# A state-of-the-art review on Crash Prediction Modelling

**Anastasios Dragomanovits, Ourania Basta, Aikaterini Deliali[1], George Yannis**

*Department of Transportation Planning and Engineering, National Technical University of Athens, 5, Heroon Polytechniou st., 15773, Athens, Greece*

## Abstract

Despite the efforts to improve road safety, road crashes constitute a major global societal problem with more than 1.25 million fatalities per year (first mortality cause for the ages 15-29). Crash Prediction Models (CPMs), including Safety Performance Functions (SPFs) and Crash Modification Factors (CMFs) and other advanced statistical models are essential tools for transport authorities and highway agencies, mostly in developed countries, to predict crashes, analyze injury severity, identify hotspots, and assess safety countermeasures. The objective of this present study is to provide a state-of-the-art review of international literature on microscopic road safety modelling methods and statistical approaches, and to identify trends and gaps of knowledge in pertinent research. A structured keyword search in online scientific databases was performed to identify the most relevant publications on crash prediction modeling. Following a set of rigorous criteria, more than 100 research publications on microscopic crash modeling were identified as appropriate for this review. A second level of categorization was applied, and included: (a) AASHTO's Highway Safety Manual and related publications, also including recent publications on its expected update, (b) development of Safety Performance Functions (SPFs), i.e. basic crash prediction models developed for "base conditions", (c) estimation and use of Crash Modification Factors (CMFs), to account for differences in geometric design/traffic control features between the base conditions of the model and local conditions of the site under consideration, and (d) stand-alone multivariate Crash Prediction Models that usually include many explanatory variables compared to SPFs, in order to consider site characteristics on their own, without the use of CMFs, and (e) multivariate models using machine learning approaches. The paper discusses the current state-of-art in the field of crash prediction modeling, emphasizing on future directions as well as strengths and limitations of the existing approaches.

**Keywords:** crash prediction; safety performance function; crash modification factors; motorways; highways.

[1] *Corresponding author.*
*Tel (+30) 6971551576*
*kdeliali@mail.ntua.gr*

## 1. Introduction

Road authorities, traffic engineers, and road safety analysts employ crash occurrence analysis with the objective to identify crash contributing factors and in turn select appropriate countermeasures. The analysis of crash records can reveal crash contributing factors related to road user behavior (e.g., seat belt or helmet use), vehicle type and vehicle conditions (e.g., motorcyclists or old vehicles are more prone to injury-related crashes), external conditions (e.g., weather and lighting conditions, time of the day, etc.), and the road infrastructure. Crash prediction models (CPMs) including Safety Performance Functions (SPFs) and Crash Modification Factors (CMFs) are essential tools to predict crashes and use this information to identify hazardous locations and/or evaluate the effectiveness of countermeasures. CPMs exist for the microscopic, mesoscopic, and macroscopic levels. Essentially, this differentiation is based on the unit of analysis. Microscopic modelling refers to models that use as basis for the analysis short homogenous road segments, or specific sites (intersections, interchange-influence areas, etc.). Mesoscopic and macroscopic models have larger units of analysis that can be at the zone-level up to the region-level.

The use of CPMs in the safety management process is of high importance as these models can be used to estimate the expected number of crashes at a site and in turn, increase the chances of correctly classifying a site as hazardous [1]. Therefore, the use of CPMs has the potential to improve safety but at the same time, ensure a better allocation of road safety funding. The objective of this study is to provide a state-of-the-art review of international literature on microscopic crash modelling methods and statistical approaches, and to identify trends and gaps of knowledge in pertinent research.

The rest of this paper is organized as follows. Section 2 presents the methodology that was followed to identify resources related to microscopic crash prediction modeling. The third section presents the findings of the review, focusing on the main approaches to develop CPMs and CMFs as well as discussing the limitations of the various predictive methods. The Discussion section centers on issues usability of CPMs and CMFs as well as practical implications. Lastly, the Conclusions present a summary of this work and discuss paths for future research.

## 2. Methodology

The literature review was conducted on documents, scientific papers, reports, project deliverables etc. discussing microscopic road safety analyses and models. As this research field is particularly extensive - a preliminary non-exhaustive keyword search on ScienceDirect, using as keywords the terms: "road" AND ("accident prediction" OR "crash prediction" OR "safety performance function" OR "crash modification") returned 1.097 results as of May 2020 - it was decided to set specific criteria for the examination of relevant references: horizontal criteria, regarding the language and the geographic origin, and vertical criteria, according to the different types of references.

Namely, the horizontal selection criteria excluded the following categories of studies:

- Publications in languages other than the English language.
- Publications from geographic origin of reports and studies other than the following areas: Europe, USA, Canada, Australia, New Zealand, and China.
- Studies on macro/planning-level applications (analysis based on jurisdiction, GDP, etc.) and on mesoscopic models.
- Publications on models focusing on specific user groups (e.g., pedestrians, bicyclists, powered two-wheelers, heavy vehicles) or on specific road elements (tunnels, bridges, railway level crossings, etc.) were also excluded.

Vertical selection criteria differ according to each category of references. Specifically:

- EU Research Project Deliverables (also including project internal reports, if available): All available references originating from EU research projects were included in the review.
- Reports, guidelines, manuals from public authorities and governmental organizations: Reports, guidelines and manuals from authorities and organizations at a national level were included in the review (e.g., American Association of Highway Transportation Officials (ASHTO), Federal Highway Administration (FHWA), National Cooperative Highway Research Program (NCHRP), AUSTROADS, New Zealand Transport Agency, etc.). Reports from regional authorities (e.g., State Departments of Transport in the US) were generally not included in the review, except for one indicative example.

- Reports from other non-governmental organizations: Reports from non-governmental organizations (academia, research institutes) were generally included in the review, if they were of national or international interest, i.e., based on data covering large geographical areas, or utilizing an interesting methodology.
- Journal Papers: The selection of journal papers to be included in the review was based on a step-by-step search process in the Scopus bibliometric database (date of query: 03/06/2020) combining keyword search, publication date and number of citations. The process is described in detail in Annex E of this deliverable.
- Conference Papers: Publications in scientific conferences were not included in the review, considering that researchers generally prefer to publish high quality work in journals with a peer review process.

It is noted that some additional literature references were included in the review (e.g., references retrieved from other reviewed papers), even if not derived from the explicit implementation of the aforementioned methodology. Such references were considered to provide important insights and significantly advance knowledge in the field of microscopic/ mesoscopic safety modelling, or they refer to geographic areas for which no other publications exist. More than 100 publications were identified as most relevant and of appropriate quality and were considered for this review. These publications refer to models on segments and/or junctions and due to space limitations, it was decided to exclude the publications that focused on junction safety (i.e., intersections, roundabouts, and interchanges).

Microscopic modelling refers to models that use as basis for the analysis short homogenous road segments, or specific sites (intersections, interchange-influence areas, etc.), depending on the model. Such models tend to use detailed explanatory variables, which, besides exposure (probably the single most important variable in all crash prediction models), are mostly related to road geometric characteristics (curvature, lane width, number of lanes, type of traffic control, etc.) but also operational characteristics (speed limits, signage, etc.). The following four groups were formed for the identified studies on microscopic models:

- Publications related to AASHTO's Highway Safety Manual Predictive Method (AASHTO, 2010, 2014) in addition to the manual itself.
- Publications on the development of Safety Performance Functions (SPFs). SPFs are basic crash prediction models developed for "base conditions", and are typically a function of only a few variables, primarily average annual daily traffic (AADT) volumes and segment length.
- SPFs cannot be used on their own for crash prediction as they require the use of Crash Modification Factors (CMFs), to account for differences in geometric design or in traffic control features between the base conditions of the model and local conditions of the site under consideration.
- Stand-alone multivariate CPMs are models that include a large number of explanatory variables compared to SPFs, in order to consider site characteristics on their own, without the use of CMFs.

## 3. Analysis and Results

The following subsections discuss existing research on microscopic CPMs and CMFs. For microscopic modeling, the most well-known approach is the AASHTO's Highway Safety Manual (HSM) which is used at the international level, or more likely has been the base of developing a handful of prediction models internationally. The HSM Predictive Method (see section 3.1) suggests that a base SPF is developed for the base conditions of a road facility (e.g., two-lane rural roads, urban arterials, rural four-leg intersections, etc.) and then, the impact of all additional road design and operational elements on crash frequency is captured by the respective CMFs. Section 3.2 present existing approaches for the development of CMFs. Besides HSM Predictive Method that proposes the use of SPF-CMF, there are numerous efforts on developing other CPMs, referred herein as stand-alone multivariate CPMs; these models are presented in section 3.3. These models do not consider base conditions separately and essentially, include multiple explanatory variables in one CMP. There is a recent trend for developing multivariate CPMs using machine learning approaches (see section 3.3.3), mainly to achieve a higher predictive performance and/or incorporate more real-time explanatory variables (e.g., traffic conflicts).

### 3.1 AASHTO Highway Safety Manual

The HSM Predictive Method [2-3] was developed in the US and has been adopted by international practitioners and road agencies. The HSM Predictive Method has stimulated pertinent research in the development of SPFs and CMFs worldwide, with several hundreds of CMFs being currently available by independent research outside the US. The next subsections present the main concept (section 3.1.1) and limitations of the method (3.1.2).

#### *3.1.1 Concept*

The basic concept of the predictive method is that of a simple Safety Performance Function for base conditions ($SPF_{base}$), later adjusted to local conditions using Crash Modification Factors (CMFs). SPFs and CMFs are

developed for certain road facility types, e.g., two-lane rural roads, urban arterials, rural four-leg intersections, etc. For each of these road facilities, the "base" conditions consider traffic volume in the form of Annual Average Daily Traffic (AADT) and segment length for the case of segments. The general form of the predictive models in HSM (AASHTO, 2010), for a given type of road facility, is shown in Equation (1) [2]:

$$N_{predicted} = N_{SPF} \text{ x } (CMF_1 \text{ x } CMF_2 \text{ x } ... \text{ x } CMF_y) \text{ x } C \quad (1)$$

Where $N_{SPF}$ is the predicted average crash frequency determined for the base conditions according to the $SPF_{base}$, $CMF_1$ .... $CMF_y$ are the crash modification factors that account for specific road design and operational characteristics, $C$ is the calibration factor to adjust the SPF for local conditions related to the network where the model is to be applied, and lastly, $N_{predicted}$ is the predicted average crash frequency.

Using Equation (1), the predicted average crash frequency of an individual site, $N_{predicted}$, can be estimated, based on geometric design, traffic control features and traffic volumes of that site. To improve the statistical reliability of the estimate for an existing site or facility, the observed crash frequency $N_{observed}$, can be combined with $N_{predicted}$, to obtain the expected average crash frequency, $N_{expected}$. This is an estimate of the long-term average crash frequency that would be expected, given sufficient time to make a controlled observation. Since the observed crash frequency in a site, roadway or network varies randomly over any period, using averages based on short-term periods (e.g., 1 to 3 years) may give misleading estimates due to regression-to-the-mean bias. The long-term average crash frequency is estimated using the Empirical Bayes.

#### *3.1.1 Limitations*

Researchers have identified several shortcomings in the HSM approaches. Regarding the Safety Performance Functions, since the HSM SPFs have been developed using data only from the states of California, Michigan, Minnesota, Texas, and Washington [4], they may not adequately reflect the geographic and traffic safety diversity across the U.S. In fact, several US state transportation agencies have used the calibration procedure and have noted issues related to poor accuracy when comparing the predictive results of the calibrated SPFs to reported crash frequencies or locally developed SPFs. For these same reasons, individual CMFs applied to an SPF might also be biased when applied to different states.

Gooch et al. have also questioned the HSM approach for horizontal curves on two-way, two-lane rural roads of applying an SPF developed using only tangent sections and adjusting this value using a CMF for curvature, as they found that the shape of the relationship between safety performance and traffic volume actually differs for horizontal curve and tangent segments [5]. Thus, they have effectively argued in favor of the development of separate SPFs for tangent and curves.

Research has also identified weaknesses in the consideration of Crash Modification Factors in the HSM predictive method, with most prominent the handling of multiple treatments. The preferred approach proposed in the HSM is to apply a single CMF that represents the combined treatments. However, such complex CMFs rarely exist in international literature, and the second and most applied procedure is to multiply the CMF values for each treatment alone to arrive at the single combined effect. Research has shown that this is an oversimplified approach and several other more appropriate, alas more complex, methods should be used for this purpose [6-8].

Finally, the calibration methodology of the HSM has also received criticism. Researchers have identified poor model fit in calibrated as per the HSM models, and they have suggested improved, yet also more complex and data intensive, methods, such as calibration per crash type, or multiple calibration factors for different components of the predictive method, SPF parameters and CMFs rather than a single calibration factor [9].

### 3.2 Development of CMFs

The development of crash modification factors (CMFs) has attracted lots of research at an international level in an effort to quantify the safety impact of various road designs, road and operational elements. Various methods have been found appropriate for developing CMFs (see section 3.2.1) while there are several considerations on how researchers and practitioners can develop robust and reliable CMFs.

#### *3.2.1 Methods for CMF development*

Several different approaches and methods are used internationally for the development of Crash Modification Factors. FHWA (2010) summarized these methods listing their strengths and weaknesses in addition to describing them (see Table 1). Before-after studies, especially when combined with the Empirical Bayes (EB) [10], are commonly considered as the best approach for developing CMFs. Compared with other methods, it is a statistically robust method that can effectively account for regression-to-the-mean bias, for traffic volume changes over time,

and for trends in the safety performance not related to the examined treatment or feature. Several researchers attempted to compare EB methods to traditional approaches for CMF development and found the former to be more appropriate [11, 12] however, EB was found to be comparable to Full Bayes (FB) [13]. This would indicate that it may not be worth the additional complexity and data needs of FB method, especially considering that for road safety practitioners EB is already too complicated. It was suggested that the FB is worth considering for situations where it is difficult to acquire a large enough reference group to calibrate SPFs required for the EB approach. Lastly, when it is not feasible implementing a before-after approach (and EB) with careful selection of sites, adequately large sample, inclusion of appropriate explanatory variables in the model, and proper assumption of the relationship between variables and their safety effects, reliable CMFs can also be developed from cross-sectional regression studies.

**Table 1: Summary of methods for CMF development *(Adopted from [14])***

| Method | Applications | Strengths | Weaknesses |
|---|---|---|---|
| Before–after with comparison group | -Treatment is similar amongst treatment sites.<br>-Before and after data are available for both treated and untreated sites.<br>-Untreated sites are used to account for non-treatment related crash trends. | Simple<br>Accounts for non–treatment-related time trends and changes in traffic volume. | Difficult to account for regression-to-the-mean. |
| Before–after with empirical Bayes | -Treatment is similar amongst treatment sites.<br>-Before and after data are available for both treated sites and an untreated reference group. | Accounts well for regression-to-the-mean, traffic volume changes over time, non–treatment-related time trends. | -Relatively complex.<br>-Cannot include prior knowledge of treatment.<br>-Cannot consider spatial correlation. |
| Full Bayes | Useful for before–after or cross-section studies when:<br>-There is a need to consider spatial correlation among sites.<br>-Complex model forms are required.<br>-Previous model estimates or CMF estimates are to be introduced in the modeling. | -Reliable results with small sample sizes.<br>-Can include prior knowledge, spatial correlation, and complex model forms in the evaluation process. | Implementation requires a high degree of training. |
| Cross-sectional | -Useful when limited before–after data are available.<br>-Requires sufficient sites that are similar except for the treatment of interest. | -Possible to develop CMFunctions.<br>-Allows estimation of CMFs when conversions are rare.<br>-Useful for crash prediction. | -CMFs may be inaccurate for a number of reasons: inappropriate functional form, omitted variable bias, correlation among variables. |
| Case-control | -Assess whether exposure to a potential treatment is disproportionately distributed between sites with and without the target crash.<br>-Indicates the likelihood of an actual treatment through the odds ratio. | -Useful for studying rare events because the number of cases and controls is predetermined.<br>-Can investigate multiple treatments per sample. | -Can only investigate one outcome per sample.<br>-Does not differentiate between locations with one crash or multiple crashes.<br>-Cannot demonstrate causality. |
| Cohort | Used to estimate relative risk, which indicates the expected percent change in the probability of an outcome given a unit change in the treatment. | -Useful for studying rare treatments because the sample is selected based on treatment status.<br>-Can demonstrate causality. | -Only analyzes the time to the first crash.<br>-Large samples are often required. |
| Meta-analysis | Combines knowledge on CMFs from multiple previous studies while considering the study quality in a systematic and quantitative way. | -Can be used to develop CMFs when data are not available for recent installations, and it is not feasible to install the strategy and collect data.<br>-Can combine knowledge from several jurisdictions and studies. | -Requires the identification of previous studies<br>-Requires a formal statistical process.<br>-All studies included should be similar in terms of data used, outcome measure, and study methodology. |

| | | | |
|---|---|---|---|
| Expert panel | -Expert panels are assembled to critically evaluate the findings of published and unpublished research.<br>-A CMF recommendation is made based on agreement amongst panel members. | -Can be used to develop CMFs when data are not available for recent installations, and it is not feasible to install the strategy and collect data.<br>-Can combine knowledge from several jurisdictions and studies.<br>-Does not require a formal statistical process. | -Traditional expert panels do not systematically derive precision estimates of a CMF.<br>-Possible complications may arise from interactions and group dynamics.<br>-Possible forecasting bias. |

After reviewing 36 studies on the development of CMFs for rural and/or urban roads it was found that 13 of them used the cross-sectional approach while 16 of them used the EB approach. The rest of the studies chose one of the above approaches.

### *3.2.2 Considerations regarding the CMF development*

In addition to the study design, several factors may result in erroneous estimations of the safety effect of the studied treatments, and in turn result in erroneous CMFs. Table 2 summarizes these factors and suggests ways for their mitigation.

**Table 2: Factors affecting the treatment evaluation**

| Factor | Mitigation |
|---|---|
| Changes in traffic flow | Traffic may experience changes over time, unrelated to the treatment implementation (e.g., general population growth). These changes should be considered when evaluating the treatment otherwise its effect will be over- or underestimated. |
| General crash trends | Crash trend may change regardless of the studied treatment (e.g., due to safer vehicle design, presence of enforcement). To simply capture the effect of the treatment it is important to include control sites in the analysis. |
| Regression to the Mean | Crashes are rare and random effects and vary over time. Capturing a small time period for the analysis may result in ignoring the long-term crash trend and working with a local minimum or maximum. Capturing larger periods of time is recommended. |
| Crash migration | Crash migration is often the result of changes in traffic flow or driver behavior associated with the implementation of the examined treatments. Essentially crashes may change type (e.g., side-swipe to rear end) or "migrate" to another sites nearby the treatment site. For changes in the crash type, the evaluation of the treatment should focus both on the total number of crashes before-after and on the number of crashes per type. |
| Adjustment period/ halo Effect | An adjustment period is in some cases encountered after the introduction of a treatment and to avoid capturing data from the adjustment period the "after" period should not overlap with the former. |
| Statistical validity | According to [15] six criteria are identified for assessing statistical conclusion validity: use of an appropriate sampling technique, having an adequate sample size, specifying whether the evaluation is to be in terms of crash numbers or injury severity outcomes, reporting the uncertainty associated with estimates, and using appropriate statistical testing techniques. Two problems that commonly complicate crash studies are zero inflation and over dispersion. |

A main issue regarding CMFs, even if they have been developed in a reliable manner following all the above recommendations, is that they may not be transferable. Previous research has shown that a CMF is not a universal constant (i.e., having the same value in all conditions), but instead it should be viewed as a random variable, the value of which depends on a host of factors (circumstances of implementation) [16]. Therefore, a CMF has a probability distribution with a mean and a variance. Thinking of CMFs as random variables allows the question of transferability to be correctly framed. To increase the effectiveness of CMF research, the variance must be reduced. Two approaches were discussed in [16]: conducting more studies and making the CMF a function of the circumstances of implementation.

### *3.2.3 Existing CMFs and their usability*

The FHWA lists existing CMFs (developed internationally) at the CMF Clearinghouse (http://www.cmfclearinghouse.org). The database is frequently updated to include the recently developed CMFs and almost 7,000 CMFs from independent studies can be found there. It is therefore obvious that an exhaustive review of pertinent literature is both unfeasible, especially for this paper. However, it is important to note that

while there is a wide range of CMFs for all road types as well as junction types (e.g., four-leg intersection and interchanges) and they can be potentially used by practitioners, there are several limitations. Some CMFs have been developed for particular conditions and cannot be transferred to different ones without proper calibration. For example, they may refer to certain crash types (e.g., single vehicle crashes or night-time crashes). Some CMFs may be based on data from rural roads and so, are not appropriate for other roads (e.g., urban roads or motorways). Therefore, it is critical to be aware of the conditions under which a CMF was developed. A second issue that becomes apparent when one reviews available CMFs on the same treatment/design element (e.g., horizontal curve radius) is that existing studies produce different CMFs even when the same type of crashes and roads are considered. These differences are attributed to local human factors (e.g., vehicle fleet characteristics, driver behavior, etc.) and critical judgment and expertise is required to synthesize the available information.

### 3.3 Multivariate CMPs

In parallel to the HSM concept, researchers have developed multivariate CPMs. These models are conceptually more complex than the SPF-CMF approach, as it is required to incorporate all significant explanatory variables in model, and this is probably the most challenging aspect when developing these models. Data availability determines a set of potential explanatory variables for the models however, the set of final model variables is usually smaller. The main reasons for excluding explanatory variables are: (i) correlation between variables, as a model that incorporates correlated variables will have a reduced performance and (ii) some variables are not found statistically significant. The determination of the final set of explanatory variables is a complex process and requires multiple iterations, to check which explanatory variable combinations maximize the model's predictive power. The following subsections presents studies that have developed models for rural road or motorway segments using conventional statistical approaches. The last subsections present crash prediction models based on machine learning algorithms.

#### *3.3.1 Models for rural road segments*

Several studies at the international level have attempted to develop multivariate crash predictions models for rural segments (see table 3). The great majority of these studies use Negative Binomial models (NB) while for rural segments it is evident that the most important explanatory variables are AADT and variables related to the presence and type of horizontal curves, while variables related to crossing flows (e.g., driveway density or presence of intersections) have also been found significant in some cases. Regarding the significant variables, it is evident from Table 3 that for the same dataset, the significant variables change from one model type to another, meaning that specifying the appropriate model type is critical for the quantification of safety.

**Table 3: Multivariate CPMs for rural road and motorway segments**

| Study | Model | Crash type | Statistically significant explanatory variables |
|---|---|---|---|
| Rural road segments | | | |
| [17] | P | single-vehicle crashes | daytime, volume/capacity ratio, shoulder width, presence of intersections and driveways, presence of passing lanes |
| | | multi-vehicle crashes | daylight conditions, number of intersections and driveways. |
| [18] | NB | POD crashes | Horizontal alignment, speed limit, visibility, road surface condition, AADT |
| | RENB | POD crashes | |
| | NB | Injury crashes | Horizontal alignment, speed limit, visibility, AADT |
| | RENB | Injury crashes | Speed limit, AADT |
| [19] | NB | all crashes | AADT, lane width, horizontal curvature, vertical curvature, density of pedestrian crossings, density of access points |
| [20] | NB | all crashes | AADT, curve ratio, speed differentials density |
| [21] | NB | all crashes | AADT, average curvature change rate, shoulder width, forest environment |
| Motorway segments | | | |
| [22] | P | all crashes | Season (snow, dry), median width, three traffic lanes, grade |
| | REP | | Season (dry, snow), grade, three traffic lanes |
| | Spatial | | Season (snow, dry), degree of curvature, median width, three traffic lanes |
| [23] | Spatial | all crashes | AADT, segment length, delay, speed limit |
| [24] | P and NB | all crashes (curve) | AADT, segment length, radius of horizontal curves |
| | | all crashes | AADT, segment length, presence of junctions |

| | | | |
|---|---|---|---|
| | | (tangent) | |
| [25] | NB | all crashes | AADT, differential speed, difference in friction, grade, tangent length |
| [26] | NB | all crashes | AADT, segment length, horizontal curvature |

*P: Poisson, REP: Random Effects Poisson, NB: Negative Binomial, RENB: Random Effects NB*

***3.3.3 Models for motorway segments***

Based on Table 3, motorway segment safety is mostly impacted by AADT, vertical and horizontal curves. The impact of AADT in all multivariate CPMs (both for segments and intersections) aligns with the HSM logic, where the base SPF has this variable in addition to segment length.

***3.3.3 Approaches using machine learnings models***

More recently, there is a growing interest on machine-learning (ML) multivariate CPMs. ML models are implemented to address some limitations of the traditional statistical models. The later may be limited by the fact that they require some assumptions for the distribution of the data and also, they assume a linear form between the response variable and the explanatory variables. Additionally, developing parametric models requires multiple trial and error tests before determining the final model structure or in other words, a set of statistically significant parameters. ML models are not affected by the data distribution, while when it comes to model specification it can be seen from the reviewed literature that trial and error processes are omitted. Lastly, ML models appear more appropriate for big data or real-time data applications, as they are able to handle well large sample sizes.

Most studies on crash prediction focus on either (i) classifying road events as crashes or non-crashes or (ii) predicting the injury severity of a crash, while very few studies deal with crash frequency prediction [27-30]. A Support Vector Machine (SVM) regression model to predict the frequency of crashes on rural roads was developed and compared to both a NB model and neural network model [27]. The authors concluded the SVM performed better than both models and they also noted that the implementation of SVM is relatively easier and faster compared to the other two model types. A similar analysis conducted by Dong et al. found that a neural network model that uses a NB model as one of its layers performs better in predicting crashes compared to both a neural network model without a regression layer and an SVM model [28]. In [29] it was shown that a single deep belief network could be trained globally with multiple datasets coming from a diverse set of roads (e.g., rural roads vs motorways from different regions) to predict the expected crash frequencies with a performance at least comparable to the traditional NB model. A multivariate, piecewise regression technique namely adaptive regression splines, was applied to build a crash prediction model and then, estimate CMFs based on it [30].

ML models to on crash frequency prediction are very few however, available approaches include SVM models or some form of neural networks. Neural networks have been found more challenging to develop as they require a lot of training, but at the same time they might be more effective. One limitation of the ML models is that they do not have a parametric form and so, are seen as "black boxes" and it is hard for safety analysts to interpret these models and develop countermeasures. One approach is to conduct sensitivity analysis for all the explanatory variables and use this information for determining which factors are more influential. In [31] the authors compared various ML classifiers for predicting injury severity level. They conducted a sensitivity analysis to see which ones of the explanatory variables are more impactful on the models' outcome. They increased the mean value of each explanatory variable (one variable at a time) by one standard deviation and recorded the proportion of each injury severity level before and after the perturbation of a variable.

## 4. Discussion

From the review it is clear that crash prediction modelling research has been very active in the last decade, and an extensive wealth of pertinent literature exists for exploration and exploitation by road safety practitioners. This section focuses on issues related to the implementation and usability of available models and the discussion aims to cover practical implications.

### 4.1 Transferability

The extend of existing research on microscopic crash prediction modeling is partially attributed to the limited transferability of the developed models and CMFs and so, even when calibrating the existing models to different conditions they have been found to have poor performance compared to newly developed models. For example, most of the State Departments of Transportation (DOT) in the US have developed their own (state-based) SPFs and CMFs instead of using the (calibrated) ones from the HSM. It was also found that when developing new models, safety researchers and practitioners are in favor of multivariate models, from which CMFs can also be estimated, e.g., the case of Pennsylvania DOT [4], the European PRACT project [32]. One limitation of the development of multivariate CPMs is their limited transferability. While the SPF-CMF approach can be potentially

transferred and calibrated, this less likely for multivariate CPMs as aspects like regression form and model specification are subject to change in addition to the model parameters.

### 4.2 Model specification

From the review it becomes evident that one of the most important aspects in crash prediction modelling, often not gathering the full attention of researchers and not adequately explained in relevant publications, is the rationale behind the choice of explanatory variables. Although largely depending on data availability (a variable cannot be included if there is no data for it), it is in many cases decisive for the model's performance. The inclusion of a variable for exposure is considered essential, as well as an assessment of the level of correlation between explanatory variables.

### 4.3 Practical implications

Crash Prediction Models are valuable tools for the road safety practitioner and decision-maker, as they effectively link risk factors to crashes in a quantitative way. CPMs also provide the added benefit of the proactive approach, i.e., identifying and improving hazardous locations before crashes occur, ultimately saving additional human lives and preventing injuries. Elvik (2003) identifies the use of CPMs coupled with Empirical Bayes as the most effective approach for safety analysis [1]. However, mainly the development of CPMs and CMFs as well as their implementation require advanced statistical skills let alone the data collection efforts. A recent study that reviewed national road authorities across Europe reported that only 30% of the involved authorities rely on CPMs for their safety management process [33]. Therefore, for the wider adoption of CPMs road agencies and authorities need to work closely with researchers and experts in statistical modeling, to better understand the applicability of CPMs and identify ways to select appropriate ones (e.g., crash type-specific, severity-specific, road type-specific) or develop new ones. Additionally, useful tools for practitioners are repositories with CMFs (CMF Clearing House) or (simple) modeling tools (e.g., AASHTOWare Safety Analyst [34] or Austroads Road Safety Engineering Toolkit [35]) have the potential to increase the adoption of CPMs.

## 5. Conclusions

This paper summarized the findings and challenges of current efforts on microscopic crash prediction models, focusing on road segments. More than 100 scientific papers, project report and guidelines were reviewed and synthesized. The most important resource on microscopic crash prediction modeling is the Highway Safety Manual Predictive Method however, many additional studies have been carried out to supplement it and/or address its limitations. While the HSM illustrates a stepwise process on the development of crash prediction models, i.e., by developing SPFs and CMFs, and provides a set of SPFs and CMFs for various road segments (e.g., two-lane roads, freeways, etc.), and provides guidelines on how to calibrate existing SPFs and CMFs, their transferability to different conditions is not always feasible. Researchers both in the US but also internationally have found that developing of new crash prediction models instead of using the HSM-based ones yields in better prediction performance. Therefore, at the international level, there are numerous efforts to develop multivariate crash prediction models and use these models to obtain CMFs. An important step during this process is related to the final model specification. This is data-dependent but at the same time, a combination of statistical expertise, engineering judgement and multiple trials are needed to conclude to the final set of explanatory variables. Paths for new research should focus on developing transferable models and assessing the transferability of existing multivariate models, while it is of equal importance to develop guidelines for using machine learning models for crash analysis.

## Acknowledgment

This research is co-funded by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation (call name: RESEARCH-CREATE-INNOVATE, project code: T7ΔΚΙ-00253)

## References

1. Elvik, R. State-of-the-art approaches to road accident black spot management and safety analysis of road networks. 2007. Oslo: Transportøkonomisk institutt.
2. American Association of Highway Transportation Officials (AASHTO). Highway safety manual, 1st Edition. 2010. American Association of State Highway and Transportation Officials, Washington, DC, USA.
3. American Association of Highway Transportation Officials (AASHTO). Highway safety manual, 1st Edition. 2014 American Association of State Highway and Transportation Officials, Washington, DC, USA.
4. Li, L., Gayah, V.V., and Donnell, E.T. Development of regionalized SPFs for two-lane rural roads in Pennsylvania. 2017. Accident Analysis and Prevention 108 (2017), pp.343–353.

5. Gooch, J.P., Gayah, V.V., and Donnell, E.T. Safety performance functions for horizontal curves and tangents on two lane, two way rural roads. 2018. Accident Analysis and Prevention 120 (2018), pp.28–37.
6. Gross, F. and Hamidi, A. Investigation of Existing and Alternative Methods for Combining Multiple CMFs. 2011. Report produced in the context of T-06-013 Highway Safety Improvement Program Technical Support.
7. Gross, F. Application of Multiple CMFs. 2019. Presentation in the FHWA CMF webinar, December 16, 2019.
8. NCHRP. Guidance for the Development and Application of Crash Modification Factors, Final Report. 2017. National Cooperative Highway Research Program 17-63.
9. Dadvar, S., Lee, Y.-J., and Shin, H.-S. Improving crash predictability of the Highway Safety Manual through optimizing local calibration process. 2020. Accident Analysis and Prevention 136 (2020) 105393.
10. Hauer, E. Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. 1997. Emerald Group Publishing Limited.
11. Persaud, B. and Lyon, C. Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. 2007. Accident Analysis & Prevention, 39(3), pp.546-555.
12. Elvik, R. The predictive validity of empirical Bayes estimates of road safety. 2008. Accident Analysis & Prevention, 40(6), pp.1964-1969.
13. Persaud, B., Lan, B., Lyon, C. and Bhim, R. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. 2010. Accident Analysis & Prevention, 42(1), pp.38-43.
14. Federal Highway Administration. A Guide to Developing Quality Crash Modification Factors. 2010. Federal Highway Administration. Report No. FHWA-SA-10-032. Gross F., Persaud B., Lyon C.
15. Elvik, R. Developing accident modification functions: exploratory study. 2009. Transportation research record, 2103(1), pp.18-24.
16. Hauer, E., Bonneson, J.A., Council, F., Srinivasan, R. and Zegeer, C. Crash modification factors: foundational issues. 2012. Transportation research record, 2279(1), pp.67-74.
17. Ivan, J.N., Wang, C. and Bernardo, N.R. Explaining two-lane highway crash rates using land use and hourly exposure. 2000. Accident Analysis & Prevention, 32(6), pp.787-795.
18. Yan, Y., Zhang, Y., Yang, X., Hu, J., Tang, J. and Guo, Z. Crash prediction based on random effect negative binomial model considering data heterogeneity. 2020. Physica A: Statistical Mechanics and Its Applications, 547, p.123858.
19. Da Costa, J.O., Jacques, M.A.P., Pereira, P.A.A., Freitas, E.F., and Soares, F.E.C. Portuguese two-lane highways: Modelling crash frequencies for different temporal and spatial aggregation of crash data. 2018. Transport, 2018, 33(1), pp.92–103.
20. Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., and Persaud, B. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. 2010. Accident Analysis and Prevention, Vol.42, pp.1072-1079.
21. Ambros, J., Turek, R., Brich, M. and Kubeček, J. Safety assessment of Czech motorways and national roads. 2020. European transport research review, 11(1), pp.1-15.
22. Ahmed, M., Huang, H., Abdel-Aty, M. and Guevara, B. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. 2011. Accident Analysis & Prevention, 43(4), pp.1581-1589.
23. Wang, C., Quddus, M.A. and Ison, S.G. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. 2011. Accident Analysis & Prevention, 43(6), pp.1979-1990.
24. Caliendo, C., Guida, M. and Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention, 39(4), pp.657-670.
25. Montella A, Colantuoni L, Lamberti R. Crash prediction models for rural motorways. 2008. Transportation Research Record. 2083(1):180-9.
26. Caliendo, C. Crash prediction models for roads including rainfall and hazardous points. 2017. International Journal of Civil Engineering and Technology, 8(9), pp.477-485.
27. Li, X., Lord, D., Zhang, Y., and Xie, Y. Predicting motor vehicle crashes using support vector machine models. 2008. Accident Analysis & Prevention, 40(4), pp.1611-1618.
28. Dong, C., Shao, C., Li, J., and Xiong, Z. An improved deep learning model for traffic crash prediction. 2018. Journal of Advanced Transportation.
29. Pan, G., Fu, L. and Thakali, L. Development of a global road safety performance function using deep neural networks. 2017. International journal of transportation science and technology, 6(3), pp.159-173.
30. Haleem, K., Gan, A. and Lu, J., 2013. Using multivariate adaptive regression splines (MARS) to develop crash modification factors for urban freeway interchange influence areas. Accident Analysis & Prevention, 55, pp.12-21.
31. Zhang, J., Li, Z., Pu, Z. and Xu, C. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. 2018. IEEE Access, 6, pp.60079-60087.
32. La Torre, F., Domenichini, L., Meocci, M., Graham, D., Karathodorou, N., Richter, T., Ruhl, S., Yannis, G., Dragomanovits, A. and Laiou, A. Development of a transnational accident prediction model. 2016. Transportation research procedia, 14, pp.1772-1781.
33. Yannis, G., Dragomanovits, A., Laiou, A., Richter, T., Ruhl, S., La Torre, F., Domenichini, L., Graham, D., Karathodorou, N., and Li, H. Use of accident prediction models in road safety management-an international inquiry. 2016. Transportation Research Procedia 14 (2016) 4257-4266.
34. Safety Analyst. American Association of Highway Transportation Officials. Available at: https://www.aashtoware.org/products/safety/safety-overview/
35. Austroads Road Safety Engineering Toolkit. Austroads. Available at: http://www.engtoolkit.com.au/