

## AN ADVANCED WEB-BASED SUPPORT SYSTEM FOR ROAD SAFETY DECISION-MAKING

Fabio Galatioto\*, Mario Catalano\*\*, Nabeel Shaikh\*\*\* and Erik Nielsen\*\*\*\*

### ABSTRACT

*This paper illustrates the results of the research conducted under the MAIA (Models and methods for collision prediction and Impact Assessment) project to develop a web-based platform for performing road safety simulation and assessing crash-related externalities.*

*After outlining the cutting-edge of accident prediction science as well as the state of the art of web-based road safety simulation, the paper describes an application in the United Kingdom, resulting in the development of the novel MAIA toolkit. This uses a set of three collision prediction models deriving from the comparison of a number of statistical parametric and machine learning methods (count data, ordered multinomial and artificial neural network models), along with an impact assessment tool. The MAIA prediction models estimate the frequency of road collisions and their casualty effects: number and severity of injuries; whilst the MAIA impact assessment tool has been designed and tested to estimate collision impacts (travel times, costs and emissions).*

*The prediction models adopt a statistical parametric framework, which has proved to perform better than non-parametric methods and have been validated on a national as well as subnational (cluster of UK districts) level.*

*In the Asia-Pacific region, hundreds of thousands of people die every year due to road accidents. In line with the United Nations “Decade of Action for Road Safety 2011-2020” Resolution, a web regional network has been created for the exchange of road safety best practices within the Asia-Pacific area. The idea behind this paper is to provide a scientific and robust set of tools, which potentially could be integrated through a web-based application to estimate the parameters of context-specific accident prediction models and support road safety decision-making at different scales.*

**Keywords:** road safety decision-making, road accident prediction, web-platform, statistical parametric models, machine learning.

### INTRODUCTION

Given the rise of motorized transport throughout the world in recent years, road accidents analysis and prevention are now a priority challenge for policy makers. In fact, in 2010 the United Nations (UN) General Assembly proclaimed the 2011-2020 period as “the Decade of Action for Road Safety” (Resolution 64/255), with the aim of stabilizing and then reducing the forecasted levels of road fatalities on a global scale (United Nations, 2010). To achieve this goal, the World Health Organization and the United Nations regional commissions, in cooperation with the United Nations Road Safety Collaboration and other stakeholders, have been asked to prepare a Plan of Action for the Decade as a guiding document to support the implementation of its ambitious objectives, fostering road safety management expertise, improving the safety levels of transport networks and vehicles, influencing the behaviour of road users accordingly and enhancing post-crash response to name few. In addition, Resolution 64/255 invited

---

\* Senior Technologist, Transport Systems Catapult, Milton Keynes, United Kingdom, E-mail:

[fabio.galatioto@ts.catapult.org.uk](mailto:fabio.galatioto@ts.catapult.org.uk)

\*\* Scientific Consultant, Transport Systems Catapult, Milton Keynes, United Kingdom, E-mail: [mariocat73@gmail.com](mailto:mariocat73@gmail.com)

\*\*\* Senior Technologist, Transport Systems Catapult, Milton Keynes, United Kingdom, E-mail:

[nabeel.shaikh@ts.catapult.org.uk](mailto:nabeel.shaikh@ts.catapult.org.uk)

\*\*\*\* Senior Technologist, Transport Systems Catapult, Milton Keynes, United Kingdom, E-mail: [erik.nielsen@ts.catapult.org.uk](mailto:erik.nielsen@ts.catapult.org.uk)

the World Health Organization and the United Nations regional commissions to coordinate regular monitoring, within the framework of the United Nations Road Safety Collaboration, of global progress towards meeting the targets identified in the plan of action through global status reports on road safety and other appropriate monitoring tools (World Health Organization, 2010).

Accident prediction modelling can provide a solid basis for road safety decision-making, in line with a culture of scientific evidence-based planning: accident prediction models can help public authorities to identify the actual sources of road hazard and estimate the potential effects of policies and management actions, in terms of crash frequency reduction and severity mitigation.

A recent study (Yannis et al., 2016) points out that most organizations involved in road accident prevention – in Europe, the Asia-Pacific region and the United States of America – do not use prediction models for road safety decision-making on a regular basis: in more detail, although about 60 per cent of the organizations interviewed (national road authorities, managing companies, academia/research institutes and highway consultants) usually derive their choices or proposals from a comparative analysis of alternative road safety measures, they rarely or never employ accident prediction models during the assessment process. This is probably due also to the shortage of easy-to-use applications and toolkits to support policy makers and practitioners in road safety promotion by the use of accurate predictive models. In fact, as the following section shows in detail, from the technological point of view, the state of the art mainly consists of simple web databases of effective safety measures (Yannis et al., 2016), which raises a serious gap hindering the global advancement of road safety levels.

This paper describes the early results of an ongoing research that, in an attempt to address the above gap, proposes a web-based platform (MAIA - Models and methods for collision prediction and Impact Assessment - toolkit) for road safety policy/measure simulation at both national and local scale. The core of this web application will consist of models to predict the main dimensions of road collision phenomena (crash frequency, number of casualties and severity) and to assess impacts of collisions in terms of non-casualty effects, such as congestion, travel delays and emissions.

In its current form, the MAIA toolkit incorporates only models for accident prediction, which derive from a complex development, validation and selection process: more than 300 models from both the microeconometrics domain and the machine learning area have been evaluated.

For the specific case of the Asia-Pacific region, the current availability of a regional web-based network to facilitate the regular exchange of road safety best practices amongst government officials and practitioners (Asia Pacific Road Safety Network) makes this research particularly considerable for the region's stakeholders. In fact, this network could be the first step of a process to harmonize and share road safety data within the area, thus creating the basis for the development of a system of accident prediction models to be integrated into the regional web network as a MAIA-like application. Every collision aspect (number of casualties, severity, etc.) could be modelled with a clustering approach, which means building several cluster-specific simulation tools for the same aspect, each cluster consisting of data from similar countries/areas. This might yield the advantage of increasing significantly the information employed to model every specific crash dimension, in terms of sample size and wealth of predictors.

The paper, in section 2, explores the scientific literature about road accident modelling, with regard to crash frequency as well as casualty and traffic congestion effects; furthermore, section 2 outlines the main web-based decision support systems for road safety decision-making. Section 3 illustrates the novel MAIA toolkit and is followed by the description of the underlying accident modelling process in section 4 and the travel time impact estimation tool in section 5. Finally, conclusions are drawn and the future steps of the research are pointed out in section 6.

## **I. LITERATURE REVIEW**

This section outlines the scientific literature on road accident and related external costs modelling and ends with the description of the most advanced web toolkits to support road safety policy-making and management. Exploring the relevant state of the art has revealed that, so far, safety scientists have mainly employed, as road accident predictors, transport variables such as traffic flows and speed, along with road geometry parameters. However, a more comprehensive approach is

emerging and may spread in the near future for the increasing availability of road operation data<sup>1</sup>, special conditions at the crash location (like, for instance, roadworks), carriageway hazards (like, for example, the presence of vehicle load on road), user behaviour (such as driving under the alcohol influence), weather conditions, land use in the area around the accident scene, etc.

The accident prediction models resulting from the research illustrated in this paper, instead, extends the set of explanatory variables far beyond the traditional one, so as to include many of the above-mentioned regressor types.

Moreover, unlike the MAIA toolkit, in general, the existing web-based road safety decision support systems do not use complex predictive models and are at most complementary to external accident models developed as single regressive equations for standard site configurations. In such cases, the web app suggests crash modification factors to adjust the crash frequency estimated for baseline conditions with external models, so as to take into account the real problem's peculiarities (Yannis et al., 2016). This is the reason why the MAIA platform, conceived as a comprehensive simulation toolkit for road safety policy and management support, represents an advancement of the current state of the art.

### A. Accident prediction models

In general, road accident prediction models are used to investigate the association between the likelihood of a certain level of crash frequency or severity and a set of potential explanatory variables, so as to elicit guidelines for public authorities to prevent collisions and alleviate their effects. So far, researchers in the field have explored a number of traditional and novel statistical methods to analyze and forecast road accidents, with particular regard to microeconometrics and machine learning models. However, they have mainly focused on transport predictors, such as traffic flows and speed, along with road design variables.

As to the application of econometric models, often the variable of interest it is of *count* nature (e.g. number of collisions, number of casualties, etc.), which requires the use of specific models of count data (nonlinear methods). The output is usually a positive probability function (e.g. probability of observing one, two, ... casualties in a single crash). This means that the number of events to forecast (e.g. likely number of casualties per crash) is estimated as the average of a random variable.

To cite just one important example from the relevant literature, consider the RIPCORDER-iSEREST research project, whose aim was promoting best practices in road accident forecasting according to the state of the art (Reurings et al., 2005). Based on a thorough investigation into the accident modelling science, generalized linear models using Poisson or negative binomial probability distributions were proposed as effective prediction tools. Also, the specific dimension of road crash severity has been investigated diffusely with statistical parametric approaches. In this case, the variable of interest can take one of several mutually exclusive outcomes (e.g. slight injury or severe injury or fatality) and, therefore, an appropriate and widely used prediction method is logistic regression (Cameron and Trivedi, 2005; Menard, 1995). An interesting example is the study by Bedard et al. (2002), who employed multivariate logistic regression to analyze crash severity and discovered that female drivers drinking, not using seatbelts and travelling at higher speeds increase the likelihood of fatal accidents.

Over the last years, various safety scientists have analyzed road collision phenomena by nonparametric methods, with particular regard to the application of machine learning to accident severity classification. In general, machine learning is employed to extract patterns from big datasets and build predictive models, without the need of assuming a specific mathematical form for the relationship between the variable of interest and its predictors.

The classification and regression decision tree (CART) is one of the most accredited machine learning methods in accidentology. An interesting example can be found in the study by Kashani and

---

<sup>1</sup> As stated in Yannis et al., 2016, "road operation data are a little less common than road design data" and, when collected, in general, "are also not publicly available".

Mohaymany (2011), who applied the CART approach to investigate rural accidents in Iran and found out that forbidden overtakes and not using seat belts are the most influential factors of collision severity.

Another notable nonparametric approach in accident severity research is the artificial neural network (ANN), as demonstrated in the work by Sohn and Shin (2001), which compares the ANN paradigm to the CART and logistic regression methods for road crash severity modelling in Korea. They showed that classification accuracy does not differ significantly across the three models and protective devices (seat belts and safety helmets) are the most determining predictors of accident severity.

## **B. Assessment of accident-related congestion costs**

Estimating the cost of road collisions has been an active topic of discussion and has been studied quite thoroughly during the past few decades (Wijnen, 2017). One of the most thorough studies retrieved was developed by the European COST313 project (Alfaro et al., 1994). This project proposed guidelines regarding the cost components that should be included as well as the methods to estimate them. Although the externality of congestion due to road collisions is mentioned, unfortunately no relevant estimation methodology is presented.

Nonetheless, only a few studies trying to explicitly quantify the cost of congestion due to road collisions could be retrieved (Chen et al., 2016; BITRE, 2009; Dowling et al., 2004; and Hallenbeck et al., 2003). The most notable is the one conducted by the Bureau of Infrastructure, Transport and Regional Economics of Australia (BITRE 2009). In this study, a deterministic queuing model was used to estimate delays due to crashes. These queuing models assume that vehicles move uninterruptedly through a bottleneck up to a certain capacity. If the number of vehicles (flow) exceeds the capacity, then queues start to build-up. Making assumptions regarding the capacity of all links in the network, the average flow on these links depending on the time of the day, the value of time of the affected road users, the effect of crash severity in terms of delay and the response time of authorities, they managed to estimate the cost of congestion due to road collisions between \$500 million and \$1.76 billion. This wide range is explained by the significant number of assumptions made to come up to this estimate. The congestion cost percent value had been estimated to be around 10 per cent of the overall accident cost in an earlier study by the Bureau of Transport Economics in Australia (BITRE 2000). Few other similar but more simplified approaches were also retrieved (de Leur et al. 2010; Scottish Natural Heritage 2011).

The key point from all the relevant studies to the estimation of road crash-related delay costs is that, so far, they rely on simplistic assumptions regarding the network conditions (flows on links) and the actual delay caused by accidents, thus their results can be very informative but not entirely accurate.

## **C. State of the art of web-based road safety decision-making support**

Exploring the pertinent literature led to find out four noteworthy web applications in the road safety field: the Federal Highway Administration CMF Clearinghouse databank (Gross, 2010; Gross, 2011; Yannis et al., 2016), the AustRoads Road Safety Engineering database (Jurewicz, 2010; Yannis et al., 2016), the iRAP Road Safety database (Yannis et al., 2016) and a recent crash risk prediction map from the Indiana State Police (Indiana State Police, 2016).

The first is directly related to the Highway Safety Manual (HSM), probably the most important guide for accident forecasting (AASHTO, 2010; AASHTO, 2014). The HSM provides predictive methods to estimate the crash frequency (by severity or collision type) of a network, facility or specific spot. The estimations are obtained with simple regression models developed for a set of base sites and named safety performance functions (SPFs). The SPFs depend typically upon only a few variables, mainly average daily traffic volumes and some roadway characteristics. The prediction performed by a safety performance function is to be adjusted if the actual features of the considered site differ from the base conditions of the model. To do this, crash modification factors (CMFs) are employed; these are particularly useful to estimate the effectiveness of safety measures. A number of CMFs are included in the HSM, but many others are provided by complementary guides and web-based platforms like the Federal Highway Administration CMF Clearinghouse databank, funded by the United States of America Department of Transportation. The CMFs provided by this platform are categorized on the basis of the site characteristics (area type, number of lanes and traffic volume) and the class of collision (angle crash, run-off-road collision, etc.).

The AustRoads Road Safety Engineering databank derive from extensive investigation into the effectiveness of several tens of road safety treatments in Australia and New Zealand. It offers transport planners the possibility of analysing accident phenomena by three criteria: crash type (head-on, rear-end, ...), safety deficiency (road lighting inadequate, unclear priority, ...) and road user (motorcyclists, pedestrians, ...). Once a specific criterion is chosen, the related collision phenomena are illustrated, along with relevant engineering and non-engineering countermeasures. Furthermore, each treatment type is hyperlinked to a more detailed description including costs, benefits, implementations issues, technical references and a general measure of effectiveness in terms of crash frequency reduction percentage.

The iRAP Road Safety databank, through ViDA (iRAP online software) allows to access content including mapping and tables of results which have been built from road inspections and assembled from surveys of many kinds as a result of the collaboration between the International Road Assessment Programme (iRAP), the Global Transport Knowledge Partnership and the World Bank Global Road Safety Facility, is very similar to the AustRoads one, even though the former provides less precise measures of treatment performance. In fact, the effectiveness of each proposed treatment is evaluated on a four scale system: 0-10 per cent, 10-25 per cent, 25-40 per cent, 60 per cent or more. Furthermore, the treatment cost (in qualitative terms) and life (in years) are indicated.

In the end, in 2016, a new web application was launched by the Indiana State Police. This provides a Daily Crash Prediction Map showing with color-coded grids how the accident risk varies across the state road network during the current day. A machine learning algorithm is used for risk prediction on the basis of historical crash data, road conditions and characteristics, traffic volumes, information on residents and workers, time of year and day, etc. However, no simulation feature is available to assess the impact of alternative intervention scenarios on road collisions. Law enforcement officials can examine the map to pinpoint the hazard hot spots and, if necessary, move first responders to the highest risk areas. Travellers can use the map to check if their planned routes are in high risk zones, thus requiring re-routing or very cautious driving behaviours.

## II. THE MAIA WEB APP FOR ROAD SAFETY POLICY AND MANAGEMENT SIMULATION

The MAIA web platform, which was developed in the first instance for the United Kingdom, gives easy and secured access to road collision model-based simulation, thus facilitating research into policy and management scenarios to improve road safety. Potentially, it promotes accident analysis and related impact assessment across different spatial and time scales and can be considered as enabling technology for data analytics in the accidentology field.

The platform is very flexible, since open source software was used for its development, which facilitates the successive integration of new features and off-the-shelf components. The users are provided with a web interface with models grouped into three categories: delay cost model for assessing the collision impact on travel times (under development), crash rate model to estimate crash frequency for a small urban or rural area, casualty and severity model to predict the number of casualties per accident and its severity. The users can select any of the three categories and apply the related models for simulation. They can also create projects, link them to a specific model, determine the set of predictors and estimate the model's parameters with new observations or with subsamples of the built-in datasets, in order to obtain updated and/or context-specific versions of the selected model (for example, changing from the national scale to a subnational area).

The MAIA toolkit consists of three main core elements (figure 1):

- 1) a comprehensive and multi-layered **integrated database** containing information on traffic volumes, speed, fleet composition, accidents as well as other data covering high spatial and temporal resolution. The web application programming interface offers the possibility of adding/removing user selected data to develop up-to-date and/or context-specific models.
- 2) A **central engine** including the different mathematical models for both prediction and assessment purposes. The prediction models, in particular, are the focus of this paper.
- 3) A **visualization layer** with useful options to visualize both inputs (from the database) and outputs (from the central engine). The visualization tool and its spatial analysis capabilities are based on GIS type tools.

The MAIA web interface contains several interactive buttons, each dedicated to a function. The main functions already included are as follows: *My datasets*, which connects to existing databases; *Input*, to import new data; *Pre-process*, which provides a table format visualization and statistical analysis tools; *Models*, to invoke the available accident prediction models and finally *Projects*, containing two sub-sections, namely *PRJ-variables* and *Results*. The latter two are better represented in figure 2 and 3 respectively.

In particular, figure 2 displays the current toolkit version with regard to the input setting session of the casualty and severity model function, where the user can define a simulation scenario in terms of a specific combination of several roadway and context variables' values (speed limit, junction type, main road class, day of week and time of day, weather conditions, presence of hazards on road, etc.). Figure 3, instead, on the right-hand side, shows the output of a simulation with the available casualty and severity models. In particular, at the top, a line plot illustrates how the probability of occurrence changes as the number of casualties per collision increases; at the bottom, the expected number of casualties (weighted average of all possible amounts of casualties, with weights equal to the corresponding likelihoods) is presented, along with the probabilities for every possible level of accident severity (slight: at least one slightly injured individual; serious: at least one seriously injured person; fatal: at least one death).

As mentioned previously, a powerful feature was added, that is the ability for the MAIA platform to estimate the accident prediction models' coefficients on its own. A coefficient determines the impact magnitude of the related predictor (e.g. speed limit) on the response variable of interest (e.g. probability of a serious crash). Hence, the toolkit is able to generalize the mathematical framework of each model taking its parameters as unknown and estimate these with new data. This implies flexibility and space scalability: in fact, the models' coefficients can be adjusted over time if structural changes occur (e.g. ways police assign injuries to serious and slight classes) and new context-specific models can be developed for those areas where collision phenomena follow anomalous patterns.

Figure 1. MAIA toolkit architecture

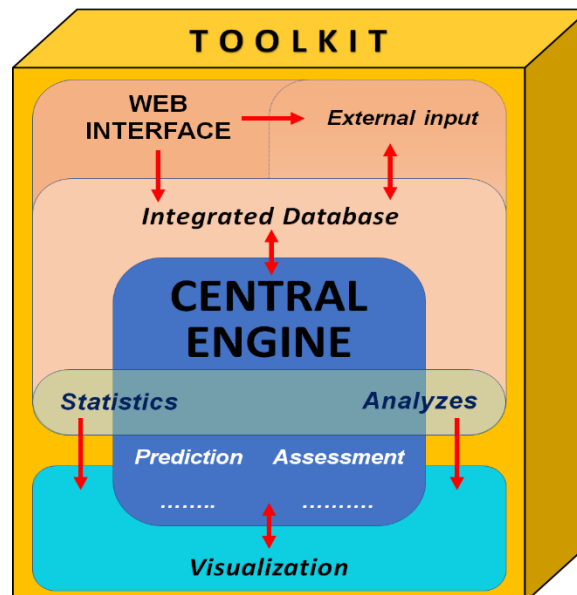


Figure 2. Example of input variable setting within the MAIA platform

**INPUT**

- PRE-PROCESS
- PROJECTS
- PRJ-VARIABLES
- RESULTS
- VISUALISER

**MODELS**

- PROJECTS
- Casualties
- Accident Rate

**Speed Limit:** 3

**Day of Week:**

- ☒ Monday
- ☐ Tuesday
- ☐ Wednesday
- ☐ Thursday
- ☐ Friday
- ☐ Saturday
- ☐ Sunday

**Time Of Day:** 1

**First Road Class:**

- ☒ Motorway
- ☐ A(M)
- ☐ A
- ☐ B
- ☐ C
- ☐ Unclassified

**Junction Detail:**

- ☐ Not at junction
- ☒ Roundabout
- ☐ Mini-roundabout
- ☐ T junction
- ☐ Slip road
- ☐ Crossroads
- ☐ More than 4 arms
- ☐ Private drive
- ☐ Other junction

**Weather Conditions:**

- ☐ Fine no high winds
- ☒ Raining no winds
- ☐ Snowing no winds
- ☐ Fine + winds
- ☐ Raining + winds
- ☐ Snowing + winds
- ☐ Fog or mist
- ☐ Other

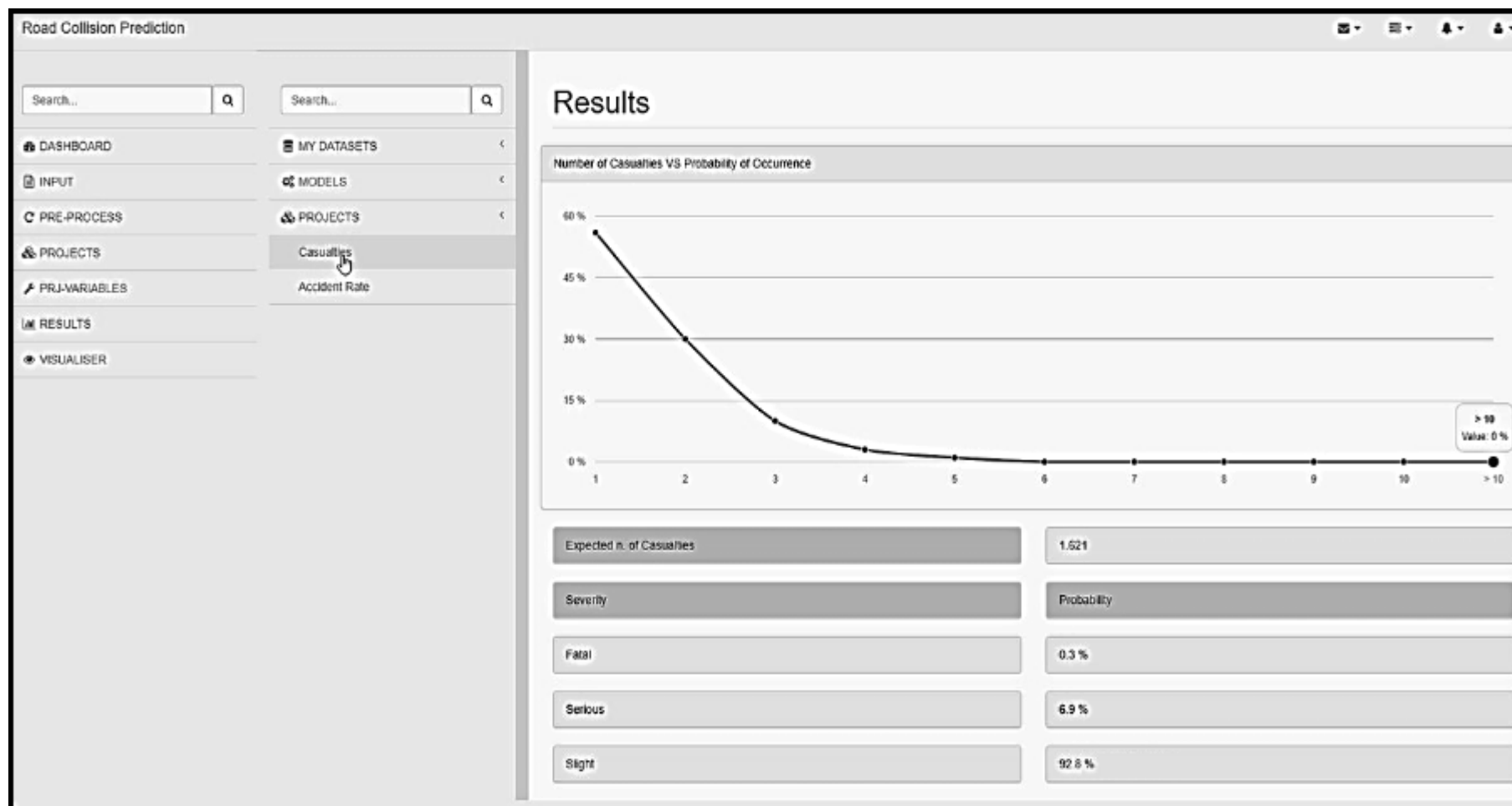
**Special Conditions at Site:**

- ☐ None
- ☐ Traffic signal out
- ☐ Signal defective
- ☒ Road sign defective
- ☐ Roadworks
- ☐ Surface defective
- ☐ Oil or diesel
- ☐ Mud

**Previous Accident:** 1

**Constant:** 1

Figure 3. Example of simulation outcome within the MAIA platform





### III. DEVELOPMENT AND SELECTION OF PREDICTION MODELS FOR THE TOOLKIT

As stated above, the set of prediction models embedded in the MAIA toolkit derives from comparing more than 300 models from both the statistical parametric and machine learning domain.

All main dimensions of collision modelling were covered: collision frequency at a disaggregate census level (Lower Layer Super Output Area or LSOA), number of casualties per collision and crash severity at the most disaggregate level (road link/junction).

The above two modelling domains were deeply explored. In particular, within the statistics area, count and multinomial econometric models were tested for collision/casualty occurrence forecasting and severity classification, respectively. As concerns the machine learning area, for all accident analysis dimensions, the artificial neural network framework was employed, since it has been gaining wide interest as data mining tool over the last decades. Moreover, the ANNs, besides being evaluated as stand-alone prediction tools, were examined also as components of ensemble models. Ensembling (or pooling) is a novel methodology in the forecasting science and is basically a multi-model approach that combines the predictions of different models to improve accuracy.

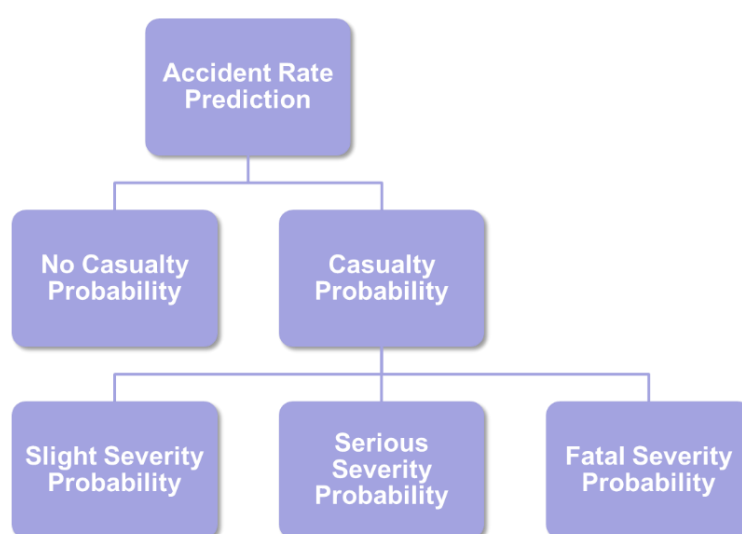
#### A. Methodological issues: problem definition, selection of predictors and data sources, model validation

The problem under investigation was addressed with a multi-step approach (figure 4):

1. Prediction of the number of road collisions per year;
2. Estimation of the probability of non-injury versus injury road collisions, so as to split the predicted number of crashes by the type of impact (only damages for vehicles versus casualties), and forecast of the number of casualties per collision;
3. Estimation of the probability of each severity level, so as to split the predicted number of accidents causing casualties among three possibilities: at least one slightly injured individual, at least one seriously injured person, at least one death.

As to the estimation of non-injury versus injury accident risks, due to the limited amount of data on non-injury crashes available at present, this part of the modelling approach is still in progress and only the model to assess the number of casualties per collision could have been developed.

**Figure 4. Multi-step approach for the accident prediction problem**



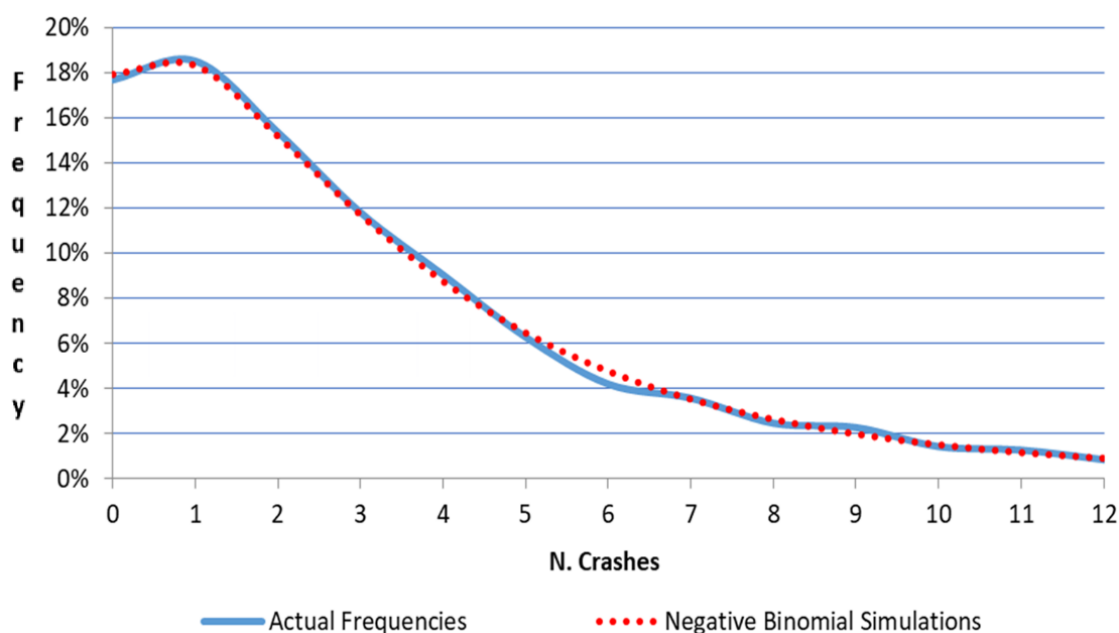
*Notes:* No Casualty: only damages for vehicles; Slight Severity: at least one slightly injured person; Serious Severity: at least one seriously injured person; Fatal Severity: at least one dead person.

In more detail, 32,844 records of collision annual frequencies (2015) for the Lower Layer Super Output Areas in England and related potential predictors - characterising the LSOA land use and road traffic, along with its population profile in terms of density, gender, age and deprivation - were analyzed and used for collision frequency modelling. Casualty and severity modelling, instead, was based on the STATS19 database, an official set of observations mainly pertaining to collision circumstances across the United Kingdom (specifically England and Wales). In particular, 140,056 records of 2015 STATS19 data were employed to model casualty occurrence and accident severity as functions of the following set of variables: traffic; speed limit; day of week and time; main road class; junction type, if the accident happens at an intersection (roundabout, crossroads, etc.); presence and type of pedestrian crossing facility (zebra crossing, footbridge or subway, etc.); weather conditions; special conditions at site (traffic signal out, roadworks, etc.); carriageway hazards (vehicle load on road, previous collision at the same site, etc.).

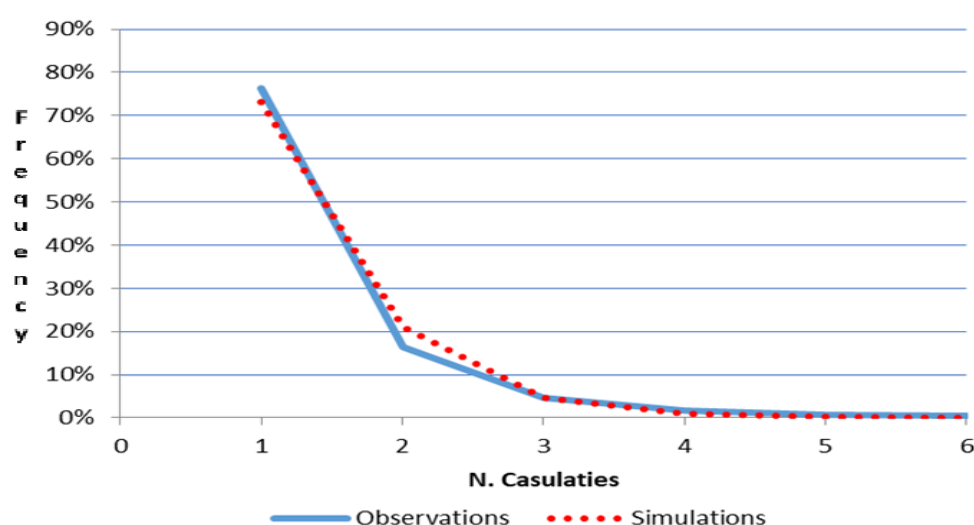
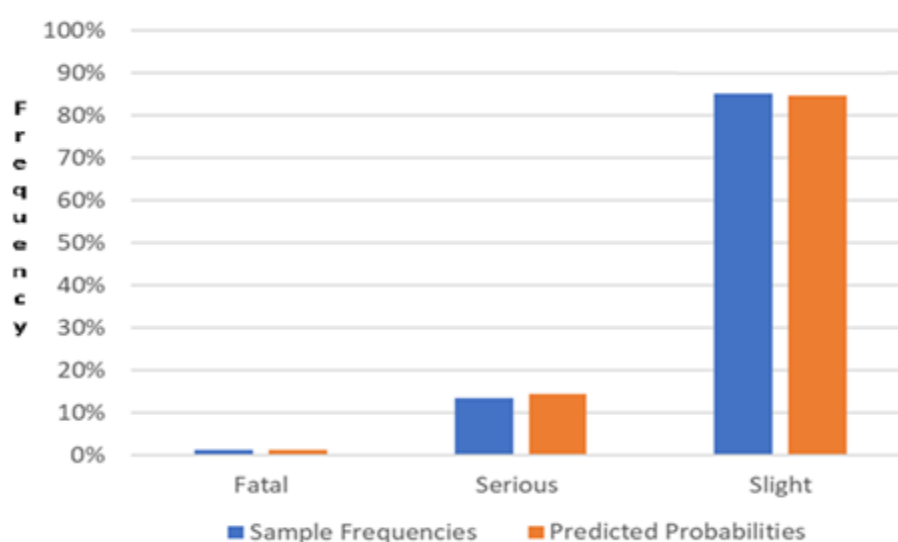
To evaluate the prediction accuracy of each model, an appropriate method is to analyze and compare the actual (from the real data) and fitted (estimated by the models) relative frequency<sup>1</sup> for every observed value of the dependent variable, in relation to a time interval different to that considered for model estimations (Cameron and Trivedi, 2005). This means, for example, comparing the percentage of real cases of serious accidents, in a testing year across the UK, with that estimated by the relevant model and check how close the prediction is to the real life.

The use of the above accuracy metric has highlighted that the statistical models - specifically, a collision rate negative binomial model, a casualty Poisson model and a severity ordered Probit model - perform better than the machine learning alternative methods, at least if compared with artificial neural networks when these are employed as stand-alone prediction tools. On the contrary, when the ANNs are combined to form ensembles of models, in very few cases, they perform a bit better.

**Figure 5. Validation of the model to predict the number of road crashes per LSOA**



<sup>1</sup> Number of cases/total of observations.

**Figure 6. Validation of the model to predict the number of casualties per crash****Figure 7. Validation of the model to predict the accident severity**

Based on a set of data different to that employed for model estimation, figures 5, 6 and 7 illustrate the validation process outcome for each of the three selected statistical models: in detail, for every value of the dependent variable, the simulated relative frequency is compared against the observed one. A satisfactory degree of forecasting accuracy emerges from model testing.

As anticipated, non-parametric and ensemble models were built and compared with the parametric ones, in relation to a test sub-sample (30 per cent of the 2015 database), for each of the accident aspects investigated (crash occurrence, number of casualties per collision and severity). In particular, for every value of the considered response variable, the comparison was based on the average percent difference between the observed relative frequency and its simulation (mean absolute percent error or MAPE).

In the crash occurrence case, the parametric approach (negative binomial model) performed better at simulating the occurrence of collisions across the LSOAs, if the number of accidents is in the 0-5 range (inclusive); whereas, beyond the 5 collisions per LSOA threshold, the non-parametric approach (ensemble of neural networks) turned out to be more accurate. However, overall the non-parametric method is a bit superior: its MAPE is 4.68 per cent against a MAPE of 5.48 per cent of the negative binomial model.

As to the number of casualties per crash, the parametric approach (truncated Poisson model) obtained better estimates for the first four levels (1-4 injuries per accident), which anyway accounted for 99 per cent of the test set of observations: the truncated Poisson average MAPE was 29 per cent against a value of 52-53 per cent of the neural models.

In the end, as regards accident severity, the parametric alternative (ordered Probit model) proved the best at forecasting the frequency of fatal and slight events (with MAPEs equal to, respectively, 48 per cent and 87 per cent of those for the best non-parametric model), while the frequency of serious collisions was best predicted by an ensemble of neural networks, whose mean absolute percent error was about 58 per cent of that calculated for the parametric model.

## **B. Models' strengths and weaknesses**

Although the statistical approach proved considerable accuracy at national level, as demonstrated in Figures 5-7, it revealed less effective in those contexts where road collision phenomena turn out to be exceptional, thus moving away from the mean behaviour on a national scale. In line with recent and seminal advancements in the relevant research field (Alikhani et al., 2013; Catalano and Galatioto, 2017; Sohn and Lee, 2003), a solution could be a self-managing model framework (metamodel). The proposed Metamodel will enable, for each specific combination of input variables, the selection of the most suitable option within a set of alternative collision prediction models, developed for a wide spectrum of situations. In this way, when unusual collision risk factors occur in a certain area, the proposed approach could select an available forecasting model that should have been designed for a cluster of similar areas where that set of factors is not unusual.

## **IV. ROAD COLLISION IMPACT ASSESSMENT TOOL**

This section focuses on the description of a novel and more accurate impact assessment method to evaluate road collision-associated impacts in terms of travel time delays. The analysis was based on data about traffic volumes and speed, as well as road collision and street work temporal and spatial information. In more detail, the data employed derive from several sources: MIDAS (Traffic volumes and speed on the primary road network), ITN OS (road and infrastructure GIS layers), STATS19 (national database of injury road accidents) and Highways England (primary road injury and non-injury road accidents).

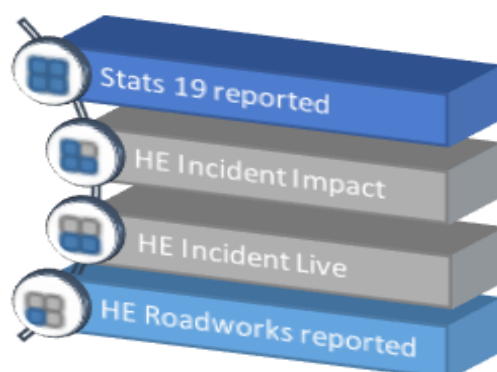
The new method can provide an accurate assessment of the delay, as total or individual vehicle time delay due to road collisions. Furthermore, it was developed to be highly transferable, with a view to making it reusable also for other types of disruption events, such as street works and breakdowns.



The development process was based on two initial phases: a temporal analysis stage to examine the time series of traffic data (speed, flow, etc.) and identify typical profiles and anomalies in the traffic behaviour (outliers), to be compared at a 15 minute resolution; a spatial analysis step to reconcile the output from site-specific on street traffic sensors (MIDAS) with the national Integrated Transport Network (ITN) link-based dataset. This resulted in a coherent and as much continuous as possible network with links associated with traffic data.

Then the available information on road collisions (from STATS19 and Highways England, which provides also non-injury collision data) was used to estimate the impact in terms of travel time delay due to individual events. It is important to note that, since the temporal analysis phase has identified anomalies (outliers) independently of the relative causes, this method can provide, as a by-product, the total impact in terms of delay from non-recurrent congestion on the monitored transport network.

The assessment was performed within an appropriate radius of influence (up to 25 km from the site of the collision) to capture anomalies around the event and not only along the main corridor affected. It was also tested using a junction-to-junction (or main node-to-main node) approach, so as to guarantee transferability to linear events such as street works.

The resulting impact assessment tool is currently based on Excel and the outputs of the analysis can be displayed and filtered by different variables. The following figure illustrates the power and usability of the method.

**Figure 8. Example of the assessment tool symbology**

The series of symbols in Figure 8 should be exploited by the reader to identify different types of road accident and events. For example, the symbol  means that, in a specific time and date, an accident reported by the STATS19 source has occurred; while the symbol  refers to the Highways England dataset, in which accidents are associated to duration values: this means that the considered symbol is reported for every time step of 15 minutes in which the accident has potentially caused an impact.

### A. Analysis of road collision impact on travel time

In figure 9, a road collision event is assessed using the impact analysis tool. From the symbols, it is clear that an accident, registered in the STATS19 database, happened in the 18:45-19:00 interval and caused a significant delay wave (measured in vehicle-hours), as the series of horizontal histograms and associated numeric values shows for a 5 km radius of influence. Observing figure 9, there does not seem to be any other collision reported in the previous few hours or after the last delay registered.

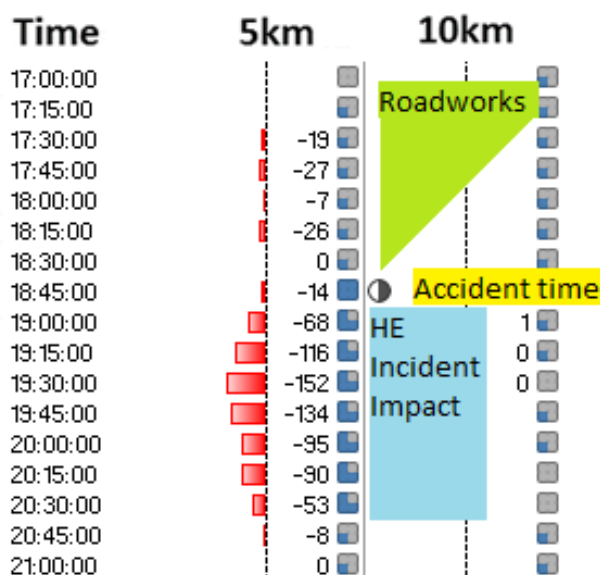
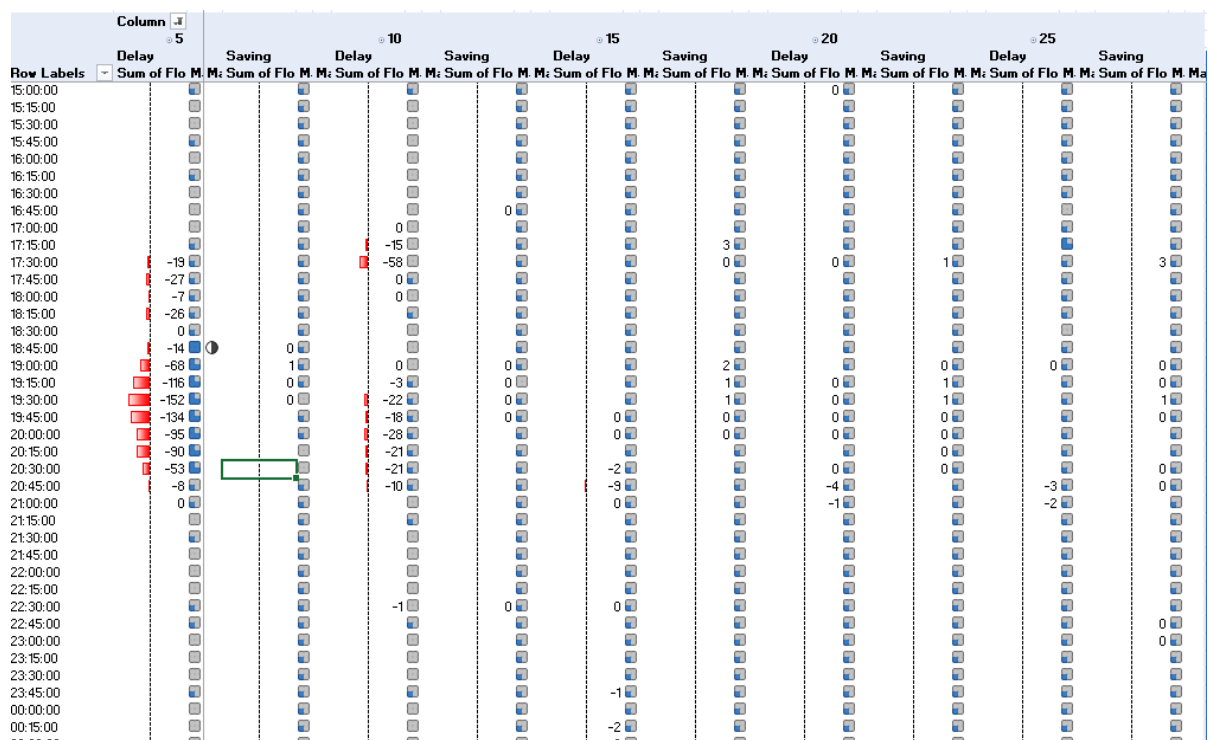
**Figure 9. Example of impact diagram for a short distance (5 km) from the collision site**

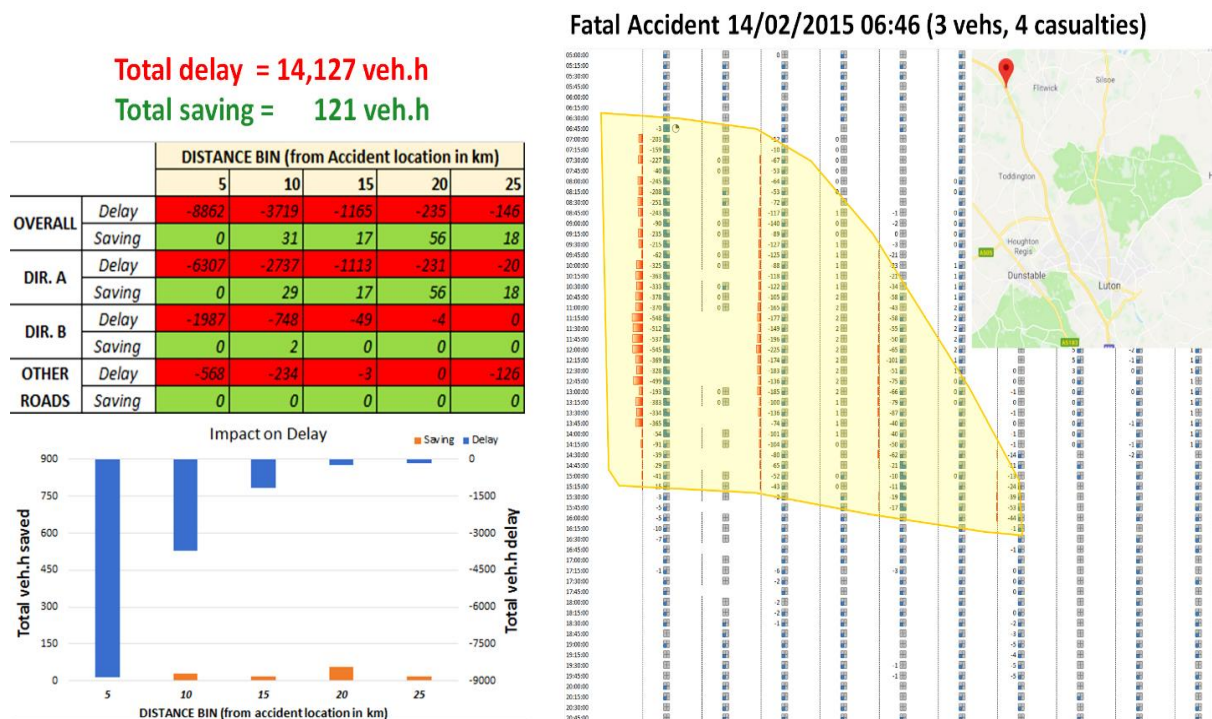
Figure 10 extends the impact analysis of the above accident event to a set of distance bands of influence (ranging from 5 km to 25 km) on a 15-minute step basis. Time savings and delays are calculated for an eight hour time window (from 15:00 to 24:00).

Figure 10. Example of impact analysis for a range of distances from the collision site



The impact assessment tool has also other functions allowing, for example, the analyst to quantify accident-related effects on travel time by direction and type of road, including “other roads” like slip roads, as represented in the summary of an event impact and its temporal-spatial evolution showed in figure 11.

Figure 11. Example of impact analysis by direction and type of road



## **B. Impact assessment tool's strengths and weaknesses**

The above method for impact analysis showed robustness and very promising results in capturing and quantifying both negative effects (delays) and benefits (savings) associated to known disruption events (namely road collisions reported by independent sources) as well as unknown events, thus giving a full picture of the impact from non-recurrent congestion. Notwithstanding, the identification of the temporal and spatial extent of the event (figure 11) is still subject to a degree of manual scrutiny and is, therefore, expected to get mostly automated in the future through the application of machine learning techniques.

Finally, the impact assessment method described was tested mainly on the primary road network, with the consequence of neglecting some types of impact such as, for example, re-routing traffic using alternative roads outside the primary road network. This, however, is currently under investigation and it will probably be possible to assess and include re-routing effects by using area wide data (e.g. link-based speed data).

## **CONCLUSION AND FUTURE STEPS**

This paper presented a set of models for road crash prediction alongside the description of an innovative web-based platform for road safety policy and management simulation, the MAIA (Models and methods for collision prediction and Impact Assessment) toolkit. In the first instance, the MAIA toolkit has been designed for the United Kingdom.

The forecasting models incorporated into the MAIA platform cover all main dimensions of road collision modelling: crash frequency, casualty occurrence and severity. These models derived from comparing the latest developments of machine learning (non-parametric domain) to econometric methods (parametric domain).

The modelling work has found out that the statistical parametric approach outperforms the non-parametric one, at least if it is compared to artificial neural networks as stand-alone models. If, instead, these are combined with ensemble techniques, in few cases, they perform a bit better than their parametric counterparts. Hence, three microeconomic models - a collision rate negative binomial model, a casualty Poisson model and a severity ordered Probit model - were chosen to simulate the main dimensions of road collision phenomena, based on a comparative analysis of more than 300 models.

A powerful feature has been added to the MAIA toolkit in the late stage of the research, that is the ability to estimate the three above statistical models on its own. This means the possibility of updating models' coefficients over time, if structural changes occur, along with space scalability.

Although parametric methods proved very good performance at national level, they revealed less effective in those contexts where road collision phenomena follow rare patterns.

The toolkit component for the assessment of collision impact on travel times showed the ability to measure traffic congestion effects for the primary road network (Motorways and A-type roads). Throughout the testing phase, the maximum distance chosen (25 km from the reported location of the event) performed well.

In quantitative terms, it was observed that even one single fatal crash can cause delays equivalent to 2 months of impact from non-extreme accidents in the same section, and road collisions effects on the primary road network section studied accounted for up to 60 per cent of the overall non-recurrent congestion.

During the analysis, a fundamental role was played by the availability of different but complementary datasets (STATS19 and Highways England data, along with information on street works), and their integration was beneficial for the overall assessment and avoiding misinterpretation.

In the end, the results achieved are of potential interest for the Asia-Pacific region, as, within this area, sensitivity to road safety issues has been growing significantly since a web regional network was created to promote the exchange of best practices and expert knowledge amongst governments

and practitioners. In more detail, a certain degree of harmonization in road accident data production within the region could be the basis for the evolution of the above web network into a road safety decision-making support platform. The MAIA project could be, on the one hand, a paradigm for such a progress, on the other hand, the starting point towards the development of the web-based road safety decision support concept. In fact, a system of accident prediction tools could be integrated into the Asia-Pacific web network as a MAIA-like application, but the high variability of collision phenomena inside the region could lead, beyond the MAIA experience, to a forecasting model framework very effective even when faced with anomalous combinations of accident risk factors.

Further research will explore the benefit of a clustering-based multi-model approach for the accurate prediction of unusual manifestations of the phenomenon (e.g. extreme collision events). Moreover, other potentially explanatory variables will be tested, such as specific road design attributes (road curvature, road width, etc.) and driver as well as vehicle-related predictors (age, driver's gender, vehicle type, etc.); the authors also intend to experiment other machine learning methods as accident simulation tools.

## ACKNOWLEDGEMENT

This work has been undertaken as part of the research project MAIA funded by the Department for Transport (United Kingdom).

## REFERENCES

- AASHTO (2010). *Highway Safety Manual*, First Edition, American Association of State and Highway Transportation Officials, Washington DC.
- AASHTO (2014). *Highway Safety Manual*, First Edition, 2014 Supplement, American Association of State and Highway Transportation Officials, Washington DC.
- Alfaro, J. L., M. Chapuis, and F. Fabre (1994). *Socio-Economic Cost of Road Accidents: Final Report of the Action*, COST 313. Available at: <http://www.worldcat.org/title/cost-313-socio-economic-cost-of-road-accidents-final-report-of-the-action/oclc/846985044?referer=di&ht=edition>
- Alikhani, M., A. Nedaie, and A. Ahmadvand (2013). "Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran", *Safety Science*, vol. 60, pp. 142-150.
- Bedard, M., M. J. Stones, J. P. Hirdes, and G. Guyatt (2002). "The independent contribution of driver crash and vehicle characteristics to driver fatalities", *Accident Analysis and Prevention*, vol. 34, pp. 717-727.
- BITRE (2000). Road Crash Costs in Australia.
- BITRE (2009). Road Crash Costs in Australia.
- Cameron, A. C., and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.
- Catalano, M., and F. Galatioto (2017). "Enhanced transport-related air pollution prediction through a novel metamodel approach", *Transportation Research Part D*, vol. 55, pp. 262-276.
- Chen, Z., X. C. Liu, and G. Zhang (2016). "Non-recurrent congestion analysis using data-driven spatiotemporal approach for information construction", *Transportation Research. Part C*, vol. 71, pp. 19-31.
- De Leur, P., L. Thue, and Brian Ladd (2010). *Collision Cost Study*, Prepared for Capital Region Intersection Safety Partnership, Available at [http://drivetolive.ca/wp-content/uploads/2014/02/Collision\\_Cost\\_Study\\_Final\\_Report\\_Feb\\_2010.pdf](http://drivetolive.ca/wp-content/uploads/2014/02/Collision_Cost_Study_Final_Report_Feb_2010.pdf).
- Dowling, R., A. Skabardonis, M. Carroll, and Z. Wang (2004). "Methodology for measuring recurrent and non-recurrent traffic congestion", *Transportation Research Record*, vol. 1867, pp. 60-68.
- Gross, F., B. Persaud, and C. Lyon (2010). *A Guide to Developing Quality Crash Modification Factors. Federal Highway Administration (FHWA) Report*, FHWA-SA-10-032.



- Gross, F., and A. Hamidi (2011). *Investigation of Existing and Alternative Methods for Combining Multiple CMFs*, Highway Safety Improvement Program Technical Support, Task A.9, Final Technical Content.
- Hallenbeck, M. E., J. M. Ishimaru, and J. Nee (2003). *Measurement of Recurring Versus Non-Recurring Congestion: Technical Report*, Research Project T2695, Task 36, Congestion Measurement, Available at <http://depts.washington.edu/trac/bulkdisk/pdf/568.1.pdf>.
- Indiana State Police (2016). Indiana Crash Risk Map, Available at <https://www.in.gov/isp/ispCrashApp/main.html>
- Jurewicz, C. (2010). *Road Safety Engineering Toolkit*, ARRB Group Ltd., Available at <http://acrs.org.au/files/arsrpe/RS07033.pdf>.
- Kashani, A. T., and A.S. Mohaymany (2011). "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models", *Safety Science*, vol. 49, No. 10, pp. 1314-1320.
- Reurings, M., T. Janssen, R. Eenink, R. Elvik, J. Cardoso, and C. Stefan (2005). *Accident Prediction Models and Road safety Impact Assessment: a state of the art*, RiPCORD - iSEREST Consortium, Internal Report D2.1.
- Scottish Natural Heritage (2011). *The Indirect Costs of Deer Vehicle Collisions*, Final Report, Available at <http://www.snh.gov.uk/docs/C298276.pdf>.
- Sohn, S. Y., and H. Shin (2001). "Pattern recognition for road traffic accident severity in Korea", *Ergonomics*, vol. 44, pp. 107-117.
- Sohn, S. Y., and H. S. Lee (2003). "Data fusion ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea", *Safety Science*, vol. 44, pp. 1-14.
- Soudris, D., S. Xydis, C. Baloukas, A. Hadzidimitriou et al. (2015). AEGLE: A Big Bio-Data Analytics Framework for Integrated Health-Care Services, SAMOS 2015, Available at <https://doi.org/10.1109/SAMOS.2015.7363682> (IEEE), pp. 246-253.
- United Nations (2010). Resolution adopted by the General Assembly 64/255. Improving global road safety, Available at [http://www.who.int/violence\\_injury\\_prevention/publications/road\\_traffic/UN\\_GA\\_resolution-54-255-en.pdf](http://www.who.int/violence_injury_prevention/publications/road_traffic/UN_GA_resolution-54-255-en.pdf).
- Wijnen, W. et al. (2017). *Crash cost estimates for European countries, deliverable 3.2 of the H2020 project SafetyCube*, Available at [https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/24949/1/D32-CrashCostEstimates\\_Final.pdf](https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/24949/1/D32-CrashCostEstimates_Final.pdf).
- Wong, S. C., N. N. Sze, and Y. C. Li, (2007). "Contributory factors to traffic crashes at signalized intersections in Hong Kong", *Accident Analysis and Prevention*, vol. 39, pp. 1107-1113.
- World Health Organization (2010). Global Plan for the Decade of Action for Road Safety 2011-2020, Available at [http://www.who.int/roadsafety/decade\\_of\\_action/plan/plan\\_english.pdf](http://www.who.int/roadsafety/decade_of_action/plan/plan_english.pdf).
- Yannis, G., A. Dragomanovits, A. Laiou, T. Richter et al. (2016). "Use of accident prediction models in road safety management - an international inquiry", *Transportation Research Procedia*, vol. 14, pp. 4257-4266.