captured.

Speed is regarded as one of the main traffic related crash contributory factors (Abdel-aty et al., 2005; Elvik et al., 2004), but research findings do not confirm this unanimously. The inconsistency between the results could be partially due to the inadequacy of annual average speed to represent the speeds at which crashes actually occurred. In fact, two crashes recorded on the same link may have occurred under entirely different traffic conditions but in a link-based approach they will be both explained by the annual average speed on the link. This can be further explained by Figs. 1 and 2. Fig. 1 shows the frequency and the cumulative distribution of the ratio of the actual speed at the crash location to the annual average speed on the corresponding road link for all 2012 motorway crashes in England. Fig. 2 is the same for traffic volume. It is obvious that the ratios are considerably different from one for a high proportion of crashes (ratio = 1 means that crash speed or volume is equal with the respective annual average), confirming that the representation of time-varying measures by annual averages is rather inadequate in many cases.

This paper introduces a new crash data aggregation concept termed as condition-based approach that aims to represent in more detail the actual pre-crash conditions in order to explore the relationship between motorway crashes and their contributory factors such as speed, volume and geometric configuration. The grouping attribute of crashes in the proposed method is the similarity of pre-crash conditions rather than a link-level spatial relationship. In this way, crash counts are represented more precisely by explanatory variables that approximate the actual conditions enabling, possibly, improved relationships. The condition-based dataset can be modelled using multivariate Poisson lognormal regression. In order to compare the two methods with respect to their outcomes, the same data are also used to build a link-based spatial multivariate Poisson lognormal regression model.

2. Literature review

A considerable amount of literature has been published on the impact of various traffic and geometric road characteristics on link-based crash frequency. Among others speed, traffic volume, number of lanes, gradient and horizontal curvature are widely studied. From a qualitative point of view, findings show that although crash severity is positively correlated with driving speed (Clarke et al., 2010; Joksch, 1975; Kloeden et al., 1997; Pei et al., 2012), the relationship between speed and crash frequency is not equally straightforward (Aarts and Van Schagen, 2006). The early study of Solomon (1964) was the first to suggest that speed and crash frequency are not proportional but their relationship can be described by a “U-shaped” curve; an idea that was supported by several other researchers (e.g., Munden, 1967; Cirillo, 1968). Solomon’s curve implies that only extremely low and high speed conditions trigger crashes. However, most of the subsequent studies find driving speeds to be linearly or exponentially related to crashes (Baruya and Finch, 1994; Fildes et al., 1991; Kloeden et al., 2002, 1997; Taylor et al., 2000). A few studies contradicted this view proposing that the speed–crash relationship is negative (Baruya, 1998; Stuster, 2004) and others reported that this relationship is statistically insignificant (Garber and Gadiraju, 1989; Lave, 1985). Some of the most recent papers that employed advanced statistical models did not find statistically significant relationships between speed and crashes (Kockelman and Ma, 2007; Quddus, 2013). Pei et al. (2012) attempted to explain the results’ inconsistencies suggesting that the crash–speed relationship that is estimated by models strongly depends on the selected measure of exposure; the relationship was shown to be negative for distance-based exposure (i.e., vehicle miles travelled) but positive for time-based exposure (i.e. vehicle hours travelled).

The relationship of speed with crashes cannot be defined without taking into account the simultaneous effect of other traffic characteristics such as traffic flow (Aarts and Van Schagen, 2006) and vehicle occupancy (Garber and Subramanyan, 2001; Lord et al., 2005a). High traffic flow (represented by AADT, hourly volume, etc.) is generally considered to increase the risk of crashes (Abdel-Aty and Radwan, 2000; Chang, 2005; Milton and Mannering, 1998). On the contrary, lower flows have been also correlated with higher speed variance that is also considered to be a significant crash precursor (e.g., Garber and Ehrhart, 2000; Elvik et al., 2004). The mechanism of its impact though is not explicitly explained because of the lack of individual vehicle-level second-by-second data that would lead to reliable estimations. Instead, current studies employed relatively highly aggregated data that lead to inconclusive results (Garber and Ehrhart, 2000; Kockelman and Ma, 2007; Quddus, 2013; Solomon, 1964). Although seldom researched, vehicle occupancy ratio was found to have a non-linear relationship with the number of crashes (Garber and Subramanyan, 2001) and was also dependent on the number of vehicles involved in the crash (i.e., single- versus multi-vehicle crashes) (Lord et al., 2005a).

Road geometric design is also believed to be related with crash frequency on the roadway (AASHTO, 2010). High crash frequency is associated with high vertical grades (Anastasopoulos...
Road Accident Database of the United Kingdom (STATS 19) and include 10,520 crashes that occurred during 2012 on the Strategic Road Network (SRN) of England (Department for Transport, 2011). The SRN consists of the main motorways and A-roads of the country and the total length is 6920 km (4272 miles). STATS 19 reports record crashes that accounted for at least one slight injury, along with information related to the crash and the involved parties (i.e., drivers, casualties and vehicles). The variables that were used here were crash severity, date, time and location.

Location is a crucial factor for crash analyses because it is closely related with the identification of the traffic and geometric conditions that are related to a crash. When crash location data are not satisfactory in terms of accuracy, the application of crash mapping techniques has been shown to significantly change the results of crash analyses (Imprialou et al., 2015). The objective of crash mapping is to determine a set of coordinates that represent the crash location as precisely as possible. In STATS 19 reports, crash locations were less accurate than desired. Thus, crashes were reallocated to refined positions estimated by a fuzzy logic crash mapping algorithm that was developed for the study area using distance, vehicle direction, road name and type. This provides a 98.9% (±1.1%) accurate matching score (Imprialou et al., 2014).

Traffic data were extracted from the UK Highways Agency Journey Time Database (JTDB) that includes link-level traffic information obtained by inductive loop detectors for the entire SRN (2505 links1) (Highways Agency, 2011). The measurement interval is 15 min resulting to a dataset of approximately 88 million observations. The variables used for this analysis are average speed (km/h), volume (vehicles) and travel time (seconds) (Highways Agency, 2011). Road configuration was determined based on the UK Highways Agency Traffic Speed Condition Survey database (TRACS) (Highways Agency, 2008). TRACS contains measurements of geometric characteristics (i.e., radius and gradient) by survey vehicles for the entire SRN divided into 10-m segments.

The data were processed separately in order to produce the datasets. Although the two datasets stem from exactly the same databases, they represent the relationship of crashes with road-related variables from entirely different perspectives and sampling frames. The sampling frame of the link-based dataset consists of road links that are actual spatial entities and is the conventional approach for safety models. The sampling frame of the condition-based dataset comprises of all the possible combinations of traffic and geometric conditions; a set of abstract/non-physical attributes that can potentially co-exist at the time and the location of a crash.

### 3.1. Link-based dataset

A link-based dataset enlists the links that comprise a road network and the total number of crashes per link. The crashes occurred on the link at different time points during the study period. Each link contains information that represents the conditions on the road defined by descriptive statistics (e.g. mean, median, maximum, etc.). Based on this aggregation method, it is assumed that the triggering factors for crashes that occurred on the same link are similar, which of course might be not true for all the cases as shown in Figs. 1 and 2.

Based on the output of the crash mapping algorithm, each road link was assigned with a number of crashes (crash counts varied from 0 to 36 per link) and one characteristic value representing speed, volume, curvature, gradient and the number of lanes. Considering the dynamic nature of the traffic variables (i.e., speed and volume) as well as the fact that a road link typically covers a

---

1 Average link length 5.23 (±4.76) km.
Considerable road length, it can be understood that both the traffic conditions and the geometric configuration of each link can only be partially represented by simple measures per link. Traffic conditions were expressed by annual averages, while road geometry was represented by categorical variables. A more detailed description of the variables can be found in Table 1. After the exclusion of the links with missing traffic or geometry data the final link-based dataset included 2356 observations (i.e., links) that represent overall 9028 crashes. Crash counts were divided by severity into crashes with Killed or Serious injuries (henceforth: KS) and crashes with Slight injuries (henceforth: SL). The split between the two severity categories was 1268 and 7760 for KS and SL crashes, respectively.

3.2. Condition-based dataset

A pre-crash condition-based dataset (henceforth: condition-based dataset) consists of every possible combination/scenario of traffic and geometric conditions that could ever be present on the network just before a crash (limited to the examined variables and their specifications). Each scenario is matched with a number of crashes (from zero to, theoretically, all the crashes of the database) that were found to occur under this particular combination of traffic and geometry conditions. Condition-based modelling attempts to represent the actual crash-related traffic and geometry conditions. In contrast to the link-based approach, the crashes that belong to the same condition scenario are spatially and temporally independent. Instead, they are similar in the sense that when they occurred the external circumstances on the road were approximately the same. Assuming that some or all of these circumstances might be related with the crash occurrences, the concentration (or absence) of crashes in some particular combinations should provide useful insights about crash triggering factors.

The formation of the condition-based dataset is quite complex relatively to the link-based dataset. Fig. 3 presents a simple flowchart describing the main processes to develop the condition-based dataset consisting of N_{\text{max}} crashes. Each step is explained in detail below.

### Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Link-based dataset</th>
<th>Condition-based dataset</th>
</tr>
</thead>
</table>
| Speed\(^a\) | Annual average of measured speeds on each link (averaged over 35,040 records) | S1. Speed up to 2nd percentile  
S2. Speed between the 3rd and the 4th percentile  
S3. Speed between the 5th and the 6th percentile  
\ldots | S50. Speed between the 99th and the 100th percentile  
Separate for each of the 50 speed scenarios:  
V1. Volume up to the 25th percentile  
V2. Volume between the 26th and the 50th percentile  
V3. Volume between the 51st and the 75th percentile  
V4. Volume over the 76th percentile |
| Volume\(^a\) | Annual average daily traffic per link (AADT) | C1. Segments that above 50% of their radius measurements are lower than 2000 m (Curve)  
C2. Segments that above 50% of their radius measurements are equal with 2000 m (Straight) |
| Curvature | C1. Links with multiple and/or sharp curves (Curve)  
C2. Links that above 50% of their radius measurements are equal with 2000 m (Straight) | C1. Segments that have more gradient measurements above 0.5% than below 0.5% (Uphill)  
C2. Segments that have more gradient measurements below −0.5% than above −0.5% (Downhill)  
C3. Segments that have more gradient measurements between ±0.5% and below 0.5% (Level) |
| Lanes | L1. Links that above 50% of their sections include more than two lanes (Lanes above 2)  
L2. Links that above 50% of their sections include up to two lanes (Lanes up to 2) | L1. Sections with more than two lanes (Lanes above 2)  
L2. Sections with up to two lanes (Lanes up to 2) |

\(^a\) Classification was based on the weighted speed and volume (S\(_w\) and V\(_w\); see Eqs. (1) and (2)).

#### 3.2.1. Traffic conditions identification

The final condition-based dataset included all the possible combinations of pre-crash-condition scenarios and the crash counts per scenario. As the scope of the creation of the alternative dataset was the representation of the conditions on the roadway just before crashes, all the examined crashes were matched with a set of traffic and geometric conditions based on the geocoded crash locations.

The pre-crash traffic conditions on the crash location were identified based on the reported crash date and time. In order to have a comparable set of measurements for all crashes, each crash was matched with traffic data equivalent to 15 min of measurements. Therefore, the speed (S\(_w\)) and volume (V\(_w\)) were estimated using a weighted average of the 15-min interval that includes the time of the crash (second interval) and its precedent (first interval).

\[
S_w = \left( \frac{t_{15}}{15} \right) S_{\text{second}} + \left( 1 - \frac{t_{15}}{15} \right) S_{\text{first}} \tag{1}
\]

\[
V_w = \left( \frac{t_{15}}{15} \right) V_{\text{second}} + \left( 1 - \frac{t_{15}}{15} \right) V_{\text{first}} \tag{2}
\]

where S\(_w\) and V\(_w\): weighted average of speed (km/h) and volume (vehicles), S\(_{\text{first}}\) and V\(_{\text{first}}\): speed (km/h) and volume (vehicles) measurements of the first interval, S\(_{\text{second}}\) and V\(_{\text{second}}\): speed (km/h) and volume (vehicles) measurements of the second interval, t\(_{15}\): time of the crash (15 min) relative to the start of the 15-min interval.
volume (vehicles) measurements of the second interval, \( t \): time difference between the start of the second interval and the reported crash time (min). It is a fact that the resolution of the traffic data is not ideal for defining the exact traffic conditions just before the crashes; within 15 min traffic conditions can change on the roadway. Even so, the traffic characteristics used here are significantly more representative than annual averages that are typically used for link-based analyses. Moreover, it should be noted that the reported time of crashes in the examined database tends to be rounded; an issue that is also reported by Kockelman and Ma (2007). In SIRTA19, crash time is reported with an hours-minutes format (i.e. HH:MM). Fig. 4 presents the distribution of the second part of the reported crash time (i.e. MM from 00 to 59) of the examined crashes. It can be seen that the distribution is clustered at the nearest 5’s. This data limitation shows that even if more disaggregated traffic data were available (e.g., 1-min resolution), it would be necessary to consider a wider temporal interval per crash so as to capture the error of reporting crash time.

### 3.2.2. Geometrical conditions identification

The configuration of the roadway a few metres before the crash location is probably also related with the crash occurrence. That is why the length of the road that was considered for each crash was defined by the stopping distance upstream of the identified crash location on the link. Stopping distance was estimated based on the annual average speed of motorways and A-roads separately using the following equation (Elvik et al., 2004):

\[
S_D = R_D + B_D = t_r n_0 + \frac{V_0^2}{2f_k g}
\]

(3)

where \( S_D \): stopping distance (m), \( R_D \): reaction distance (m), \( B_D \): braking distance (m), \( t_r \): reaction time (here: 1.5 s), \( n_0 \): average speed (m/s), \( V_0 \): average speed (km/h), \( f_k \): friction (here: 0.8, average tire on dry pavement), \( g \): gravity acceleration (here: 9.8 m/s²).

Based on the above equation, the stopping distance was estimated 97 and 75 metres for motorways and A-roads respectively. To correct for errors in the crash location, the final road segment for each crash included the length of the stopping distance upstream of the crash location and 20 m downstream (error distance). Fig. 5 is a schematic illustration of the road segment that is considered for obtaining the geometrical conditions of each crash. The final road segments included a number of successive radius and gradient measurements that were converted to categorical variables so as to keep the number of scenarios of the final dataset relatively low. Thus, crashes were considered to occur on curves if the majority (above 50%) of the radius measurements of the segment were less than 2000 m and on straight segments otherwise. Similarly, crashes that occurred on uphill segments were considered those that included more gradient measurements above 0.5% than below 0.5%, on downhill those that include more gradient measurements below −0.5% than above −0.5% and otherwise on level segments. The road width was represented with another dummy variable that separated road segments with more than two lanes from segments with up two lanes.

#### 3.2.3. Final condition-based dataset

After each crash was matched with a set of traffic and geometric pre-crash conditions, the initial 10,520 crashes of the database decreased to 9310 (1310 KS and 8000 SL crashes) due to missing or illogical values in one or more variables. Crashes left in the analysis were classified according to their prior conditions to a spreadsheet that included all the possible combinations of pre-crash conditions. Apart from the crash data, to generate a condition-based dataset it is necessary to employ all data that describe the conditions on the network. The scenarios of a condition-based dataset should represent all the condition combinations that existed on the network regardless of whether these were associated with crashes or not. That is why before generating a condition-based dataset the range and the distribution of the measurements of the explanatory variables that will be used should be known. The process of the development of this dataset might not be the only way for building a condition-based dataset. However, the presentation and comparison of different data combination methods fall out of the scope of this paper.

To facilitate controlling for the exposure, all the scenarios of the condition-based dataset should be chosen to have equal likelihood of occurrence during the examined study period. To do this, the continuous variables that were included in the dataset (i.e., speed and volume) were divided into equal frequency groups defined by percentile ranges with a constant step \( n \) (e.g. from the \( N \)th percentile to the \( N + n \)th percentile, from the \( N + n \)th percentile to the \( N + 2n \)th percentile, from the \( N + 2n \)th percentile to the \( N + 3n \)th percentile...). Each group was represented in the dataset by a representative value (i.e., a central tendency measure). In this way, for every continuous variable \( C_i \) there were a number of \( K_i \) equally likely distinct groups of observations (where \( K_i = 100/n \)). Every discrete variable \( D_j \) had by default a number of categories, \( L_j \). To develop a dataset that includes every possible variable combination the number of scenarios \( S \) that should be generated is:

\[
S = \prod_{i=1}^{l} K_i \prod_{j=1}^{L_j}
\]

(4)

The number of scenarios of the dataset can be empirically adjusted so as to serve the analyses needs by selecting a smaller step \( n \) that decreases the number of scenarios and vice versa.

Traffic characteristics were grouped into categories of equal frequency. The speed groups were defined by dividing the cumulative speed distribution of the entire network into 50 equal parts (i.e., \( K_{\text{speed}} = 50 \)) with a 2-percentile step (i.e., \( n_{\text{speed}} = 2 \)) (see Table 1). Following, the volume, for each speed group separately, was split into the quartiles of its cumulative distribution (i.e., \( K_{\text{volume}} = 4 \) and \( n_{\text{volume}} = 25 \)). The number of groups was decided to be higher for speed than for volume because this paper mainly focuses on the impact of speed on crashes. Some different combinations of numbers of groups for speed and volume that have been attempted (that are not presented here due to brevity) did not seem to significantly change the modelling outcomes. Speed and volume per group were represented by their medians. Other measures were also tested such as the mean and the 85th percentile that did not exhibit any statistical difference in the modelling results. To keep the number of combinations relatively low, all the geometric variables were represented by categorical variables. As it can be seen
in Table 1 curvature and lanes are divided into two categories (i.e. \( L_{\text{curvature}} = L_{\text{lanes}} = 2 \)) and gradient into three (i.e. \( L_{\text{gradient}} = 3 \)). Using Eq. (4) the number of scenarios \( S \) was estimated to be:

\[
S = K_{\text{speed}} \cdot K_{\text{volume}} \cdot L_{\text{curvature}} \cdot L_{\text{gradient}} \cdot L_{\text{lanes}}
\]

\[
= 50 \cdot 4 \cdot 2 \cdot 3 \cdot 2 = 2400
\]

Overall, the spreadsheet contained the 2400 unique combinations of pre-crash scenarios (e.g. speed is between the 40th and the 42nd percentile with the median value of 93 km/h, the volume is between the 50th and the 75th percentile for these speed conditions with median 112 veh/lane, on a straight and downhill section with up to two lanes). The distinct values of each categorical or continuous variable had equal frequency with the other values of this variable (e.g., 800 scenarios were on uphill segments, 800 scenarios on downhill and 800 scenarios on level). Each crash was classified to one of these scenarios with respect to its traffic and geometric conditions and the severity of its consequences. The final output of this process was a dataset with 2400 observations that represent all crash counts by severity (i.e., KS and SL). Table 2 presents the summary statistics of the explanatory variables of both the datasets.

### 3.3. Exposure

In order to enable meaningful comparisons in terms of crash risk between the observations of crash models it is necessary to take into account one exposure variable. The use of an offset in a count model indirectly transforms the dependent variable from a number of events to a rate of events per the exposure measure. Exposure in link-based approaches attempts to express the total amount of travel on each link. The most appropriate measures of exposure for link-based modelling have been broadly discussed in the literature (e.g. Qin et al., 2004; Pei et al., 2012; Lord et al., 2005a) as there is a plurality of surrogate measures of exposure such as link length, average annual daily traffic, vehicle-miles travelled, vehicle-hours travelled, etc. Link length, that is one of the most commonly used exposure variables in crash analyses, was employed for the link-based model in this paper.

The way of expressing exposure in a condition-based approach is similar, however not identical. The condition-based dataset that is developed here divides traffic conditions based on the percentiles of their occurrence on the entire network (Table 1). In other words, in terms of the traffic conditions, all scenarios had equal occurrence frequency on the study network during the study period. The fact that all the scenarios are equally likely to occur, though, does not mean that they have equal crash probability, so the exposure cannot be considered as uniform among condition scenarios. The probability of crashes is proportional with the probability of crash prone interactions between vehicles on the network (e.g. Chipman et al., 1992; Navon, 2003). The number of vehicle encounters at a particular condition scenario increases as the number of vehicles and the duration of their stay under these conditions raise. In order to control for this effect, the offset variable for the condition-based dataset was set to be the average vehicle-hours per kilometre for each scenario. Vehicle hours per kilometre were estimated by multiplying the mean of all the travel time per kilometre measurements of a scenario with the corresponding average volume.

### 4. Methodology

Despite the difference in data generating mechanism, both the link-based and the condition-based are count datasets. Poisson regression and its extensions is the most suitable family of models for modelling crash counts, in terms of statistical properties (Lord and Mamer, 2010). One of the ways to control for over-dispersion (i.e., variance of the dependent variable is higher than its mean), that appears practically to most count datasets due to heterogeneity, is to add a random effect to the Poisson regression model. When the Poisson parameter is lognormally distributed the regression model transforms to a Poisson lognormal (PLN). The PLN model was found to be adequate for the data at hand, since the maximum percentage of zeros was 65% and the skewness for all the datasets was below 3.0 (Vangala, 2015; Vangala et al., 2015).

The main objective of this paper is the examination of the relationship of speed with motorway crashes for two severity levels. Different crash types cannot be considered independent of each other and modelled as such because they are both subsets of the total crashes on a road network (Park and Lord, 2007). For simultaneous modelling two or more crash categories multivariate Poisson lognormal (MVPLN) regression is proposed. MVPLN controls simultaneously for over-dispersion and the correlations between the categories (Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009; Ma et al., 2008; Park and Lord, 2007).

The observations of the link-based dataset cannot be considered as spatially independent. Consequently, in the link-based model the effects of unobserved spatial relationships between adjacent segments should be taken into account by adding a random effect using a multivariate conditional autoregressive priors (CAR) model in a hierarchical Bayesian approach (Aguero-Valverde, 2013; Barua et al., 2014). As mentioned above, the observations of the condition-based dataset are not spatial entities and thus at this case unobserved spatial correlation does not need to be considered. The models below are presented including the random effect for spatial correlation that should be taken as zero for the condition-based dataset.

For a crash count dataset containing \( n \) observations (links or pre-crash scenarios) the number of crashes by severity is Poisson distributed:

\[
y_k \sim \text{Poisson}(\lambda_k), \quad i = 1, 2, \ldots, n \quad k = 1, 2, \ldots, K
\]
where $i$: index of observation, $k$: index of severity type, $y_{ik}$: observed number of crashes of $k$ severity for the $i$th observation and $\mu_{ik}$: the expected mean of crashes of $k$ severity for the $i$th observation. The expected mean $\mu_{ik}$ is a function of the model’s explanatory variables (link function):

$$\ln(\lambda_{ik}) = \beta_0 + \sum_{m=1}^{m} \beta_{km}X_{ikm} + \ln(e_{i}) + \epsilon_{ik} + u_{ik}$$ \hspace{1cm} (7)

where $\beta_{0}$: intercept for severity $k$, $\beta_{km}$: coefficient of the $m$th explanatory variable for severity $k$, $X_{ikm}$: value of the $m$th explanatory variable for the $i$th observation and severity $k$, $e_{i}$: offset/exposure variable, $\epsilon_{ik}$: unobserved heterogeneity for the $i$th observation and severity $k$ and $u_{ik}$: random effect for the spatial correlation between the $i$th observation and its neighbours for severity $k$. In order to take into account the correlations within the unobserved heterogeneity, $\epsilon_{i}$ has a multivariate normal distribution:

$$\epsilon_{i} \sim \text{MVN}(0, \Sigma)$$ \hspace{1cm} (8)

where $\Sigma$ is the variance–covariance matrix of the unobserved heterogeneity.

The $u_{ik}$ term as proposed by Besag (1974) is:

$$u_{ik} | y_{ij} \sim \text{MVN} \left( \frac{\sum_{i \neq j} w_{ij} u_{ij} w_{ij}}{\sum_{i \neq j} w_{ij}}, \frac{\Omega}{\sum_{i \neq j} w_{ij} w_{ij}} \right)$$ \hspace{1cm} (9)

where $w_{ij}$: adjacency weight matrix that denotes $w_{ij} = 1$ if the links $i$ and $j$ are first order neighbours (they share a common boundary), or $w_{ij} = 0$ otherwise, $\Omega$: variance–covariance matrix for the spatial correlation.

$$\Omega = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix}$$ \hspace{1cm} (10)

As the direct computation of the marginal distribution of crash counts is not possible, because it requires the computation of a $K$-variate integral of the Poisson distribution with respect to $\epsilon_{ik}$, the parameter estimation was done via Markov chain Monte Carlo (MCMC) in a Bayesian framework (Barua et al., 2014; Ma et al., 2008; Park and Lord, 2007). The prior distribution for $\beta$ is:

$$\beta \sim \text{MVN}(\beta_0, R_{\beta0})$$ \hspace{1cm} (11)

The conjugate prior distribution of the inverse of the variance–covariance matrix for the heterogeneity an the spatial correlation is usually Wishart (Aguero-Valverde and Jovanis, 2009; Aguero-Valverde, 2013; Barua et al., 2014; Ma et al., 2008; Park and Lord, 2007):

$$\sum_{-1} \sim \text{Wishart}(R, d)$$ \hspace{1cm} (12)

$$\Omega^{-1} \sim \text{Wishart}(S, d)$$ \hspace{1cm} (13)

where $\beta_0, R_{\beta0}, R$ and $S$ are known non-informative hyperparameters and $d$ is equal to the degrees of freedom (number of the examined crash severity types: $d = 2$).

5. Estimation results

The model presented in Eqs. (6)–(13) was fitted to both the link-based and the condition-based datasets using WinBUGS 1.4.3 (Spiegelhalter et al., 2003), an open-source software that is suitable for full Bayes model estimation using the Markov Chains Monte Carlo (MCMC) method. The posterior distributions were obtained from 50,000 iterations of two Markov chains. The first 20,000 iterations were discarded from the final estimations as the
burn-in sample. Convergence was visually detected from Markov chain history graphs of the models’ coefficients. The multivariate models for both the link-based and the condition-based datasets showed improved statistical fit (based on the Deviance Information Criterion) compared to univariate models estimated by severity group.

As there is no clear evidence about the form of the relationship between speed and crash occurrences, three different functional forms in the link function were tested for both datasets: (a) a linear (e.g. $\beta_1 \cdot \text{Speed}$), (b) a logarithmic (e.g. $\beta_2 \cdot \ln(\text{Speed})$) and (c) a quadratic (e.g. $\beta_1 \cdot \text{Speed} + \beta_2 \cdot \text{Speed}^2$). The same strategy was applied to traffic volume. To control for a possible interaction between speed and volume on crash frequency, a multiplicative interaction term (i.e. speed$\times$volume) was also investigated. This results in a total of nine different specifications\(^2\) of the link function. The functional form with the best goodness-of-fit statistic (i.e. the functional form with the lowest Deviance Information Criterion (DIC) score) is considered as the most accurate representation of each dataset thus for brevity only these models are presented here. The best fitting specification for the condition-based model was the quadratic speed and the quadratic volume along with their interaction term and for the link-based linear speed and logarithmic volume without the interaction term. Tables 3 and 4 present the posterior means, standard deviations, Monte Carlo error (MC error) and the 95% credible intervals of the estimated coefficients. The functional forms of the models are described in the next section.

### 6. Discussion

From Tables 3 and 4, it can be seen that the results derived from the two models are significantly different. The estimated coefficients for some of the variables have different signs indicating that the data aggregation concept has a considerable impact on the results of crash prediction models. The relationship between the traffic variables and crashes cannot however be interpreted solely based on the signs of their coefficients due to the variable transformations and the speed-volume interaction term. To facilitate the interpretation of the outcomes, Figs. 6, 7, 10 and 11 provide a graphical representation of the crash rate as a function of speed for three distinct volumes and the reference categories for geometry (i.e., for a straight (Curve = 0) and level segment (Downhill = Uphill = 0) with 2 or less lanes (lanes above 2 = 0). Figs. 8, 9, 12 and 13 illustrate the variations of crash rate as a function of the entire range of speed and volume. The corresponding KS and SL crash rates for the link-based approach can be shown as follows:

#### KS crashes

$$\text{Link Length}_{\text{KS}} = \exp(-0.0231 \cdot \text{Speed} + 0.1310 \cdot \ln(\text{AADT}) - 0.0551)$$  \hspace{1cm} (14)

#### SL crashes

$$\text{Link Length}_{\text{SL}} = \exp(-0.0290 \cdot \text{Speed} + 0.6848 \cdot \ln(\text{AADT}) + 0.6075)$$  \hspace{1cm} (15)

---

\(^1\) Statistically significant coefficients at the 95% credible interval.

Table 4
Multivariate coefficient estimates for crashes with killed and serious injured (KS) and crashes with slightly injured (SL) crashes for the condition-based dataset.

<table>
<thead>
<tr>
<th></th>
<th>KS crashes</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>MC Error</td>
<td>2.50%</td>
<td>5.00%</td>
<td>Median</td>
<td>95%</td>
<td>97.5%</td>
<td></td>
</tr>
<tr>
<td>Speed</td>
<td>0.02414</td>
<td>0.00897</td>
<td>0.00047</td>
<td>0.00720</td>
<td>0.01043</td>
<td>0.02356</td>
<td>0.04045</td>
<td>0.04281</td>
<td></td>
</tr>
<tr>
<td>Speed squared</td>
<td>-0.00014</td>
<td>0.00005</td>
<td>0.00000</td>
<td>-0.00025</td>
<td>-0.00024</td>
<td>-0.00014</td>
<td>-0.00006</td>
<td>-0.00004</td>
<td></td>
</tr>
<tr>
<td>Volume</td>
<td>-0.02377</td>
<td>0.02201</td>
<td>0.00009</td>
<td>-0.02417</td>
<td>-0.02363</td>
<td>-0.02036</td>
<td>-0.01695</td>
<td>-0.01621</td>
<td></td>
</tr>
<tr>
<td>Volume squared</td>
<td>0.00004</td>
<td>0.00000</td>
<td>0.000002</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.00004</td>
<td>0.00005</td>
<td></td>
</tr>
<tr>
<td>Speed · Volume</td>
<td>0.00002</td>
<td>0.00002</td>
<td>0.000001</td>
<td>-0.00001</td>
<td>0.000001</td>
<td>0.00002</td>
<td>0.00005</td>
<td>0.00006</td>
<td></td>
</tr>
<tr>
<td>Curve</td>
<td>0.08056</td>
<td>0.06534</td>
<td>0.00009</td>
<td>-0.04698</td>
<td>-0.02665</td>
<td>0.08047</td>
<td>0.18860</td>
<td>0.20870</td>
<td></td>
</tr>
<tr>
<td>Straight (reference)</td>
<td>0.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Uphill</td>
<td>2.12500</td>
<td>0.15890</td>
<td>0.00441</td>
<td>1.82500</td>
<td>1.87000</td>
<td>2.12200</td>
<td>2.39400</td>
<td>2.45100</td>
<td></td>
</tr>
<tr>
<td>downhill</td>
<td>2.95200</td>
<td>0.15460</td>
<td>0.00451</td>
<td>2.65900</td>
<td>2.70400</td>
<td>2.94900</td>
<td>3.21400</td>
<td>3.27000</td>
<td></td>
</tr>
<tr>
<td>Even (reference)</td>
<td>0.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Lanes above 2</td>
<td>-0.64770</td>
<td>0.06954</td>
<td>0.00099</td>
<td>-0.78550</td>
<td>-0.76190</td>
<td>-0.64710</td>
<td>-0.53300</td>
<td>-0.51090</td>
<td></td>
</tr>
<tr>
<td>Lanes up to 2 (reference)</td>
<td>0.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.23900</td>
<td>0.43100</td>
<td>0.02135</td>
<td>-4.15100</td>
<td>-4.01600</td>
<td>-3.20200</td>
<td>-2.58600</td>
<td>-2.47900</td>
<td></td>
</tr>
<tr>
<td>Ln (Exposure)</td>
<td>1.00</td>
<td>6.534</td>
<td>0.00099</td>
<td>-0.04698</td>
<td>-0.02665</td>
<td>0.08047</td>
<td>0.18860</td>
<td>0.20870</td>
<td></td>
</tr>
</tbody>
</table>

* Statistically significant coefficients at the 95% credible interval.

Fig. 7. Predicted SL crashes per kilometre as a function of speed for links with average annual daily traffic: (a) 20,000, (b) 40,000 and (c) 60,000.

Fig. 8. 3D contour plot of the predicted KS crashes per kilometre as a function of speed and average annual daily traffic.
Overall, the results of the link-based model were hard to interpret and to a certain extent counterintuitive (see Figs. 6–9). Speed was found to be inversely proportional with all crashes. Although some other studies have presented similar findings (Baruya, 1998; Lave, 1985), none of the researchers has given a very good explanation of why higher average speeds are overall safer. Some of the main arguments to support this idea are the increased design standards of high speed motorways and the longer available distances between vehicles at high speed conditions. However, the vast majority of studies that examined the number of crashes before and after speed limit changes (consequently changes in average speed) suggest that higher speeds are related to more crashes (e.g., Elvik et al., 2004).

Higher AADT is related with more crashes, however, considering the estimated coefficients AADT has stronger impact on SL crashes that on KS, a result that is in-line with most of the existing studies. As for the geometrical features of the links, they mostly seem to be statistically insignificant apart from links with more than two lanes for all crashes and downhill links for SL crashes only. The use of dummy variables for geometry could possibly affect the estimated coefficients. However, the signs of the coefficients of the most important variables (i.e., speed) did not change even when the geometrical characteristics were represented by continuous variables, that is not presented due to brevity. These results possibly indicate the inability of average measures of time-varying variables that are frequently used in the link-based approaches to accurately explain the variation in crashes and that this inefficiency might have a direct impact on the modelling results.

On the other hand, the outcomes of the condition-based models are quite different (see Figs. 10–13). Speed was found to be proportional with both crash frequencies (i.e., KS and SL crashes). The shape of the curves shows that the number of crashes increases proportionally with speed until a point (e.g. 85 km/h at a volume of 100 vehicles/lane) and then either it stabilises or decreases. This can be potentially explained by the decrease of crash prone reactions that increase while speed reaches very high values (Navon, 2003). Comparing the maxima of the curves between Figs. 10 and 11 it can be seen that, not surprisingly, crashes which occur under higher speed conditions tend to have more serious outcomes; a finding that is consistent with the literature (e.g., Kloeden et al., 1997; Pei et al., 2012). The KS and SL crash rates for the reference cases of categorical independent variables, i.e., Curve = 0, Uphill = Downhill = 0, Lanes above 2 = 0; see Table 4):

\[
\frac{\text{KS crashes}}{\text{VehHours per km}} = \exp(0.0241 \cdot \text{Speed} - 0.00014 \cdot \text{Speed}^2 - 0.0204 \cdot \text{Volume} + 0.00004 \cdot \text{Volume}^2 + 0.00002 \cdot \text{Speed} \cdot \text{Volume} - 3.24) \tag{16}
\]

\[
\frac{\text{SL crashes}}{\text{VehHours per km}} = \exp(0.036 \cdot \text{Speed} - 0.0002 \cdot \text{Speed}^2 - 0.0076 \cdot \text{Volume} + 0.000025 \cdot \text{Volume}^2 - 0.00003 \cdot \text{Speed} \cdot \text{Volume} - 3.01) \tag{17}
\]

An interesting finding of the condition-based model was that the frequency of crashes is higher at low volume conditions than that of at high volume conditions, ceteris paribus. More specifically, the relationship between crash rate and volume is described as an
approximate U-shaped curve with the minimum crash rates were found to be at 241 and 211 vehicles per lane for KS and SL crashes respectively at average speed conditions (see Figs. 12 and 13). This outcome is consistent with the results for speed, because high volume is usually associated with congested, low speed conditions when crashes are less likely to be severe and reported (Lord, 2002). Another explanation for this finding could be that low volumes can be related with higher speed variations (when traffic is building up) that may increase the probability of crashes (Garber and Ehrhart, 2000). This is because when the volume decreases drivers have more freedom to choose their own speed and so speed patterns on the roadway tend to be less uniform leading to more encounters between vehicles (Elvik et al., 2004). Additionally, low volumes occur more often during off-peak periods, such as night time, that is related to insufficient light conditions and extreme driving behaviours (e.g. drinking and driving) that are also factors proved to trigger crash occurrence (Chang and Wang, 2006; Clarke et al., 2010; Jonah, 1986).

Curvature is not shown to have a statistically significant relationship with KS crashes but it increases the likelihood of SL crashes. The finding for SL crashes is consistent with other studies on the relationship of horizontal alignment with crashes (Ma et al., 2008; Milton and Mannering, 1998; Park et al., 2010a). However, the outcome for KS crashes is not expected as literature suggests that curvature is associated with higher crash severity (Geedipally et al., 2013; Ma and Kockelman, 2006). The high design standards of the study area could be a possible explanation why curvature is not statistically significant for KS crashes (i.e. small radius curves are relatively rare for motorways and major A-roads) although it has been suggested that curvature is linked with more crashes even on freeway segments (Park et al., 2010a). Another explanation could be that speeding and other risk-taking actions might be more unlikely on curved sections. Vertical alignment of the road section just before a crash is found to be associated with more crashes. The existence of both positive and negative slope seems to triggers crash occurrence although, based on the coefficient values, the latter has higher impact. This outcome is in line with findings of existing literature (e.g., Milton and Mannering, 1998). Finally, roads with more than two lanes are related to lower crash counts for all crash severities. This is similar to the findings of Ma and Kockelman (2006) who reported that the number of lanes decreases crash counts for non-fatal crashes and the results by Bonneson and Pratt (2008) and Park et al. (2010b) who found that 6-lane freeways are less crash prone than 4 or 8-lane but opposite to the majority of current literature (Chang, 2005; Milton and Mannering, 1998). A possible explanation for that could be that wider roads allow more manoeuvres for crash avoidance during a crash-prone encounter. Moreover, this result can also be explained by the inclusion of crashes that occurred on undivided (single) carriageways. Over half of the examined crashes occurred on A-roads that include some single carriageways which are related with hazardous vehicle interactions that may lead to crashes with severe consequences (e.g. head-on collisions).

Considering the variations between the results of the two models, it is clear that aggregation bias that occurs at link-based approaches might lead to significant errors, meaning that the data aggregation concept plays a major role to the outcomes of safety analyses. This subject has been disregarded by most researchers, who mainly focused their research on developing more advanced statistical models; however it seems that the way crash data are
prepared for the statistical analysis is important too. The link-based and the condition-based models cannot be directly compared to each other neither using goodness-of-fit statistics nor based on the interpretability of their outcomes. However, it can be argued that the condition-based model gives a significantly more accurate representation of the crash-related conditions and so its results apart from being more reasonable might be also more reliable.

7. Conclusions

This paper presented a novel crash modelling approach in re-examining crash–speed relationships based on a new concept that overcomes some existing limitations of the conventional approach. The originality of the work lies in the development of an alternative data aggregation concept that defines the pre-crash traffic and geometric conditions as the crash aggregating factors. Compared to the approaches that assign crashes into groups based on their spatial relationship with road entities, the new method addresses the inherent problem of over aggregation of time-varying traffic variables and relevant information losses that may affect the modelling outcomes.

The new modelling approach was employed to all the crashes that occurred on the Strategic Road Network (SRN) of England during 2012. Pre-crash condition identification was based on geocoded crash locations obtained by a crash mapping algorithm that was previously developed for the study area. In order to compare the traditional modelling approach with the proposed approach, link-based and condition-based datasets were developed based on identical crash, traffic and geometry data. Bayesian multivariate Poisson lognormal regression was employed for modelling both datasets by injury severity taking into account first order spatial correlation for the link-based model. The models explored the optimal variable specifications as well as for potential interactions between speed and volume.

Speed has been found to be a significant contributory factor for the number and the consequences of crashes when the data are modelled with the condition-based approach. In contrast to that, according to the results of the link-based model speed has a negative relationship with crash occurrences for all severity types. From a methodological point of view, the difference in the results of these approaches reveals that the data aggregation method is an important decision before conducting a crash data statistical analysis. Thinking that the link-based approaches include observations that often lack details and tend to mask the crash contributory factors, link-based models are very likely to have limited explanatory potential. Condition-based approaches, on the other hand, focus on the crash time and location and can be considered as more representative of the actual circumstances. That is why they provide more explainable, logical and possibly more credible outcomes.

Condition-based crash modelling, according to the results presented on this paper, is a new and promising approach that can increase the insight about various crash triggering factors and by indicating hazard–prone traffic conditions contribute to the assessment and the development of road safety measures. The method is flexible and transferable to other study areas and can be implemented using different combinations of variables and, preferably, higher resolution traffic data. Future work should include assessment of the two methods through comparison of the predicted values between link-based and condition-based models. Instead of employing condition-based as a substitute of link-based methods, it would be also useful to research whether and how these approaches could work complementary of each other towards the quantification of crash risk from different perspectives.

Acknowledgements

The authors would like to gratefully thank Prof. Benjamin Heydecker and Prof. Mike Maher of University College London for their thoughts and comments during the development of this work. This research was partially funded by a grant from the UK Engineering and Physical Sciences Research Council (EPSRC) (Grant Reference: EP/R018894/1). The authors take full responsibility for the content of the paper and any errors or omissions. Research data for this paper are available upon request from Maria-Ioanna Imprialou.

References


