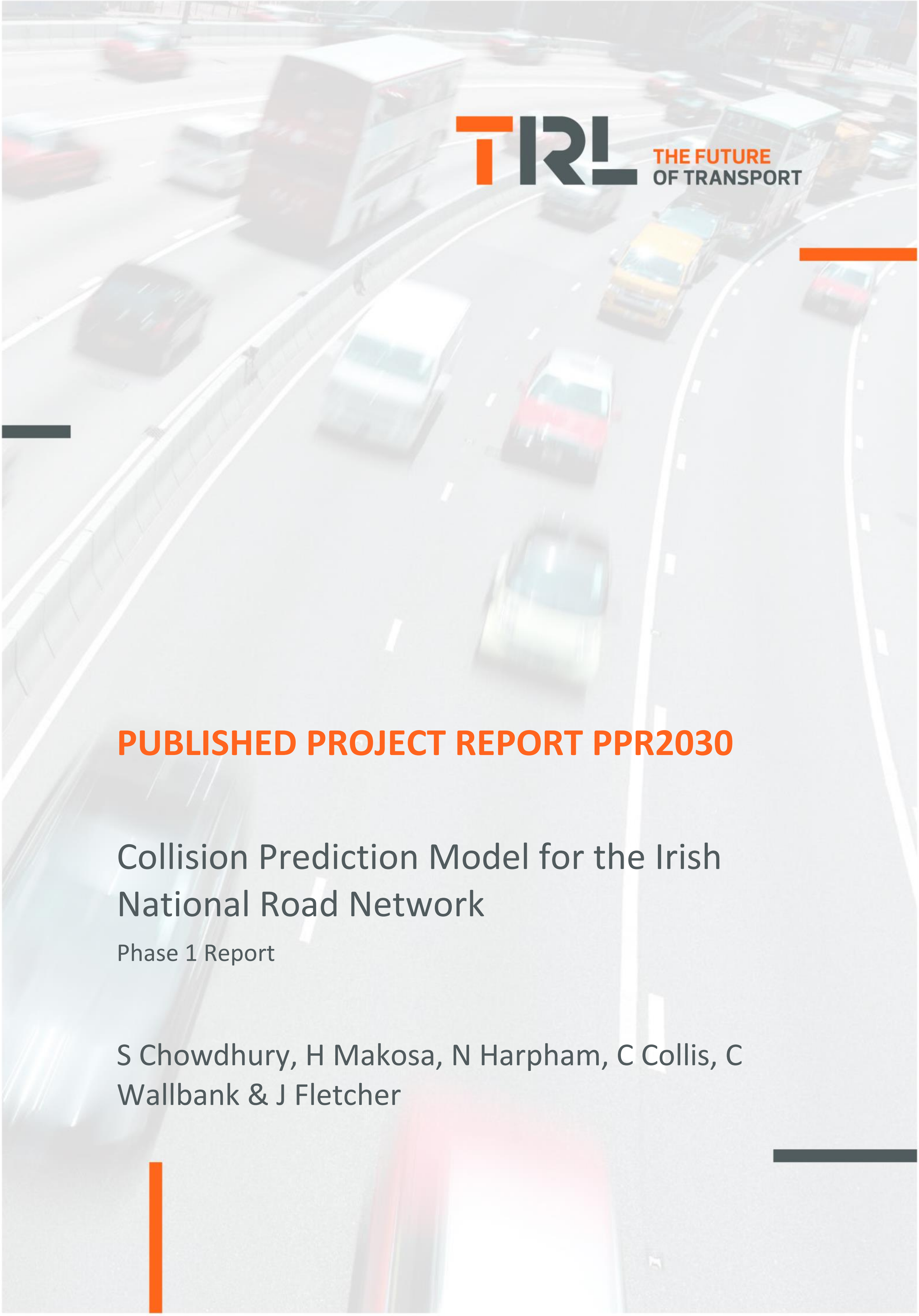


TRL
THE FUTURE
OF TRANSPORT

PUBLISHED PROJECT REPORT PPR2030

Collision Prediction Model for the Irish National Road Network

Phase 1 Report

S Chowdhury, H Makosa, N Harpham, C Collis, C Wallbank & J Fletcher

Report details

Report prepared for:	Transport Ireland Infrastructure, Suzanne Meade		
Project/customer reference:	TII268 Collision Prediction Modelling		
Copyright:	© TRL Limited		
Report date:	November 2023		
Report status/version:	Final - Version 4.0		
Quality approval:			
Warsame Mohamed (Project Manager)		Lynne Smith (Technical Reviewer)	LSmith

Disclaimer

This report has been produced by TRL Limited (TRL) under a contract with Transport Ireland Infrastructure. Any views expressed in this report are not necessarily those of Transport Ireland Infrastructure.

The information contained herein is the property of TRL Limited and does not necessarily reflect the views or policies of the customer for whom this report was prepared. Whilst every effort has been made to ensure that the matter presented in this report is relevant, accurate and up-to-date, TRL Limited cannot accept any liability for any error or omission, or reliance on part or all of the content in another context.

When purchased in hard copy, this publication is printed on paper that is FSC (Forest Stewardship Council) and TCF (Totally Chlorine Free) registered.

Executive summary

TII (Transport Ireland Infrastructure) wishes to develop Accident Predictive Models (APMs) in order to use these to assist engineers to better manage the safety of physical road features across its trunk network. The development of APMs is not a simple or cheap undertaking as they require processing, and potentially specific additional collection of a wide range of road and traffic data. The use of existing infrastructure and flow datasets can reduce the cost of developing APMs, however, the comprehensiveness and quality of data available for the modelling are very important. These data sets are then statically tested to understand the mathematical relationship of specific infrastructure elements to the collision numbers occurring on different road types.

Once the models have been developed using regression-based analyses approaches, the output can be used in a predictive way to assess how collision occurrence might change if the road elements were changed. The APMs can therefore be used by road designers and safety engineers to understand whether a road section is performing well with respect to safety, e.g. collision occurrence matches that predicted, or if it requires investigation as the collisions are excessively higher than might be expected on a road with those characteristics. How different road features in schemes designs will affect expected collision numbers can also be used in economic appraisal of proposed schemes. This is because the APM effectively generates estimates for local crash modification factors (CMFs) – for those variables included in final models (which individually account for significant amounts of the variance in collision occurrence). **Developing APMs which assist local engineers to manage infrastructure safety better is the over-arching aim of this project.**

The feasibility of developing successful APMs depends on a range of complex issues. Phase 1 of this project has investigated the potential for developing APMs for the Irish trunk network. The key dependencies are on the range of datasets that are currently available and also the densities of collision occurrence on different roads. Both these factors will significantly impact the likelihood that statistically significant and comprehensive models can be developed.

TRL, together with ARUP, have been commissioned to carry out a two-phase project (sandwiched by a breakpoint). This report covers the three tasks that constitute the first phase, prior to the breakpoint:

- Task 1 reviewed the statistical approaches that other researchers have used to develop APMs, with an emphasis on approaches applied on roads similar to those that makeup the Irish trunk network.
- Task 2 reviewed the range and formats of relevant data sources available for the Irish trunk network. Their quality, consistency and potential for linking to shorter road sections was a main focus.
- Task 3 brings the findings of Task 1 and Task 2 together to indicate, with as much certainty as possible, the potential that useful APMs can be developed for the Irish trunk network given the client’s main requirements and aims.

Based on the Task 3 report findings and recommendations (this report), a decision by the client will be made, with advice from the consultant, on whether it is sensible to proceed with development of APMs, and associated practitioners' tools, for Ireland (Phase 2).

Task 1 summary

Task 1 reviewed more than 25 published papers and reports. The aim was to understand the practical aspects of APM development (e.g. the datasets, variables and methods for assigning these to the network) and the statistical approaches used to develop the models.

The following summarises the main issues identified in Task 1:

Most papers reviewed used five or six years of collision data and modelled all injury collisions. The time period used is a trade-off between obtaining sufficient collisions per road segment for the modelling and reducing the impact of there being major differences in the road factors present over time. Having many road segments with zero collisions is problematic for statistical approaches. The length that is modelled crucially affects the average collision number per segment.

Traffic flow is always the most significant factor that explains collision occurrence in APMs. Using systematically collected flow data from modern counter stations (Annual average daily traffic, AADTs) is more cost effective for modelling rather than conducting specific traffic surveys.

The approach applied to divide the roads into segments which are used to model collision occurrence is very important to the overall results. It is important to define segments with relatively few zero collision counts whilst capturing enough variability in the other explanatory parameters. This 'zero inflation' causes problems for the statistical approach in assigning variance because segments with zero collisions may 1) actually be safer or 2) this may be a result of segments being short coupled with there being a generally low density of collisions. However, less variation in road features on longer segments (due to variation being averaged) may lead to models with poor statistical power to explain patterns in collision occurrence.

A recommended approach is to divide the network into segments with the same flows and/or other specific features, mainly curvature. Alternative approaches are to divide the road into segments of equal length, again selecting lengths which minimise numbers with zero collision.

The most commonly occurring variables that were significant in developed APMs were:

- Lane dimensions
- Shoulder dimensions
- Median dimensions
- Curvature related variables
- Gradient related variables

The most common statistical approach for the development of APMs was Generalised Linear Modelling (GLMs). Studies from the 1990's typically assumed that collisions followed Poisson distribution (mean equals to the variance in the distribution). More recent studies tended to assume a Negative Binomial distribution for collisions, which better describes these data when the mean and variance are not equal. Where significant numbers of segments have no (zero) collisions zero-inflated models were considered.

Main road types (motorway, dual carriageway and undivided roads) will have differing characteristics such as flow levels (presence, absence of shoulder, number of lanes etc.). For this reason most studies developed models for specific road types. Major junctions were generally excised from the main road link sections for the modelling. These can be modelled separately. Some studies modelled different collision types and/or injury types separately, however this will reduce collision numbers on the segments for the modelling process (leading to more zero counts).

Task 2 summary

Task 2 sought firstly to identify all possible sources of road infrastructure and traffic data available for the trunk network. The requirement was to assess the robustness, consistency and extent of the coverage for each data type. Ways to process the datasets so that values can be assigned to specific road sections were also evaluated.

14 main data sources were identified, obtained and assessed including network mapping data, pavement management system data, traffic data and collision data.

Collision records will be modelled as the response variable (the variable the model aims to understand/predict from the explanatory variables). These have co-ordinates which permit flexible linking to segments. Including damage only collisions in addition to injury incidents increases the total number of collisions on the network from 7,641 to 53,873. This changes the average number of collisions per kilometre from 1.5 to 10.2. Given the density of injury collisions alone there is strong reason to include the damage only collisions when modelling: with segments of length 100m, the number of zero collision segments is 67% compared with 10% for a length of 2km. In the absence of robust information on travel direction, it will not be possible to assign collisions to a particular side of the carriageway.

Aggregation of the potential explanatory variables will be required for assigning values to segments:

- Weighted averages may be used for variables such as AADT and speed
- Ranges or groupings may also be applied and rolling averages can reduce noise in the data (for example with curvature)
- Taking an average, minimum or maximum may be appropriate for geometric or road condition variables such as gradient and curvature
- Density values are more appropriate for counting the number of (minor) junctions or considering the number access points
- Most of the network is rural; urban sections are typically much shorter
- For safety barriers, a ‘% of segment covered’ variable is useful

- There may also be strong correlations between some variables (such as speed limit and peak speeds) which need to be investigated prior to building the model to avoid including correlated factors

When defining segments according to metrics such as curvature and AADT:

- Different thresholds may be required for different road types or regions to ensure that segments are of appropriate length
- Segment lengths (and therefore number of collisions) may vary greatly if using curvature or AADT; a minimum or maximum length threshold could be applied for greater consistency
- A combination of variables could also be used to define segments, though thresholds may need to be wider to ensure segments are sufficiently long

Task 3 summary

As indicted, Task 3 is a synthesis of the understandings gained from both Tasks 1 and 2. The following main recommendations have been developed.

Either Poisson or Negative Binomial Generalised Linear Modelling will be used as the main statistical method, depending on which of the distributions best fits the collision data. Alternative modelling approaches take account of time trends, but these require data for the explanatory road features for the time periods of collisions modelled, which are not available.

A zero inflated GLM approach may be applied if the number of modelled segments with zero collisions on them is high. However, it will be best to define homogenous road segments which are long enough to avoid this requirement, as this approach has disadvantages.

There are four main road types with specific characteristics on the road network identified by the client, which will be modelled separately, these being:

- Motorway
- Dual carriageway
- Non legacy single carriageway
- Legacy road (subnet 3 and 4)

This approach aligns with the findings from the literature, where separate models were developed by road type because many of the significant explanatory variables will correlate with this variable. The most obvious and important example is AADT: at certain flow levels undivided roads will have additional lanes added and will be divided (forming dual carriageway), and then at even higher flows these will be developed into motorways. These correlations between variables cause problems for the GLM approach for assigning the relevant variance, which leads to unstable models.

The mainline segments should be modelled removing sections with larger junctions present. To model junctions comprehensively would require far more flow and geometric data than is available for these currently and would always be a separate task.

A very important aspect of the project is how the road is divided into the short sections ('segments') that will be modelled. The recommended approach is to divide into segments with similar traffic flow (AADT) and curvature (likely in two states: straight/gently curving or curving). These characteristics will be used to give segments which are homogenous for these features. This approach for generating segments was trialled successfully in Task 2 on two small sections of the network.

To reduce the frequency of segments with a zero collision count, a minimum segment length will be imposed. A maximum segment length will also be considered.

Six years of collision data (2014 to 2019) has been identified as the most appropriate data source to be used for the collision modelling. This approximates to the generally applied five years used by other workers. Importantly these specific years avoids the impacts of Covid restrictions which would introduce additional unexplainable variance to the modelling process. Average AADTs will be developed for the six year period matching the years the collision data are taken from.

A clear recommendation is that all collisions, including damage collisions, are modelled since using injury collision numbers alone was shown to lead to very low average collisions per segment and too many zero collision segments. Another related factor is that collisions cannot reliably be assigned to particular carriageways or directions of travel. This means that both sides of divided roads will be modelled together and this requires that the features on both carriageways are represented in the model in the variables.

A simple base model will be developed initially including the key variables listed below:

	Variable
Base model	AADT
	Road segment length
	Number of lanes (where this varies)

The variables identified from Task 2 that can be included in the modelling are as follows:

Variable	Variable detail
Speed	Modelled AM peak and inter-peak speed
Road geometry and condition	Gradient
	Crossfall
	Radius (curvature)
	SCRIM value
	Junction density (major/minor, or by junction type)
Roadside features	Safety barrier: location
	Safety barrier: material
	Access density

	Urban or rural
Other variables that impact collisions	Proportion of traffic which are heavy vehicles
	Proportion of rear-end collisions

This table includes many of the variables identified to be most frequently tested to develop APMs such as, curvature and gradient related variables. Some are not available for inclusion, such as lane, hard shoulder and median dimensions. These however may not be relevant if roads are generally constructed to consistent standards meaning there is no variation in them.

Task 3 has identified that the model may assist safety engineers and designers to better understand the impacts of a range of interventions which will potentially relate to the parameter values generated in the various APMs. This does however critically depend on whether the final models include the parameters others have identified as significant variables; this cannot be known with certainty until the modelling is actually performed.

Phase 1 conclusions

Developing APMs is technically challenging. Until the Phase 2 models are completed, it cannot be guaranteed which variables will be identified as significant in the final models. It also cannot be guaranteed at this point exactly how much of the variation in collision occurrence will be explained by the models. However, we have identified that the available explanatory datasets can be linked to road segments, and these align well with those used by others who have successfully developed APMs on similar road networks.

Using damage collisions (which are understood to be reported well in Ireland) in addition to injury collisions, is indicated to give a high enough density of collisions on short segments so that a GLM approach to develop the APMs should be feasible. However, this approach is not entirely risk free. As indicated, specific variables of interest to TII which are tested may not be statistically significant in the final model. There is an option to produce 'practitioners' models' where the normal level for significance is relaxed. Another option to increase the practical benefit of the models to safety engineers, is that the final models could be used in conjunction with specific design elements of interest using CMFs derived from other sources in the tools that will be developed.

Table of Contents

1 Purpose of this project

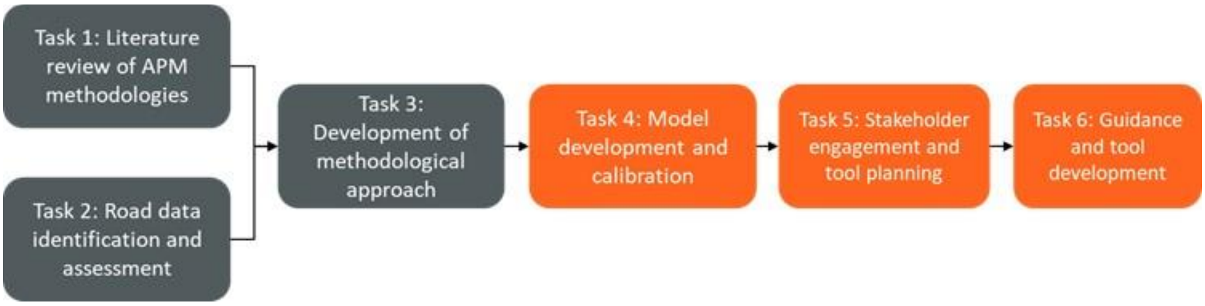
The aim of this work is to develop Ireland's first Accident Prediction Model (APM) to provide Irish Crash Modification Factors (CMFs) for the benefit of Transport Infrastructure Ireland (TII), local authorities and road safety practitioners who wish to identify effective road safety interventions and measures to reduce road traffic collisions. CMFs are also important for the economic appraisal of countermeasures.

Section 2 gives an introduction to APMs.

The project will investigate:

- The extent to which APMs can feasibly be developed to form the basis for the identification of effective infrastructure improvements, by helping staff to programme targeted, cost-effective and proactive interventions.
- How the data behind the APM development and the decision tool for practitioners, which will contain the models, can be enhanced to deliver timely and effective information into the future.
- An alternative option of implementing an APM calibrated using CMFs derived from other sources, to be considered following the decisions made in following Task 3.

There are six tasks to address these aims:


Task 1: Literature review of APM methodologies

Task 2: Road data identification and assessment

Task 3: Development of methodological approach

Task 4: Model development and calibration

Task 5: Stakeholder engagement and tool planning

Task 6: Guidance and tool development

This report covers the findings from the first three of these: Section 3 presents a discussion around the key findings identified from the literature review (Task 1). Section 4 presents the results of the data review to identify data sources available in Ireland for use in the modelling (Task 2). Finally, Section 5 assesses the feasibility of developing robust APMs for the Irish national road network and makes recommendations on the best approach given the methodological review and the data available (Task 3).

The later tasks (following a project breakpoint after Task 3), will develop, calibrate and validate the model(s) then incorporate these into a model for practitioners use. Suitable guidance for the tools use will also be developed and tested with the end users.

2 Introduction: Accident Predictive Models

This section introduces the basic principles of Accident Predictive Models (APMs).

2.1 The case for quantifying the safety of road features

The physical features present on roads, along with traffic flows and speeds are known to fundamentally influence the risks present to road users. There are two main forms of infrastructure elements. Firstly those designed and formally constructed (numbers of lanes, shoulder etc.), secondly other aspects including ‘natural’ features in the environment such what is present at the nearside (e.g. slopes/ trees). Features of the terrain can also lead to compromise in what is constructed; for example, sharp bends may be required to avoid immovable objects and the landscape can lead to extreme gradients. These challenges can potentially be ‘designed out’; however, that is not always possible due to economic constraints.

The kinds of features outlined in the previous paragraph will have differing levels of impact on road collision occurrence and the resulting casualties. It is therefore useful to quantify the relationship between different road features and the occurrence of road collisions and injuries. With this knowledge, decisions can be made to change design standards and practices and to remove and replace harmful features. Coupled with the costs of construction and maintenance more subtle decisions based on economic appraisal can be made to assist with scheme designs and prioritisation to optimise these for safety, given constraints on funds.

Understanding the relationship between collisions and the road features will help TII to deliver against the requirements of the EU Road Infrastructure Safety Management (RISM) directive, which requires network wide safety assessments be carried out and followed up by targeted road safety inspections or direct remedial action. In particular, this activity will enable TII to “identify road sections where road infrastructure safety improvements are necessary and define actions to be prioritised for improving the safety of those road sections” (EUR-Lex, 2019).

One approach to understand how design features affect road collisions and casualties is to perform a statistical analysis. Initially this was done as a straight ‘naïve’ comparison of the average collision rates before and after a change to a number of road segments was made. This does not result in a reliable estimate of the impact of a design element (see Hauer (Hauer, 2007) for a discussion of the problems of this approach). Hauer (Hauer, 2007) identified that many factors can change between the before and after periods, chief amongst these is the traffic flow which is known to have a fundamental impact on collisions (Elvik, Høye, Vaa, & & Sørensen, 2009). A better statistical approach is to include comparator road segments in a cross-sectional analysis. These comparator segments should have the same characteristics as the treated sites but were not altered. Non-parametric statistics such as the Chi Squared test can then be used; however, this approach still does not take account adequately of the range of variables that can affect collision occurrence. Elvik et al (Elvik, Høye, Vaa, & & Sørensen, 2009).

A well established and understood way to analyse the relationships between a number of (explanatory) variables/factors and the variation in a (response) feature is multiple linear regression. This is an extension of simple linear regression which fits a line of best fit to the data and works well where there is a single factor (x) affecting changes in the response parameter (y). Multiple linear regression extends this to understand how multiple explanatory variables and factors (x_1, x_2, x_3 etc.) influence the value of the response parameter (y). For example, such a model may explain how traffic flow (x_1), number of lanes (x_2) and the gradient of the road (x_3) influence the number of collisions (y). However, multiple linear regression makes a number of assumptions about the data, and other approaches have been shown to better model these type of data.

2.2 Why more complex statistical approaches are required to analyse road collisions

A common approach for modelling count data such as collisions or casualties is the use of generalised linear models (GLMs); one approach is Poisson regression. The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space, when these events occur with a known constant mean rate and independently of the time since the last event. This assumption is valid for collisions, which generally occur independently of previous collisions.

Another type of regression which is commonly used for count data is Negative Binomial regression. This is preferred over Poisson regression when the data are over dispersed (i.e. the variance in the data is higher than expected by the Poisson distribution). There are tests which can be performed on the data to determine which approach is better for a given dataset.

These models are also preferred over the linear regression approaches since they do not allow the model to predict a negative number of collisions or casualties. They can also be extended to model distributions in which there are frequent zero-valued observations (which may be appropriate for some collision/casualty models).

2.3 Crash Modification Factors

Crash Modification Factors (CMFs)[1] are used by road safety decision makers to understand the relative change in collision frequency due to a change in one specific condition. For example, they can be used to estimate the change in number of collisions due to an intervention, when all other conditions and site characteristics remain constant. The CMF is calculated as the ratio of the expected collision frequency after a measure is implemented to the estimated collision frequency if the change does not take place.

CMFs can be generated from APMs and from Before and After studies. They can be transferable between countries in some circumstances (OECD, 2012) but it is often better to

[1] These are also known as Accident Modification Factors and act as a multiplicative factor to compute the expected number of collisions after implementing a given improvement. The term ‘Crash Reduction Factor’ (CRF) is also common in the field of road safety and provides an estimate of the percentage reduction in collisions due to an improvement. CMF = 1 - (CRF/100).

Alternatively, a Crash Modification Function is an equation that calculates a CMF based on the characteristics of the site to which it will be applied. These are often used to determine the effects of interventions which alter site characteristics (e.g. an interventions which increases the lane width of a given section).

have locally derived and robust estimates of the impact of countermeasures. CMFs for various measures can be found on the Pract repository (https://www.pract-repository.eu/) and on Clearinghouse.(https://www.cmfclearinghouse.org/). Ireland does not currently have any county specific Crash Modification Factors.

3 Task 1: Literature review of Accident Prediction Model methodologies

This section presents the findings from the literature review of APM methodologies. The aim of this review was to:

1. Gather information on the practical aspects and methods used for the development of Accident Prediction Models, including the datasets used by other workers, and their sources, segment lengths modelled, variation in collision numbers in modelled segments and the variables used in the statistical modelling (Section 3.1).
2. Identify and understand the statistical approaches that have been applied to develop APMs using different datasets, and to assess how these perform given the nature of the available data (Section 3.2).
3. Understand the approaches that have been used to develop crash modification factors (CMFs) from APMs (Section 3.3).

An overview of the literature review method is provided in Appendix B.

Accident Prediction Modelling is a tool that allows road safety practitioners, road authorities and other organisations to quantify the relationship between aspects of the road system with the occurrence of collisions and/or casualties. Because APMs state the way that collisions and casualties will change with the presence or absence of different road features it is a way to predict how system changes will impact on safety numerically (in terms of collision/ casualty occurrence) (Yannis, et al., 2016). At their simplest APMs are regression equations which are generated by modelling the impact of just traffic flow and road segment length on collisions and/or casualties. These simple models are also called Safety Performance Functions (SPFs). These SPFs are developed and 'work' for distinct road cross section types. Their predictive power is limited since they estimate the expected collisions on road segments in relation only to traffic flow. However, traffic volume always explains by far the greatest variation in collision occurrence in APMs (Elvik, Høye, Vaa, & & Sørensen, 2009).

A wide range of physical road features are understood to influence safety in addition to the impacts of flow, from, for example, before and after statistical evaluations of design elements on roads. More complex APMs have also been formulated which identify the numerical impact of a wider range of road variables in addition to traffic flow. These models are typically regression models which have a mixture of variables and factors that are found to have a statistically significant association with collision occurrence. That is the parameters in the model all account for statistically significant amounts of the variation in the dependant variable (e.g. collisions/ casualties) and were consequently found to improve the overall model fit[2]. These more complex models allow the prediction of collision occurrence related

[2] Model fit refers to the predictive performance of the model. It is a measure of how well a statistical model generalizes to similar data to that on which it was developed. A well-fitted model produces more accurate predictions for the response variable compared to an under-fitted or over-fitted model which do not match the data accurately.

to changes in the infrastructure design elements present or absent on roads, providing that the relevant mathematical relationships have been derived from an APM (CEDR, 2013).

A main aim of the project is to develop APMs which will summarise the quantitative relationship between ideally a range of safety critical road design features and safety. For this reason, the literature review has focused on papers which develop APMs using more complex road geometry variables rather than the simpler base models (e.g. SPFs). The process also aimed to review materials that reported on practical examples of the approaches to develop APMs rather than works which sought to develop the theory of the methods in less applied ways.

3.1 Data collection and characteristics

The type and form of the data available for the modelling process is fundamental. The ideal situation is that the widest possible range of road features that vary to any extent would be tested in the modelling process. However, in practical terms this is not possible from a cost and effort perspective. When a limited number of the potential explanatory variables and factors are tested the final model fit may be poorer. This can occur when a variable that influences collision occurrence significantly is absent from the modelled dataset.

The data used in the development of APMs can be divided into several essential types. These being:

- Collision data – the dependant variable
- Traffic flow information
- Analysis segment length
- Variables and factors representing the (physical) road features

In general, most studies used similar approaches to collect this information.

3.1.1 Collision data

Historic studies conducted in the UK, primarily by TRL, tended to model around four to five years of collision data (Summersgill & Layfield, Non-junction accidents on urban single-carriageway roads, 1996), (Walmsley, Summersgill, & Payne, Accidents on modern rural dual-carriageway trunk roads., 1998), (Walmsley, Summersgill, & Payne, Accidents on modern rural dual-carriageway trunk roads. TRL report 335, 1998a), (Taylor, Baruya, & Kennedy, 2002), (Pickering, Hall, & Grimmer, 1986), (Walmsley & Summersgill, The relationship between road layout and accidents on modern rural trunk roads., 1998). This collision data was obtained from the UK Department for Transport's 'STATS19' system[3] using data collected from the police via standard reports that they fill for collision they attend.

Some of the TRL studies used periods of collision data longer than five years. For instance, (Walmsley, Summersgill, & Binch, Accidents on modern rural single-carriageway trunk roads,

[3] More recently the Department for Transport introduced an online self-reporting system for registering collision details with the police. This has been implemented across 23 forces and further information can be found here.

1998) and (Walmsley, Summersgill, & Payne, Accidents on modern rural dual-carriageway trunk roads. TRL report 335, 1998a) used more than 11 years of data to produce APMs for modern rural single-carriageway trunk roads and dual carriageway trunk roads respectively; however, the papers did not explain the reasoning behind this[4]. In these papers however, the authors did control for the background trends in accident rate through the addition of a variable representing the average yearly percentage decrease in collisions to the model.

The approach taken to collect collision information was similar across Europe and the US. Cafiso et al., (2010) developed APMs for motorway networks by using collision data from police reports in Spain for a five-year period. They were limited to collisions with at least one fatality or injury; this resulted in a total of 279 collisions on the 168.2 km network of two-lane local rural roads (with 640 injured persons and 16 fatalities) being included in the analysis. In another study Cafiso and D'Agostino, (2012) used six years of collision data (both fatal and injury) and empirically demonstrated that periods longer than five years could reduce the accuracy of the models as they were likely to introduce time-related trends that the traditional modelling approach might not be able to account for.

In the US, Labi (2011) used accident data from the Indiana State Police database from 1997 to 2000 (four years). This dataset included collision location, severity, type, and the assigned primary cause. Similarly, (Ambros, Havranek, Valentova, Krivankova, & Streigler, 2016) used six years of accident data but this analysis included all severity categories (including damage-only).

Things to consider for model development

- The period of collision data needed for the modelling
    - Most of the studies identified modelled around five years of collision data as this provided sufficient collision numbers, whilst not introducing time related trends.
    - The adequacy of the chosen period will also depend on the modelled segment lengths.
- Whether to model injury only collisions or include damage only incidents too
    - Most studies modelled injury collisions obtained from police reported crashes
- Whether to model collision or casualty numbers
    - All studies identified modelled the number of collisions; this is common practise as the number of casualties per collision can vary.

[4] Personal Communication (John Fletcher): this long time period used in modelling was selected to ensure few sections had zero collision counts; interpretation based on working with Summersgill.

3.1.2 *Traffic data*

In most of the TRL studies, which developed APMs for a range of junction types and non-motorway roads, traffic flows were generally manually sampled over 12-hour periods. The counts conducted on normal working days (between 07:00 and 19:00) were factored up to give annual average daily traffic (AADT). Traffic flow make-up by vehicle type and turning numbers at junctions were also recorded (Maher & Summersgill, 1996).

For modelling in the US, traffic flow data was taken from the roadway inventory dataset (Labi, 2011).

Another study conducted in Portugal, Vieira Gomes *et al*. (2012) aimed to develop methods to estimate the safety performance of various components of the urban highways system for vehicle collisions only (e.g. pedestrian-related collisions were excluded). In this study, data on geometric design characteristics, collision data and traffic volumes were collected at signalised and unsignalised intersections. However, apart from the collision data, which was available from the accident database, the traffic data and characteristics of the junctions were collected manually from on-site visits which meant a smaller sample of sites were modelled.

Things to consider for model development

- The availability and coverage of available traffic data
    - Some historic studies conducted by TRL collected traffic data using manual counters; this is likely to provide accurate traffic data at each site but is costly to collect.
    - Provided it is available for the whole network, using online (automated) databases will significantly reduce the development costs for APMs in Ireland.
- The format of the available traffic data
    - Annual average daily traffic (AADT) figures are commonly used in these models.
    - Depending on the section lengths over which these figures are calculated, and the segment lengths chosen for the model, the AADT figures may need to averaged (e.g. as a flow-weighted average).

3.1.3 *Division of the network into segments*

APMs for road networks are based on modelling the relationships between collision values as individual observations in road segments. In the following text we will use the terms sections and segments. Segments refer to sections of the road that were used for modelling (‘modelling segments’), whereas sections refer to the road sections defined in the raw data. The way in which these road segments are determined is therefore of great importance. Although not mentioned in every study, Labi (2011) and Cafiso et al (2010) both highlighted the need to have longer segments that were homogenous in nature to avoid having many segments with zero collisions. This suggests that most studies in this review avoided ‘zero-inflation’ models by having longer segments of road in the main statistical model.

Maher & Summersgill (1996) provided an overview of the methodology followed to develop APMs for a range of junctions including roundabouts (3 and 4 arm), rural T junctions and urban crossroads. The same technical approach for data collection and site selection was applied across all these studies. The first step was to conduct national reconnaissance surveys to identify suitable sites. Next, a stratified sampling approach based on the variables, pedestrian and vehicle flows, was used to randomly select study sites. The studies ensured that sites had not been modified over the duration of analysis period. Collisions that occurred at or within 20 metres of the junction were identified and used in the statistical models. The overall approach to site and link selection for non-junction studies was the same as those applied for the junction studies (Maher & Summersgill, 1996). Reconnaissance surveys were conducted to identify suitable sites based on vehicle and pedestrian flow. A stratified sample based on vehicle and pedestrian flow was selected from these sites, and speed limit and whether the link was one or two way was taken into account to ensure samples were representative.

More recent studies created road segments by identifying segments of varying lengths which had consistency in specific features to be modelled. This results in road segments of varying lengths. Turner et al. (2012) highlighted that models developed using homogeneous road segments were more likely to have better accuracy than those developed using fixed road segment lengths. The variables used to determine the homogeneity of segments vary by study.

A summary of the variables used to determine homogeneity of segments is presented in Table 1.

Table 1: Summary of variables used to determine homogenous road segments

Study	Variables	Method
(Ambros & Sedonik, A Feasibility Study for Developing a Transferable Accident Prediction Model for Czech Regions, 2016)	AADT Speed limit reductions Road category Number of lanes Paved shoulder	A change in any of the variables marked the end of a segment and the beginning of another segment.
(Garach, de Ona, Lopez, & Baena, 2016)	AADT Road Width Curvature Change Rate (CCR)	Segments with constant CCR were identified. For AADT and road width values intervals were selected. A new segment started when the value moved from one interval to another.
(Cafiso, Di Graziano, Di Silvestro, La Cava, & Persaud, 2010)	AADT CCR Average paved width Roadside hazard rating (RSH)	Segments where RSH values were considered constant were identified by minimising the sum of squared deviations of the RSH values with respect to the mean. A t-test with 15% significance level was conducted in this process. No information on the AADT, CCR and Paved with variables given.
(Turner, Singh, & Nates, 2012)	Radius	Segments were split into curved and straight elements using a 30m rolling average of the radius and the direction of the radius. Curve segments had an average radius less than 800m and all three 10m sections had the same direction. A straight segment occurred when the rolling average was greater than 800m or there

		were differences in the sign of the radius in the three 10m sections.
(Cafiso & D'Agostino, Safety Performance Function for Motorways using Generalized Estimation Equations, 2012)	AADT Curvature Slope and grade downhill Roadside hazard Viaduct presence Embankment presence	Segmentation was carried out to maintain all the variables constant within each segment.

Routes were divided into homogeneous segments by grouping 10m adjacent segments on the basis of whether they were straight or curved by Turner et al. (2012). This was determined based on a 30m rolling average of the radius and direction of the radius. Garach et al (2016) used a small subset of the available variables: AADT, average paved width, and curvature change rate (CCR), whereas Cafiso and D'Agostino (2012) used all available traffic and road geometry variables to determine homogenous segments, although the process applied was not defined in detail. Garach et al (2016) identified homogeneous road segments by identifying change in AADT, analysing distribution of paved width (both shown in Table 2) and using the formula below to identify segments with constant CCR:

$$CCR_{sect} = \frac{\sum_i |y_i|}{L_{HS}}$$

Where CCR is segment curvature change rate, y is deflection angle for a continuous element I (curve or tangent) and L_{HS} is road segment length.

A road segment was classified as homogenous when all three variables were constant. In addition, a minimum segment length of 2km was applied.

Table 2: Factors determining homogenous road segment (from Garach et al. (2016))

Variable	Range
AADT (veh/day)	[500-1000]
	[1000-3000]
	[3000-5000]
	[5000-10,000]
	> 10,000
Paved width (m)	<=5
	[5-6.5]
	> 6.5

Labi (2011) defined a two-lane rural road segment as a “section of road between major intersections or where there was a significant change in geometric characteristics”. Therefore, any road segment within 200 ft from an intersection was excluded from the analysis. Furthermore, any small segments (less than 0.1 mile) were excluded to avoid the possibility of zero-inflation of collisions in the dataset. In contrast to the studies above, Ambros et al.

(2016) segmented the road network between settlements. However, this approach led to complications in the modelling as there were road segments with less than 5 collisions on each segment.

Depending on the homogeneity criteria used, the resulting segments could have wide variety of lengths. Most studies set a minimum segment length. Ambros and Sedonik (2016) set a minimum length of 50 metres, Cafiso and D'Agostino (2012) set a minimum length of 70 metres, Ambros et al. (2016) set a minimum length of 50 metres, and Garach et al (2016) set a minimum length of 2 kilometres. Similarly, Ambros and Sedonik (2016) also set a maximum segment length of 500 metres. There were other criteria that were used to determine if segments were acceptable. In (Garach, de Ona, Lopez, & Baena, 2016) a minimum traffic flow of 500 vehicles per day was used to determine if segments were to be included in the data or not. An additional point to consider when creating segments was the removal of junctions. Some studies excluded junctions and segments of roads leading up to junctions. The distance from the junction up to which the road segments were excluded depended on the study, and this was usually calculated from the centre point of the junction. For instance, Vieira Gomes et al., 2012) (2012) used 40 metres, and Daniels et al. (2011) used 100 metres from the centre of the roundabout as the junction segment length.

Things to consider for model development

- The way in which road segments are determined impacts the number of collisions on a given segment and thus the modelling method needed.
    - If there are lots of segments with zero collisions, then modelling methods which account for zero-inflated counts may be needed.
- How to identify homogenous road segments (longer road segments that are homogenous in nature are less likely to have zero-inflated collision numbers).
    - The variables used to determine homogenous road segments varied depending on data collected and distribution of variables in the dataset.
        - AADT, road width and curvature change rate were the most common variables used to determine homogeneity
        - Other variables like road category, number of lanes and paved shoulder were sometimes used in conjunction with the variables above.
    - The method used to define homogenous road sections also varied:
        - In some studies, road segments could vary in length as long as selected features modelled were consistent within each section.
        - In other studies, curvature variables were used to split segments into straight and curved.
        - Some studies defined the links between each intersection as a segment.
        - In addition to the methods described above, it was common for a minimum section length to also be applied. This varied by study from anywhere between 50m and 2km.
        - Exclusion of junction sections: some studies removed junctions and varying lengths up to the junction.

3.1.4 Variables representing the physical features

Ambros and Sedonik (2016) discussed the variables that were included in the model and their form. In addition to AADT and segment length, the model included average curvature change rate, density of intersections with minor roads, density of roadside facilities, road width category, number of lanes, hard shoulder and speed limit reductions. The study also highlighted that exploratory analysis, such as cumulative residual graphs, might be necessary to decide the form of the variable. For instance, both segment length and traffic variables are commonly used in power form but could be included exponentially. The most appropriate form depended on the statistical distribution of data used for modelling purposes.

Another study (Cafiso, Di Graziano, Di Silvestro, La Cava, & Persaud, 2010) used GPS (Global Positioning System) survey and road safety inspections to collect road geometry data. It included the common variables such as curvature change rate, shoulder and median widths, lane width, number of lanes, and average paved width. They also included speed differential (a measure of variation between the segments and average operating speed)[5]. The study also included various context variables such as roadside hazard ratings (used to describe roadside conditions) and driveway density. In the follow-on study, Cafiso and D'Agostino (2012) included roadside hazard (values ranging from 1 to 6 in order of potential risk), slope of grade downhill, lack of cross slope of the segments analysed, variables relating to whether it was embankment or trench, and curvature of the road elements in the statistical models.

Summersgill (2000) summarised the variables used in the historic UK studies for building APMs on both junction and non-junction sites. These included geometric variables such as road width, hard strip characteristics, quality factor, hilliness coefficient and context variables such as the number of major and minor junctions.

In the US, Labi (2011) used roadway inventory and road alignment datasets to extract road geometry data. The variables to model for rural two-lane highways included lane width, shoulder width, friction number, pavement conditions, horizontal and vertical alignment of road segment. Their final road segment dataset was developed using spatial integration of the previously mentioned datasets using Geographical Information System (GIS) map layers. Another study conducted by Vogt and Bared (1998) extracted geometric data from various databases and photologs in the United States. Apart from AADT, the main variables collected in the study were lane width, shoulder width, number of driveways or intersections, shoulder type, lighting presence or absence, terrain information, weather conditions, and horizontal and vertical alignment.

Another study developed APMs for rural highway roads in New Zealand (Turner, Singh, & Nates, 2012). In this study, an extensive pilot study was used to identify 12 key variables to be included in the model out of 28 initially considered. These were AADT, unsealed shoulder width, seal width, combined point hazards, combined accesses, distance to non-traversable slope, average absolute gradient, average curvature, SCRIM coefficient and horizontal consistency (percentage change in speed), roadside hazard rating. The modelling also took account of regional differences/ similarities. The study grouped regions based on the following conditions:

- 85th percentile speed
- Regional under-reporting of serious collisions
- Percentage of state highway collisions in wet weather
- Percentage of state highway midblock 100km/h alcohol-related collisions
- Percentage of state highway midblock 100km/h collisions in dark conditions
- Percentage of state highway midblock 100km/h collisions relating to cornering

[5] JF: not clear to me what this means (note for SM)

Things to consider for model development

- The availability and coverage of explanatory variables to be included in the model
  - In the more complex APMs, explanatory variables relating to road geometry data are included. These variables typically include both constructed or engineered features and natural aspects of the road system can also be included. The most commonly occurring variables in this category across the studies were:
    - Lane dimensions
    - Shoulder dimensions
    - Median dimensions
    - Curvature related variables
    - Gradient related variables
- The exploratory analysis to identify variables to be considered for inclusion in the model
  - One study used cumulative residual plots to decide on the most appropriate form for the variable (e.g. exponential, categorical etc.)
  - One study considered regional differences and similarities, grouping similar regions based on various conditions.

3.2 Statistical models

3.2.1 Types of models

The need for complex statistical approaches to model collision rates are discussed in Section 2.2. The most common statistical models used to develop APMs were generalised linear models (GLMs) with the outcome variable following either a Poisson or negative binomial distribution, due to the non-negative and random nature of the count data models. Older studies such as those carried out at TRL in the 1990s ((Summersgill & Layfield, Non-junction accidents on urban single-carriageway roads, 1996) and (Walmsley, Summersgill, & Payne, Accidents on modern rural dual-carriageway trunk roads. TRL report 335, 1998a)) use Poisson models whereas more recent studies tend to consider both Poisson and Negative Binomial models (Ambros & Sedonik, A Feasibility Study for Developing a Transferable Accident Prediction Model for Czech Regions, 2016), (La Torre, et al., 2016), (Garach, de Ona, Lopez, & Baena, 2016), (Cafiso, Di Graziano, Di Silvestro, La Cava, & Persaud, 2010).

Empirical Bayes estimates expected accident frequencies from existing models which can improve the fit of GLM models. Ambros et al. (2016) compared traditional reactive accident-based approaches (black spot identification using accident data only), state-of-the-art Empirical Bayes (EB) methods, and proactive preliminary road safety inspection (based on data collected by an instrumented vehicle) to identify hazardous road stretches. They recommended that the risk based and EB approaches were more valid than the blackspot approach, particularly on roads with low flows. Maher and Summersgill (1996) and Ambros and Sedonik (2016) mention Empirical Bayes modelling in the context of before and after studies because this deals rigorously with the Regression to the Mean[6] issue when collision numbers are low (see also (Hauer, 2007)).

Moving away from the more traditional GLMs, Geedipaly et al. (2012) highlighted that there could be a number of sites where no collisions were observed over a long period of time. This could result in the collision data containing a large number of zeros and a long-tailed distribution. In these cases, zero-inflated models can be used for both Poisson and negative binomial distributions. These models assume that zeros are generated in two states: one is a zero or safe state and the second is a non-zero state. However, Geedipaly et al. (2012) point out that this modelling technique has not been previously used to analyse a zero-inflated collision dataset and may come with some challenges such as the safe state having a long-term mean equal to zero. The US study used two collision datasets from Indiana and Michigan to model APMs with negative binomial and zero-inflated distributions; they also tested a negative binomial Lindley model. It is important to note that the reason for zero-inflated nature of the collision dataset is explained by the smaller segment lengths used in the study (majority being less than 0.3 miles).

Another study (Cafiso et al., 2010), used Generalised Estimating Equations (GEE) in addition to GLM models and compared the results between both modelling techniques. GEEs were mainly used to understand if changes over time (such as annual variation or trend in calibration of SPFs due to the influence of factors which may change over time) improved model results. The study found that GEEs improved the goodness of fit and accuracy of regression parameters. However, a drawback of this approach is that it required more attention if there are many missing values of the explanatory variables.

[6] Regression to the mean (RTM) is a statistical phenomenon which can occur when locations for implementation of road safety schemes are selected on the grounds of high numbers of collisions. Road collision counts are influenced by various causal factors as well as by random variation; this random element means that collision numbers fluctuate between higher and lower values, about an overall long-term average.

If a site is selected for a scheme due to a high number of collisions, this high collision rate may be part of this random fluctuation. It may be expected that the number of collisions will reduce (i.e. regress towards the longer term average) irrespective of whether or not any road safety interventions are implemented. This means that a simple study that compares the number of collisions before and after implementation of an intervention, without controlling for RTM, may overestimate the intervention effectiveness.

Things to consider for model development

- The type of model developed – this will depend on the data available
  - The most common model used to develop APMs were Generalised Linear Models (GLMs). Older studies typically assumed that the outcome variable followed Poisson distribution whereas more recent studies assumed a Negative Binomial distribution (which is more appropriate when the data are over dispersed and the mean and variance are not equal).
  - Another approach used in one study was Generalised Estimating Equations which accounted for time trends. The study found that GEEs provided better model fit compared to the traditional GLMs, but there are challenges if the explanatory data has missing values (i.e. the value of each variable cannot be identified accurately at each time point).
  - In cases where sections did not have any collisions over a long period of time, zero-inflated models (using either Poisson or Negative Binomial) were used. However, this assumes zeros are generated by two processes which may not be appropriate for collision data.
- An alternative approach uses information from existing models to improve the fit of new models
  - Although the search identified a few papers using Empirical Bayes before-and-after comparisons, the primary focus of these papers was around reliability of the EB method rather than development of APMs. Therefore, this review cannot draw any conclusions around using Empirical Bayes to estimate expected accident frequencies from existing models to improve the fit of GLM models.

3.2.2 Models for specific road types or collision types

Most studies focused on the development a model for a specific road type. Cafiso and D'Agostino (2012) and Pei et al (2016) focused on models for motorways. Cafiso et al. (2010) focused on two-lane rural highways. Some studies include additional restrictions, Garach et al. (2016) focused on two lane rural highways over flat terrain, whereas Turner et al. (2012) narrowed the focus even further by using rural highways with a speed limit of 100 km per hour, with no narrow bridges or railway crossings present. There were also studies that focused on lower volume, two-lane undivided roads (Ambros, Havranek, Valentova, Krivankova, & Streigler, 2016).

Some studies included separate models for different collision types (Summersgill, The availability of accident predictive models for inter-urban roads., 2000), and injury types (Ambros, Havranek, Valentova, Krivankova, & Streigler, 2016). The study by Taylor et al (2002) took an interesting approach by classifying road links into groups using Principal Component Analysis (PCA) and linear discriminant analysis. They then developed individual models for each group separately.

Things to consider for model development

- How many models are needed to cover the TII road network and the scope/coverage of these
    - Due to differences in the characteristics across the network and the effect of these on collision risk at each location, most studies developed models for specific road types.
    - Different junction types were often modelled separately from the main road links.
    - None of the studies in this review developed a single model for the entire network encompassing different main road types.
- Whether is it beneficial to model different collision types separately
    - Some studies modelled different collision types and/or injury types separately.

3.2.3 Procedure for building the APMs

The same model building procedure is followed in most studies (Summersgill & Layfield, Non-junction accidents on urban single-carriageway roads, 1996), (Walmsley, Summersgill, & Payne, Accidents on modern rural dual-carriageway trunk roads. TRL report 335, 1998a), (Vogt & Bared, 1998), (Turner, Singh, & Nates, 2012). First, a base model is created using the flow and segment length variables. The flow (AADT) and segment length variables are most commonly included in power form in the model equation:

$$Crashes = C * flow^{\alpha} * length^{\beta}$$

Ambros and Sedonik (2016) investigated the effect of using different functional forms, such as the power and exponential, for the AADT and segment length variables. They found that AADT in the exponential form was not statistically significant in any of several different model variants[7].

Next, additional variables are added to the model in exponential form (Garach, de Ona, Lopez, & Baena, 2016) (Taylor, Baruya, & Kennedy, 2002) (Vogt & Bared, 1998) using forward selection[8]. These usually include all the variables extracted from the various datasets and associated with the road segments that are the unit modelled.

There is a wide variety of criteria that have been used to select which variable is added to the model during a forward selection pass. Garach, et al., (2016), Vogt & Bared (1998) both used

[7] Where tested, flow is indicated to be the most significant determinant of collisions (Elvik, Høye, Vaa, & & Sørensen, 2009)

[8] This refers to the following method: 1. Start with a base model. Create new models that include base model variables and each of the possible additional variables. 2. Select the best of the new models according to some criterion. 3. The selected new model becomes the new base model. 4. These steps are repeated until the new models do not offer an improvement on the base model.

p-values for variable selection. Garach, et al. used p-value=0.05 as the criteria for significance and Vogt & Bared (1998) used p-value<0.05 as strongly significant and p-value<0.15 as moderately significant. Garach, et al. (2016) used several additional criteria these being the t-statistic significance for each parameter at 95% confidence level, engineering judgement, and low correlation with other independent variables to avoid the issue of multicollinearity[9].

Turner et al. (2012) developed 10 APMs using Poisson or Negative Binomial error structures for loss of control, head-on and driveway-related collisions on straight and curved rural roads. It is interesting to note that they developed two types of models. Firstly, one which was based on statistical expertise and used 'Akaike Information Criterion' (AIC) and Bayesian Information Criteria (BIC) for variable selection (model selected with 95% confidence levels). However, these models did not often identify variables that were of interest to road safety experts. Therefore, a second type of model was built (labelled as practitioner's models) which did not always achieve the 95% confidence level (for individual variable fit) but included variables of practical interest (the confidence level tended to be over 70%). The final results were compared across both models and below are some of the variables that were identified to be significant across most APMs:

- AADT
- SCRIM coefficient
- Region (a proxy variable to capture the effect of socio-economic and weather effects)
- Sealed width
- Gradient
- Roadside hazards (for instance, in the form of ratings)

In the study conducted by (Cafiso, Di Graziano, Di Silvestro, La Cava, & Persaud, 2010), two sets of modelling approaches were used: first, a simpler SPF was formulated (using AADT and segment length) and a second more complex model was developed using variables that relate to physical characteristics of the road segments. The AADT and segment length were included in power form in both models, in the complex model the additional variables were included in exponential form. Sets of non-correlated variables were identified using Pearson's correlation criterion to avoid any issues of multicollinearity. These sets were then used in the model development. Only variables within a single set were considered for inclusion in the model. The study applied two modelling techniques: Generalised Linear Models (GLMs) and GEE (Generalised Estimating Equations) to understand if incorporating time trends (such as annual variation in collision data or trend in calibration of SPFs due to the influence of other factors which could change over time) improved the model results. A total of 19 models were developed and the key variables that were identified to be significant were:

[9] Multicollinearity is an issue in statistically models where multiple explanatory variables are highly correlated to each other. This results in less reliable statistical inferences as the modelling will be unable to assign variance clearly to specific variables, all variables included in the model should therefore have low correlation (be independent of each other).

- AADT
- Driveway density
- Curve ratio
- Roadside hazard
- Speed differential density[10] in the homogenous segment.

The comparison of the two modelling techniques showed that including the time variable improved the goodness of fit of the GEE model compared to the more traditional GLM approach. It also improved the accuracy of the regression parameters and standard errors, thus improving the quality of the model. However, the GEE model required more attention when it came to missing values for the explanatory variables[11] and therefore required better quality of data compared to the traditional approach.

In the studies where several models were created, they had to be compared to select the most appropriate model. This was often done using one or more goodness-of-fit criteria. Cumulative residual plots were used in Cafiso *et al.* (2010) to check the performance of the model. Garach et al. (2016) reported these to be essential to developing a good model. The purpose of cumulative residual plots is to evaluate the variance of the system and the trend of the variation of variable residuals. This process identifies any abnormal deviations of the model used and evaluates how it fits to the dataset. If the cumulative residuals exceed the limits of +/-2 times the variance of the residuals, then the analysis suggests that the fit of the model is poor.

Some studies such as Summersgill and Layfield (1996) and Walmsley et al. (1998) used the scaled deviance criterion. Vogt and Bared (1998) used R-squared and its different variations such as weighted R-squared and Freeman-Tuckey R-squared. Turner et al (2012) used the Bayesian Information Criterion (BIC). Cafiso et al. (2010) used Pearson's Chi-squared values and Akaike's Information Criterion (AIC). Geedipaly et al. (2012) used the Deviance Information Criterion (DIC), Mean Absolute Deviation (MAD), and the Mean Squared Prediction Error (MSPE) to assess the model's predictive performance.

[10] Difference in the 85th percentile of speed between subsequent/contiguous elements of the homogenous section.

[11] If missing values exist then the estimates of the coefficients may be biased.

Things to consider for model development

Based on the findings from the literature review, below is a summary of the considerations when developing a procedure for model development.

Prior to model development:

- All explanatory variables checked for multicollinearity.
  - One approach would be to use Pearson's correlation criterion to identify independent variables which would be tested for inclusion in the main model.

During model development:

- Develop a base model using flow and segment length as these are known to be the key variables effecting collision risk.
- Explore various functional forms (power, exponential, categorical) for the other explanatory variables, depending on the distribution of the data.
- Add variables to the model using forward (or backward) selection techniques:
  - P-values could be used for variable selection. Some studies used p-value<0.05 as the criteria for significance, whereas other studies used higher p-values determined by the practitioner.
- Alternatively, AIC or BIC measures could be used to determine which variables to include. These are measures of goodness of fit and generally the lowest value implies better model fit.

Checking model fit:

- In cases where multiple models are being compared, various goodness-of-fit criteria should be explored to identify the best model explaining the variation in the response variable. For example, R-squared values and its variations, and scaled deviance criterion.
- Cumulative residual plots used to check the validity of each model.
  - In general, the residual standard deviation shows the largest improvement (decrease) when the important variables are added to the model and reduces as other variables are included. This could be used to identify the main variables affecting collision numbers.

Checking predictive performance of the model:

After the model has been built and fit to the data, the predictive performance of the model will need to be tested. This can be done by:

- Comparing the predicted values to the actual values.
- Calculating the Deviance Information Criterion (DIC), Mean Absolute Deviance (MAD) and/or Mean Squared Prediction Error (MSPE) measures.
- Although none of the studies identified discussed this, more recent approaches such as splitting the data and using a test set for model building and a training set for model testing should also be considered in the next phase.

3.3 Development of crash modification factors from APMs

After an SPF (or APM) has been developed, any modification to the predictions from the model must account for geometric design or traffic differences between the base conditions of the model and the conditions of the site being considered. This is determined using a Crash Modification Factor (CMF). In the simplest form, if an intervention is estimated to reduce injury collisions by 20% (or X%), then the CMF is 0.80 (1- X/100) (OECD, 2012). The study by OECD (2012) suggests that the CMF always refers to target collisions of a specific type and specified injury severity or severities. It also depends on various details and the circumstances under which it was estimated. For instance, the CMF for the radius of road curvature depends on multiple factors such as approach speed, angle between tangents and road type (urban or rural). The study also recommends testing multiple functional forms when developing CMFs.

Labi (2011) developed APMs using the negative binomial distribution for two-lane rural roads in the United States. The model found the following variables to have a significant effect on collision frequency:

- Road segment length
- Lane width
- Shoulder width
- Pavement surface friction
- Pavement condition
- Horizontal curvature
- Vertical grades

Once the model was established, crash reduction factors were estimated using the following equation for predicting nonlinear change in collisions due to changes in the independent variable:

$$CRF_{x_j} = (1 - e^{(\beta_j \Delta x_j)})$$

Where, CRF is the Crash Reduction Factor associated with the j-th independent variable, Δx is the change in magnitude of the variable (defining the intervention) under consideration and β is the estimated parameter for the j-th independent variable.

A value greater than 1 indicates an increase in collisions and a value below 1 indicates an expected reduction.

The standard error can be estimated using the square root of the variance, and the confidence interval can be estimated by using the formula below:

$$Confidence\ Interval = CMF \pm (Cumulative\ Probability\ \times Standard\ Error)$$

The cumulative probability for 95% confidence interval (a frequent standard) is 1.96.

Another study (Gross, Persaud, & Lyon, 2010) used the same mathematical formula for estimating CMFs. The study suggested developing CMFunctions rather than a singular CMFactor as safety effectiveness varies depending on a range of site characteristics. This is a recommendation supported by OECD (2012).

The development of a reliable CMF is costly and can often take multiple years (OECD, 2012). Therefore, the study looked into the transferability of CMFs between countries. In order to do so, the study looked at two interventions, effects of road lighting on injury accidents and speed enforcements on accidents. Next, the estimates of the effects of each intervention were gathered from a number of studies around the world and assessed based on consistency of results. A simple consistency score was developed based on the degree of overlap between the confidence intervals of the CMF of each study compared to the other. The higher the consistency score (closer to 1), the better the replication was across countries or years. The study found that the consistency score for road lighting (0.9) was much higher than speed enforcement (0.7). Therefore, the study (OECD, 2012) concluded that if there have been multiple studies looking at a particular intervention over decades and different countries and all these studies have shown highly consistent estimates of the intervention then it is reasonable to assume that the results would also apply to a different country. However, this assumption relies on a number of conditions being fulfilled and it is better to have locally derived and robust estimates. Firstly, all studies should apply the same approach and be of high methodological quality. Secondly, there should not be any trends over time in research findings which may suggest that CMFs are not transferable.

Things to consider for model development

- CMFs should be estimated for a specific collision type and injury severity.
- Multiple functional forms of the variable should be tested when developing CMFs.
- CMFunctions may be preferred over CMFactors as they account for a range of site characteristics.
- Developing a reliable CMF is costly and time-consuming. Therefore, studies have looked into the transferability of CMFs. While it is possible to apply CMFs developed for other countries, it depends on a number of factors and can only apply to specific interventions. Therefore, it is recommended to develop reliable estimates based on data available for the intended country.

Once the APMs have been developed, the next phase of this work should consider the interventions for which CMFs should be developed.

3.4 Summary

The aim of the literature review was to summarise the data and variables that are considered in the development of APMs, understand the modelling techniques used to develop APMs, and how the APMs can be converted to CMFs. This section presents a summary and discussion around the findings from the literature review.

Collision data and traffic flow information were the most common factors used to develop the basic APM (also known as Safety Performance Function). Older studies used manual approaches to collect information on traffic flows whereas, more recent studies were able to get this information from transportation databases. In general, the majority of the studies used about five years of collision data in order to limit any biases that may occur due to changes made to road sections over a longer period of time.

The choice of road segment length is important when developing these models. Some studies highlighted the need to have longer segments to avoid having zero-inflated data (where there are lots of segments with zero collisions during the time period of interest) in the dataset. Most studies avoided the issue of zero-inflation because they set a minimum length of acceptable road segments; and this varied from 50 metres up to 2 kilometres. Only one study did have the issue of zero-inflated collision numbers This was due to having a large number of very short road segment lengths (less than 500 metres), resulting in about 70% of the segments with zero collisions. This study compared the performance of negative binomial, zero inflated negative binomial and negative binomial Lindley models to model these data. It found that the negative binomial Lindley model performed best with highly zero inflated data.

In addition to criteria on the length, most studies developed road segments by grouping other variables included in the model in a homogenous manner. The variables used differed by study; some used a smaller subset of variables whereas other studies used all the variables available. Furthermore, some studies used additional criteria such as removing junctions or having a minimum amount of traffic flow on each segment.

In most of the studies, once the base models which included traffic and length were developed, additional variables on the physical road features were used to create more complex APMs. This data captured more detailed information on the dimensions and spatial characteristics of the roads. Whilst the range of variables tested for inclusion in the model in each study varied, depending on the road or junction type modelled, there was some overlap in the variables: this is summarised in Table 3. Common variables include those that have been used in two or more studies.

Table 3: Road geometry variables included in various studies (apart from AADT and segment length)

Study	Common variables	Additional variables
(Ambros & Sedonik, A Feasibility Study for Developing a Transferable Accident Prediction Model for Czech Regions, 2016)	Curvature change rate Density of intersections Road width Number of lanes Hard shoulder width Speed limit restrictions	Density of roadside facilities
(Cafiso, Di Graziano, Di Silvestro, La Cava, & Persaud, 2010)	Curvature change rate Hard shoulder and median width Lane width Number of lanes Paved width Speed differential density Average operating speed	
(Cafiso & D'Agostino, Safety Performance Function for	Roadside hazard rating Curvature of road	Slope of grade downhill Lack of cross slope

Study	Common variables	Additional variables
Motorways using Generalized Estimation Equations, 2012)		Embankment or trench
(Summersgill, The availability of accident predictive models for inter-urban roads., 2000)	Road width Number of major and minor junctions	Hardstrip factor Quality factor Hilliness coefficient
(Labi, 2011)	Lane width Shoulder width Pavement conditions Horizontal and vertical alignment of roads	
(Vogt & Bared, 1998)	Lane width Shoulder width Number of driveways or intersections Horizontal and vertical alignment of roads	Shoulder type Road lighting Terrain information Weather conditions
(Turner, Singh, & Nates, 2012)	Unsealed shoulder width Average absolute gradient Average curvature SCRIM coefficient Roadside hazard rating Horizontal consistency (percentage change in speed)	Seal width Combined point hazards Combined accesses Distance to non-traversable slope Regional groups

The modelling technique used was fairly standard across all studies. Generalised Linear Models (GLMs) with the outcome variable following a Poisson or negative binomial distribution was the approach used in almost all studies reviewed. However, it is crucial to note that no singular model was used for the entire road network and most studies focused on developing APMs for specific road types. One study applied Generalised Estimating Equations (GEEs) to better capture changes over time. Whilst this model fit the data better than a tradition GLM, there were drawbacks when it came to missing values and the approach required better quality data.

The variables that were identified to be statistically significant did not vary substantially between studies. These were:

- AADT
- Road segment length
- Curve ratio
- Roadside hazard
- Lane width
- Shoulder width

- SCRIM coefficient
- Horizontal and vertical conditions

After the APM has been developed, Crash Modification Factors (CMFs) can be calculated from the model output. A CMF always refers to a target category and injury severity and therefore, depends on details and factors affecting site characteristics. For example, a CMF for radius of road curvature depends on multiple factors such as approach speed, angle between tangents and road type (urban or rural). A CMF can be estimated using the formula

$$CRF_{x_j} = (1 - e^{(\beta_j \Delta x_j)})$$

Where, CRF is the crash reduction factor associated with the j-th independent variable, Δx is the change in magnitude of the variable (defining the intervention) under consideration and β is the estimated parameter for the j-th independent variable. A value greater than 1 indicates an increase in collisions and a value below 1 indicates an expected reduction.

It must be noted that calculating a reliable CMF is time consuming and costly process. While studies have looked into the transferability of CMFs, they are extremely difficult and depend on the intervention being applied. Therefore, it is recommended to develop a CMF specific to the country.

One key observation from the literature review is that none of the studies developed one model that could be used for a wide variety of road types, and thus could be applied to a whole network. Most studies focused on an individual road types, and some also narrowed down the types of collisions that were modelled. Of the studies that did cover a wide range of roads, most did this by generating individual models for each type of road. The glossary on the PRACT website[23] makes the following observation about APMs built using the regression approach: “The model cannot be safely applied to sites significantly different from the ones it was developed. For example, if a model was developed for four-lane motorways (two-lanes per direction), and lane number is not an input variable in the model, it cannot be safely used for six-lane motorways.”

3.5 Limitations of the literature review

During the literature search process, it was observed that the distinction between Accident Prediction Model (APM) and Safety Performance Function (SPF) has disappeared, and the terms are being used interchangeably. This meant that only using APM as a search term generated a large number of papers related to the simpler SPF models that were not relevant to the focus of this literature review.

The literature review identified a number of studies that applied Empirical Bayes before-and-after comparisons to develop APMs. However, the primary focus of these studies was around performance or reliability of the Empirical Bayes method rather than the process followed to develop the APM. Therefore, these studies were not included in this review.

Another type of study that muddled the search results were studies in which real time accident prediction models were developed. They included similar keywords to the studies that were of interest to this project, but the approach is very different. These studies were screened out at the abstract review stage.

Full review of the short-listed papers highlighted some other limitations and challenges: some studies focused on specific aspects of the modelling process. For example, Pei et al. (2016) did not provide any information on how road segments were determined, and variable selection techniques applied during the model building process. Instead, it focused on the bootstrap resampling approach used to deal with excessive zero crash counts, which isn't the focus on this review. Similarly, Ambros et al. (2016) focused on comparing three different black spot management models and did not go into details about model development. This meant that the remaining aspects of the modelling process were described in a brief manner and much of the information of interest was not available in the paper.

4 Task 2: Road data identification and assessment

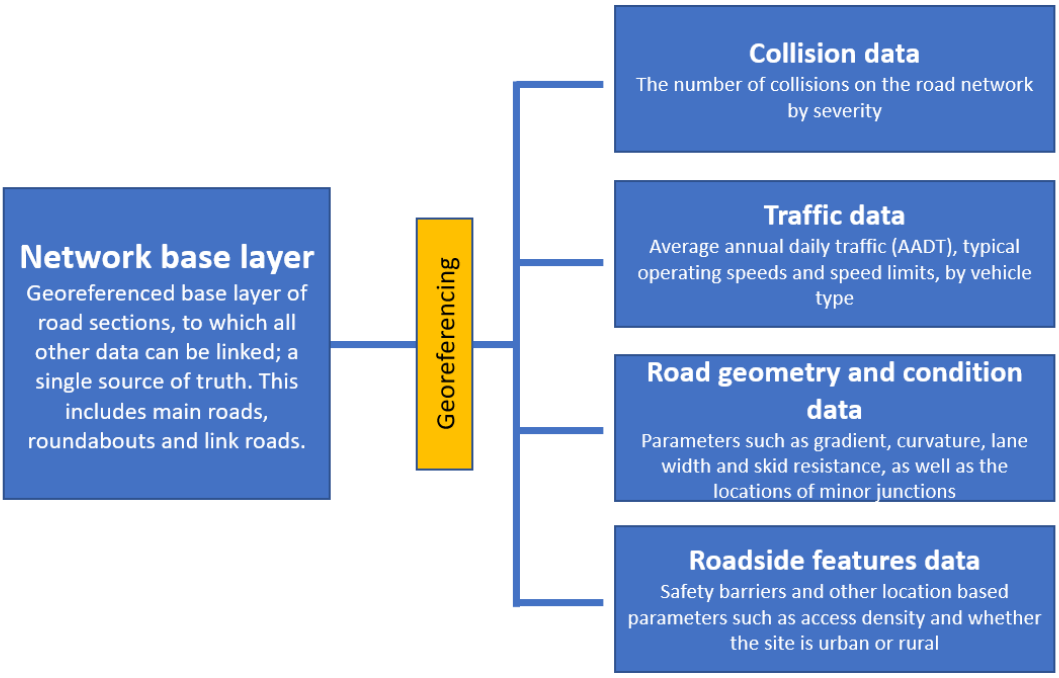
The availability of explanatory data sources affects the potential for different APMs pertaining to the TII road network to be developed successfully. Task 2 was concerned with the identification and full characterisation of the range of datasets for the TII road network that could potentially contribute to the statistical modelling exercise. The assessment of crash occurrence and patterns was particularly critical to this task.

Section 4.1 outlines the data sources available, including the nature and extent of the data, and the variables present. Section 4.2 discusses the data in more detail, analysing the collision data and identifying the explanatory variables suitable for inclusion in the modelling process. Section 4.3 describes how the data can be linked using GIS to build a central dataset for modelling. Finally, Section 4.4 outlines some feasible options for defining segments according to different metrics.

For the avoidance of confusion, when discussing sections of the road **for modelling,** these are referred to as **segments** (‘modelled segments’), as distinct from sections of the road defined by other means, for example as given in the raw data sources.

4.1 Data sources reviewed

The main data required for the development of the models is illustrated in Figure 1. Data were received and assessed relating to each of these categories. For use in the modelling, it is crucial that all data are georeferenced to a network base GIS layer which acts as a single source of truth.



Network base layer
Georeferenced base layer of road sections, to which all other data can be linked; a single source of truth. This includes main roads, roundabouts and link roads.

Collision data
The number of collisions on the road network by severity

Traffic data
Average annual daily traffic (AADT), typical operating speeds and speed limits, by vehicle type

Road geometry and condition data
Parameters such as gradient, curvature, lane width and skid resistance, as well as the locations of minor junctions

Roadside features data
Safety barriers and other location based parameters such as access density and whether the site is urban or rural

Figure 1: Data sources relevant for the modelling

Table 4 presents an overview of the different datasets assessed during Task 2, including the nature and extent of the data and the variables present which may be relevant for the modelling.

Table 4: Datasets reviewed during Task 2

Dataset	Description	Nature and extent	Key variables present
TII GIS base data	The entire national road network, primary and secondary; the linear referencing system used by TII.	A lines GIS layer of the national road network and points GIS layer of junctions on the network. There are a small number of roads not mapped in this data that are mapped in other datasets.	• Road category (main line, ramp, link road and roundabout) • Route ID, junction name and number
Ordnance Survey Ireland (OSi) PRIME2 data	The primary database (2021) for Osi spatial data containing an extensive amount of data on roads and buildings; acts as a referencing platform.	Data is within a buffer of the national road network. It is node to node and broken up at intersections. An extensive amount of detailed information is present on roads and structures within the buffer zone.	• Road class • Road and junction type • Locations of structures such as buildings and car parks
PMS lane width data	PMS collated paved width data, using the PRIME2 polygon data for motorways and dual carriageways	A lines GIS layer of the network with lane width for each 100m section. 'Chainage from' and 'chainage to' columns are included to give an ordering. There are a few small areas (approx. 1%) without data as lines are not present.	• Paved width for each 100m section • Urban yes/no field
PMS overall survey data	The latest (2021) PMS data from SCRIM, Road Surface Profiler (RSP) and Laser Crack Measurement System (LCMS) surveys relating to geometric parameters	Data across the national road network as a GIS point layer. Points are located every 10m and are referenced by road tag (name and number) and chainage for ordering.	• Radius (km), crossfall and gradient data (both in degrees) • SCRIM coefficient • Data on deterioration such as cracks in the road
PMS asset inventory data	Miscellaneous data from PMS related to the features of carriageways and hard shoulders	Data is a lines GIS layer. Network coverage is variable, for example there is only data on hard shoulder locations for most motorways and some dual carriageways.	• Location of hard shoulders (incomplete data) • Subnetwork values ('2' being urban, and '3' and '4' being legacy roads)
PMS Junctions data	Location of the approaches to minor junctions on the national road network	The approaches to junctions, typically 50m, as a lines layer – rather than point features of the junctions themselves. Coverage across the national road network.	• Junction type: crossroads, T-junctions, stop signs, and junctions with side roads
Traffic data	National transport model (NTpM) traffic data on peak speeds and flows across the national road network – for 2015 - 2019	Lines layer with one line representing undivided roads and one or two lines for divided rows (depending on year of data), with values presented for either direction. Data is across the entire national road network and a chainage field gives an ordering.	• Modelled AADT, AM peak and inter peak speeds for light and heavy vehicles in both directions • Number of lanes in both directions • Speed limit in km/h

Dataset	Description	Nature and extent	Key variables present
Speed limit data	Speed limit for roads on the national road network	Speed limit for major roads, linked to road names and georeferenced. There is data missing for approximately 3-5% of the network.	• Speed limit in km/h
Road type data	Data on road type and function for all roads on the national road network	Data for the whole network covering main roads, slips and roundabouts, split up by entire road length	• Carriageway type (e.g. 'Motorway', '3 Lane Dual') • Road name and route ID
Vehicle Restraint Systems (VRS) data	A combined MMaRC (2020) and local authority (2014) dataset on safety barriers	Data referenced by route ID and road. Data is quite extensive with a number of different descriptive fields with missing entries (nulls).	• Location, material and height of safety barriers
GeoDirectory data	Data from the definitive database of buildings, matched to a unique postal address – from 2021.	A point layer with each structure assigned geographical co-ordinates. Data is within a 1km buffer of the national roads.	• The location of, and general information on, structures such as schools, colleges, flats • An urban or rural field
2016 census data	Mapping of population densities and other demographic variables from the 2016 census	Data broken down into regions represented by polygons.	• Population (density) • Car ownership • Permanent dwellings
1km collision rates data	Data from 2014-18 on the number of collisions by severity on 1km sections of the road network, including an indicator of how that section compares with other sections with similar attributes	Lines GIS layer of the national road network broken into chunks of 1km, with chainage field for linking sections.	• Collision numbers split by severity (fatal, serious, minor and damage only) • Threshold (e.g. 'twice above average rate') • AADT • Speed limit
Raw collision data	Collision and vehicle level data on all collisions occurring on the national road network from 2014 to 2019	Csv files with one line per collision (collision level) and one line per vehicle (vehicle level). Vehicles are linked to collisions by a collision ID and collisions can be mapped using their latitude and longitude co-ordinates. Some collisions do not have co-ordinates.	At a collision level: • Location and time of collision • Collision severity • Other information such as road surface condition, lighting, junction and weather At a vehicle level: • Vehicle type • Vehicle action • Damage to vehicle

Additional data sources were also identified but subsequently not explored in detail. This was due to unreliability of the data, irrelevance to the modelling or difficulties in obtaining them. For example, the Road Safety Inspections dataset (which focuses on identifying hazards impacting the likelihood and consequences of collisions) was discussed with TII, but it was decided not to use this as historically this data has been qualitative with a lot of inconsistencies and subjective judgement of hazards. A more recent quantitative approach to gathering this data is not complete with surveying of the non-motorway network due to finish over the next few years. The potential to use climate data was also discussed due to the prevalence of rainfall in areas of the road network; however, suitable data was not available.

The following subsections discuss the reviewed datasets and their relevance to the modelling in more detail.

4.1.1 Network base Layer

The TII GIS base dataset is an almost completely comprehensive network base layer that other variables can be linked to, with **main roads, link roads, roundabouts** and **ramps** identified and mapped. Route IDs and junction names also act as reference points in this dataset. See Figure 2 for a screenshot illustrating the coverage of this dataset at a national level. There are very small sections missing from this dataset that are mapped in other datasets (such as the road type dataset and asset inventory dataset), hence combining datasets can provide a more comprehensive base layer. By joining the road type dataset, each road can be categorised as **dual or single** carriageway with a specified **number of lanes**. The subnetwork classifications in the PMS asset inventory data can also be used to identify road types, with classifications '3' and '4' being the **legacy** single carriageway roads (see Figure 9).

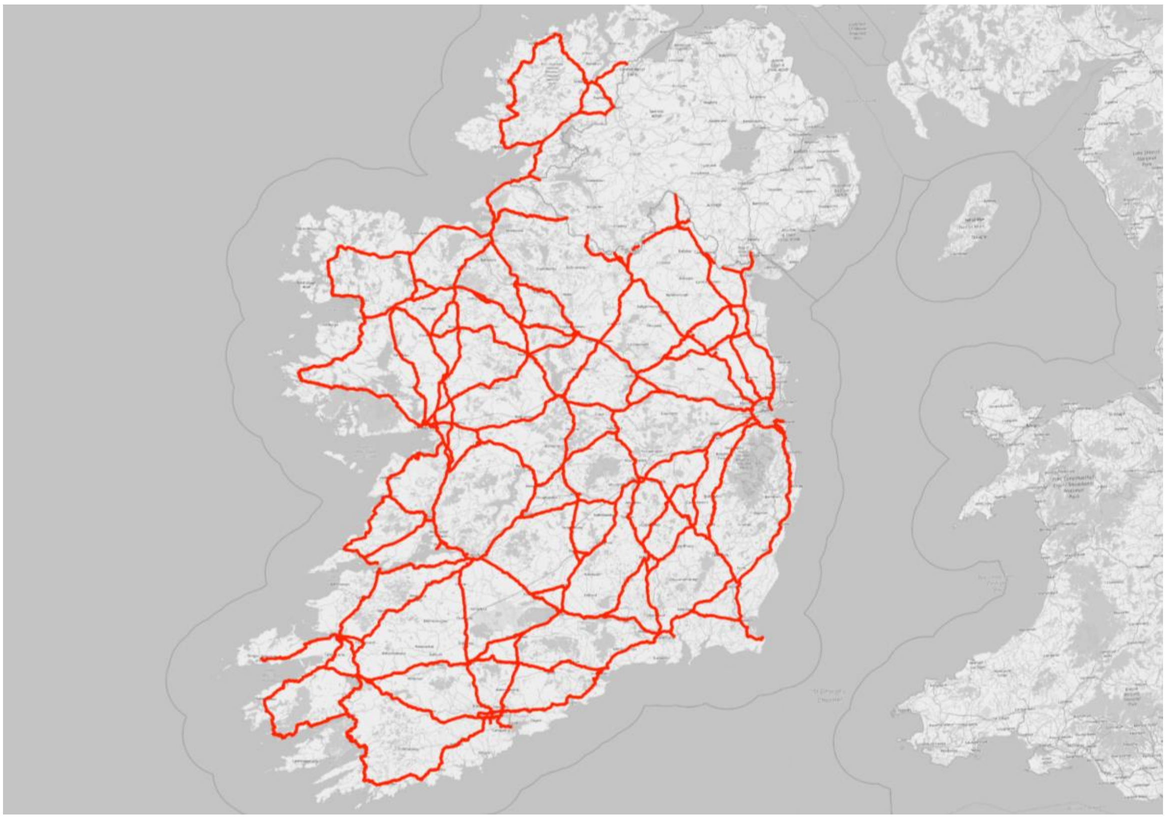
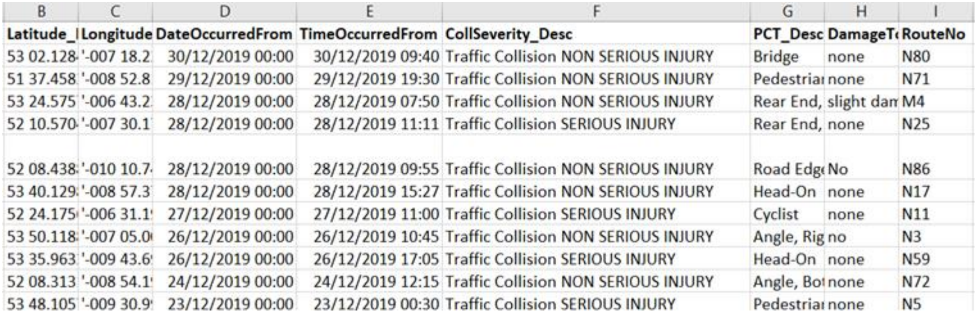


Figure 2: TII GIS base layer – network coverage

The PRIME2 data is another possible base layer, with road class, function and junction types present in the dataset. However, from analysis this data is more suitable as a backup base layer as the variables in the TII GIS data are more useful and well defined. As the PRIME2 data is given in a buffer around the national road network, separating out the national roads for modelling also creates an extra step to processing this data into a base layer.

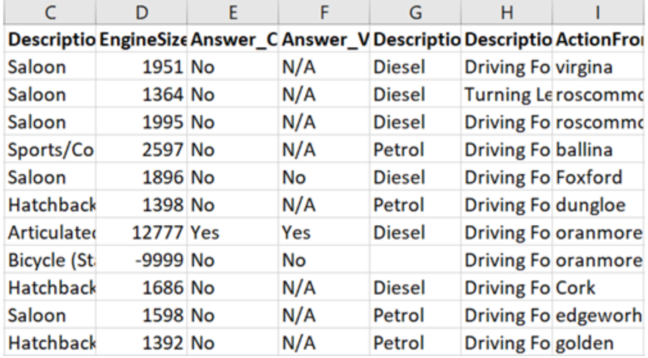
4.1.2 Collision data

There are two sources of data on the number of collisions on the national road network. The raw collision data (as csv files) contains data on every collision on the national road network (collision level) from 2014 to 2019 and all the vehicles involved in each of these collisions (vehicle level). Vehicles can be linked to the collisions they were involved in by a unique ID. See Figure 3 for a screenshot of this data at the collision level and Figure 4 for a screenshot at the vehicle level, with one row per collision and one row per vehicle respectively.



B	C	D	E	F	G	H	I
Latitude_I	Longitude	DateOccurredFrom	TimeOccurredFrom	CollSeverity_Desc	PCT_Desc	DamageT	RouteNo
53 02.128	'-007 18.2	30/12/2019 00:00	30/12/2019 09:40	Traffic Collision NON SERIOUS INJURY	Bridge	none	N80
51 37.458	'-008 52.8	29/12/2019 00:00	29/12/2019 19:30	Traffic Collision NON SERIOUS INJURY	Pedestriar	none	N71
53 24.575	'-006 43.2	28/12/2019 00:00	28/12/2019 07:50	Traffic Collision NON SERIOUS INJURY	Rear End,	slight dan	M4
52 10.570	'-007 30.1	28/12/2019 00:00	28/12/2019 11:11	Traffic Collision SERIOUS INJURY	Rear End,	none	N25
52 08.438	'-010 10.7	28/12/2019 00:00	28/12/2019 09:55	Traffic Collision NON SERIOUS INJURY	Road Edg	No	N86
53 40.129	'-008 57.3	28/12/2019 00:00	28/12/2019 15:27	Traffic Collision NON SERIOUS INJURY	Head-On	none	N17
52 24.175	'-006 31.1	27/12/2019 00:00	27/12/2019 11:00	Traffic Collision SERIOUS INJURY	Cyclist	none	N11
53 50.118	'-007 05.0	26/12/2019 00:00	26/12/2019 10:45	Traffic Collision NON SERIOUS INJURY	Angle, Rig	no	N3
53 35.963	'-009 43.6	26/12/2019 00:00	26/12/2019 17:05	Traffic Collision SERIOUS INJURY	Head-On	none	N59
52 08.313	'-008 54.1	24/12/2019 00:00	24/12/2019 12:15	Traffic Collision NON SERIOUS INJURY	Angle, Bo	none	N72
53 48.105	'-009 30.9	23/12/2019 00:00	23/12/2019 00:30	Traffic Collision SERIOUS INJURY	Pedestriar	none	N5

Figure 3: Raw collisions data – collision level (more columns not shown)



C	D	E	F	G	H	I
Descriptio	EngineSize	Answer_C	Answer_V	Descriptio	Descriptio	ActionFro
Saloon	1951	No	N/A	Diesel	Driving Fo	virgina
Saloon	1364	No	N/A	Diesel	Turning Le	roscommc
Saloon	1995	No	N/A	Diesel	Driving Fo	roscommc
Sports/Co	2597	No	N/A	Petrol	Driving Fo	ballina
Saloon	1896	No	No	Diesel	Driving Fo	Foxford
Hatchback	1398	No	N/A	Petrol	Driving Fo	dungloe
Articulate	12777	Yes	Yes	Diesel	Driving Fo	oranmore
Bicycle (St	-9999	No	No		Driving Fo	oranmore
Hatchback	1686	No	N/A	Diesel	Driving Fo	Cork
Saloon	1598	No	N/A	Petrol	Driving Fo	edgeworh
Hatchback	1392	No	N/A	Petrol	Driving Fo	golden

Figure 4: Raw collisions data - vehicle level (more columns not shown)

There are various fields in the data which are critical to the modelling process, such as the **latitude** and **longitude** co-ordinates for linking with a network base layer and the **collision severity**. Nearly 7,000 collisions (13%) do not have suitable co-ordinates, with this field being either blank or containing a default value, and hence cannot be linked geographically to a base layer. The route number can link these collisions to a particular road; however, in most cases this will not give precise location information. Other fields, for example the collision type field, provide a greater understanding of the nature of the collisions and will therefore be informative to the modelling process. Fields such as 'speed limit' and 'junction' may be useful for cross-checking or filling in gaps in other data sources.

The second data source on collisions is the 1km collision rates data, currently available from 2014-18. This dataset splits up the network into (up to) 1km sections, presenting the number of collisions on each section by severity. One of the purposes of this data is to identify locations with a high number of collisions, hence there is a variable indicating how the section compares with other sections of similar features (with values such as 'twice above average rate'). Other information including speed limit and AADT is also present at a section level. It is important to note that each section combines collisions from both sides of the carriageway, so it is not possible from this dataset to assign collisions to one side of the carriageway. Figure 5 illustrates the distribution of injury collisions in the 1km sections across the network. There are 1,613 sections with zero collisions, which makes up 29% of the sections on the national road network.

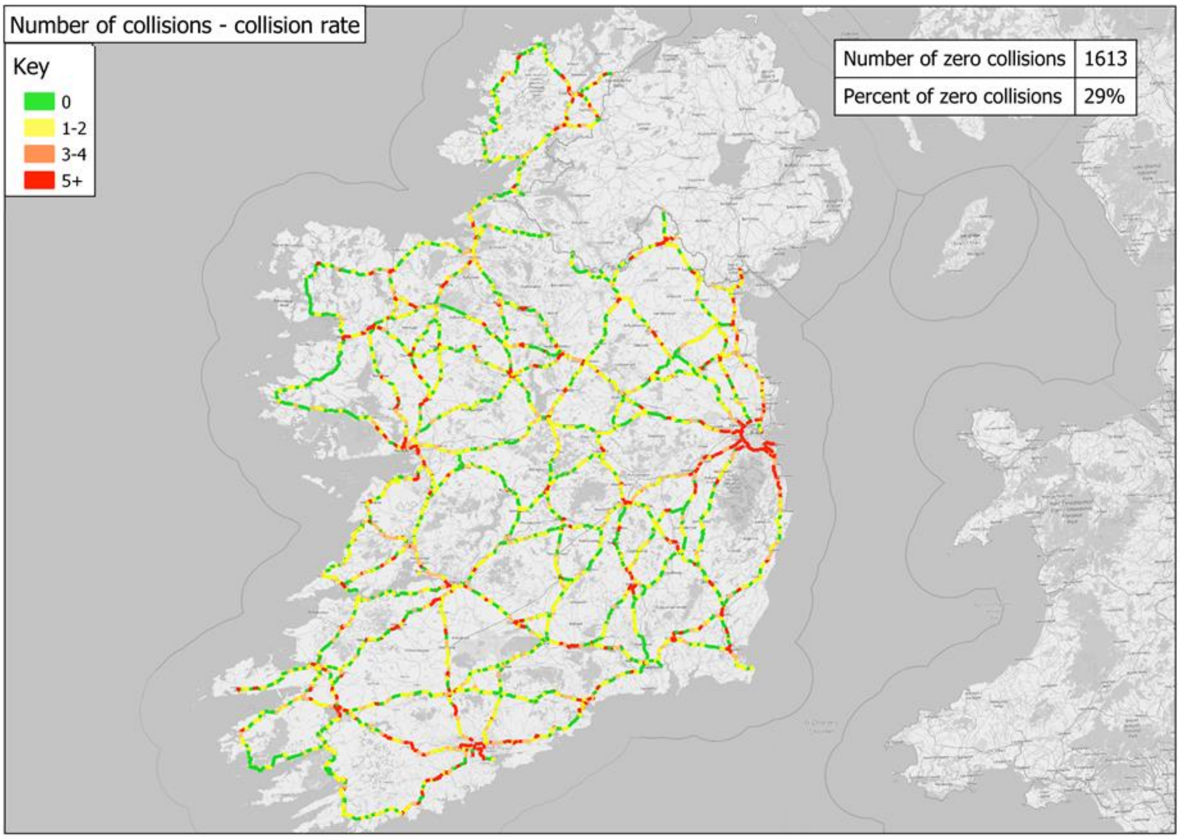

Number of collisions - collision rate
Key
🟩 0
🟨 1-2
🟧 3-4
🟥 5+
Number of zero collisions 1613
Percent of zero collisions 29%

Figure 5: Distribution of injury collisions in the 1km collision rates dataset

The main difference between these two collision datasets is how the collisions are mapped. The co-ordinates in the raw data give the location of every individual collision, whereas in the 1km collisions data every collision is only known to be within a 1km section of road. Also, 2019 data is currently only available in the raw format (TII could provide the 1km mapping for 2019 if required). The collision data is discussed in more detail, emphasising implications to the modelling process, in Section 4.2.1.

4.1.3 Traffic data

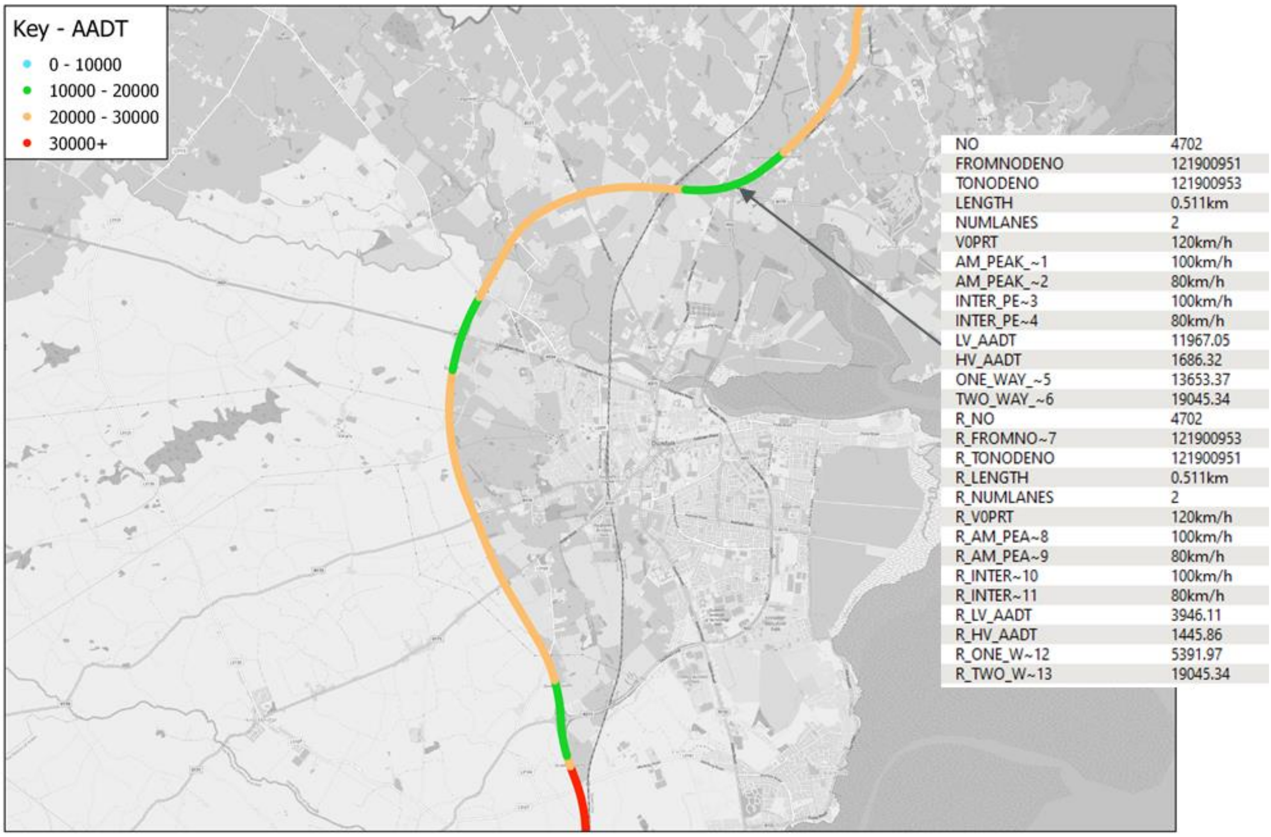
The traffic data from the National Transport Model has the following important variables, all given for both directions:

- **AM and inter peak speeds** – modelled for light vehicles (typically motorbikes, passenger cars and vans) and heavy vehicles (buses, caravans and HGVs) separately. Speeds are determined by the traffic flow, speed limit and link type; each link has a capacity, maximum allowable speed and volume-delay function. ‘Peak’ speeds are the modelled actual speeds for the given time period; AM speed being modelled between 7am and 10am and ‘inter’ generally meaning 12pm to 2pm.
- **AADT** - estimated from AM peak and inter peak traffic flows using expansion factors, for light vehicles and heavy vehicles separately. Values are calibrated with Traffic monitoring units regularly, using the 300 traffic counters from across the country.
- **Speed limit**

The modelling is done at a strategic level and data is aggregated into relatively long sections hence there are not big changes in traffic flows and speeds at junctions. Section lengths vary substantially from 200m to 20,000m and are dependent on where the changes in flow or speed occur. The traffic data received during task 2 spans 2015 to 2019; however, the 2014 data is also available to align with the time period of the raw collision data, if required for the modelling.

Figure 6 shows a screenshot of the 2019 traffic data on the road network around Dundalk, colour coded according to two-way AADT. The fields in the data are also illustrated for a section of this road; those beginning with an ‘R’ give the values for the second direction of travel.

Data for the speed limit could be taken from the traffic data, the separate speed limit dataset, the raw collisions data (on sections where there are collisions) or a combination of these datasets as required.



NO	4702
FROMNODENO	121900951
TONODENO	121900953
LENGTH	0.511km
NUMLANES	2
V0PRT	120km/h
AM_PEAK_~1	100km/h
AM_PEAK_~2	80km/h
INTER_PE~3	100km/h
INTER_PE~4	80km/h
LV_AADT	11967.05
HV_AADT	1686.32
ONE_WAY_~5	13653.37
TWO_WAY_~6	19045.34
R_NO	4702
R_FROMNO~7	121900953
R_TONODENO	121900951
R_LENGTH	0.511km
R_NUMLANES	2
R_V0PRT	120km/h
R_AM_PEA~8	100km/h
R_AM_PEA~9	80km/h
R_INTER~10	100km/h
R_INTER~11	80km/h
R_LV_AADT	3946.11
R_HV_AADT	1445.86
R_ONE_W~12	5391.97
R_TWO_W~13	19045.34

Figure 6: Traffic data for M1 around Dundalk, grouped according to two-way AADT

4.1.4 Road geometry and condition data

Most of the important road geometry and condition data is available in the PMS overall survey dataset. **Radius, crossfall, gradient** and **SCRIM** coefficient presented at 10m intervals gives flexibility with data linking and segment creation (see Section 4.3). Radius data can easily be used to estimate curvature; Figure 7 illustrates the range of radius values on the M1 road network around Dundalk, with values smoothed according to a 100-point rolling average. The dataset also includes information on deterioration such as cracks in the road, which are likely to be more temporally varying and hence less useful for the modelling. It is also difficult to take localised deterioration data to make a general statement about an entire section.

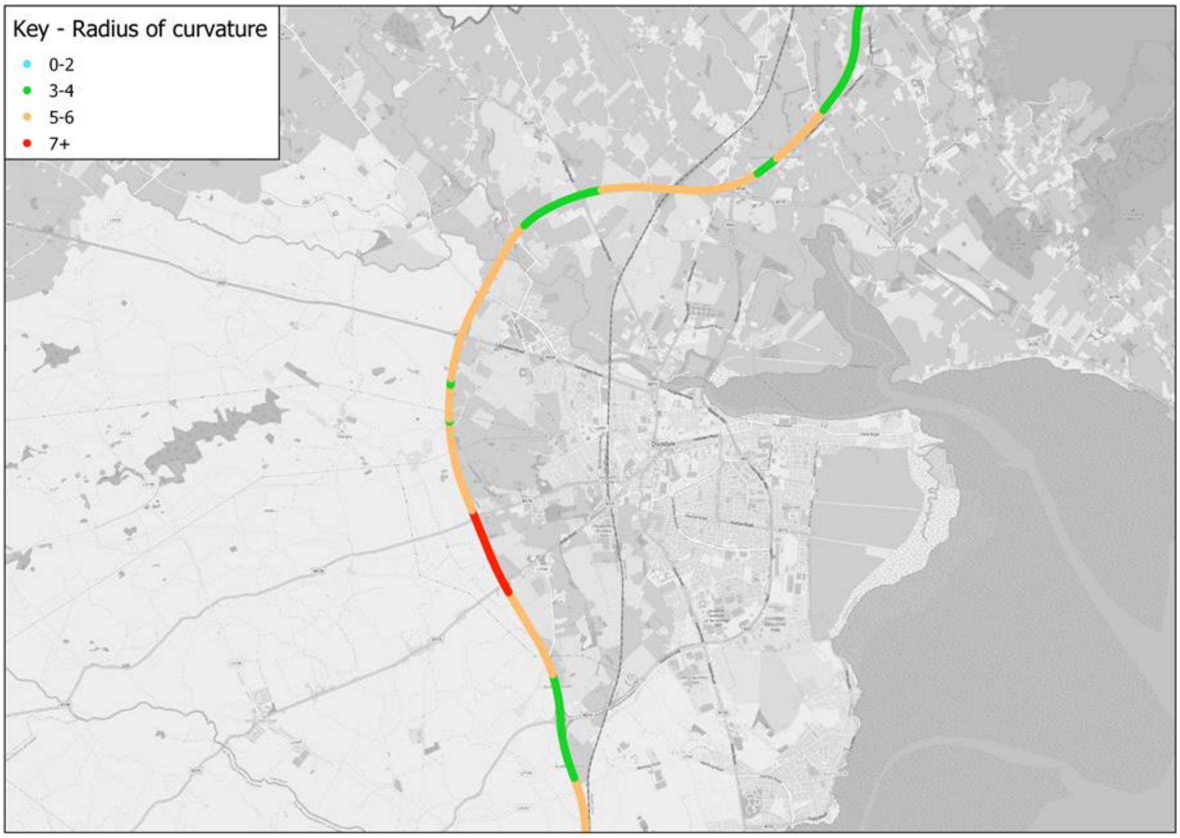


Figure 7: Grouped radius (km) values on the M1 around Dundalk, smoothed with a rolling average

Lane width across the network was also provided by PMS and the small sections in this data with missing values can potentially be estimated using information on the number of lanes and the fact that each lane is typically around 3.5m. PMS junctions data gives the location of all **minor junctions** such as crossroads and T-junctions.

4.1.5 Roadside features data

The PMS asset inventory dataset has data on the location of **hard shoulders** for most motorway sections and some dual carriageways. The rest of the network is being surveyed this year. There are fields in the data giving more information on hard shoulders such as surface and width, though much of this is incomplete and not very detailed. There is no information on presence of hard shoulders or hard strips for other road types.

The GeoDirectory dataset can potentially act as a proxy for **access density** on the road network, as structures are all geolocated. For example, the number of structures can be counted within a certain buffer of the network, either at the 1km level as given in the data, or smaller. Information on building use (such as 'residential') and type (such as 'terrace', 'townhouses') is also available. See Figure 8 for an illustration of the buildings within 1km of the network around Dundalk, categorised by use.

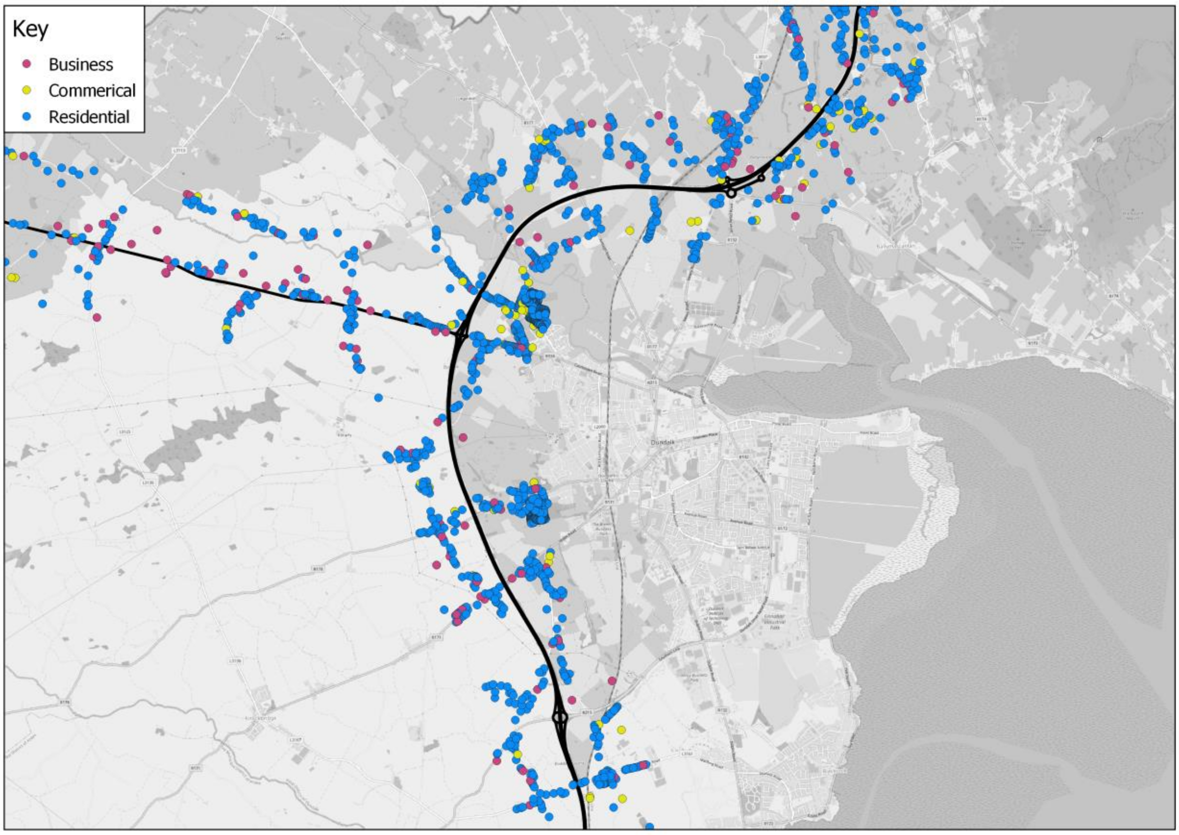


Figure 8: Buildings mapped in the GeoDirectory data within 1km of the network

The 2016 census data on the number of households with cars also has information regarding access density. The following information is contained in this dataset at a regional level (each region represented by a polygon in GIS).

- Population
- Permanent dwellings and the number of these that are occupied
- Number of permanent dwellings with different numbers of cars (‘No car’ up to ‘4 or more’ cars).

However, each region is quite large so the usefulness of this data in assigning variation to segments is limited.

Further, the PRIME2 dataset geolocates buildings with attributes such as form (such as ‘detached building’, ‘hotel’) and function (such as ‘residence’, ‘commercial/retail’).

Classification of locations into **urban or rural** is available in multiple different datasets. There is an urban ‘yes’ or ‘no’ field in the PMS lane width data and the GeoDirectory data. There are also subnetwork groups in the lane width data and the asset inventory data, where subnetwork classification ‘2’ refers to an urban site. The subnetwork distribution in the asset inventory data is illustrated in Figure 9, according to the following groups:

- Subnet 0 – Motorway and Dual Carriageway Network – c.1200km

- Subnet 1 – Engineered Single Carriageway – c.1200km
- Subnet 2 – Urban Areas – c.700km
- Subnet 3 – Legacy Pavement High Traffic – c.1250km
- Subnet 4 – Legacy Pavement Low Traffic – c.1000km

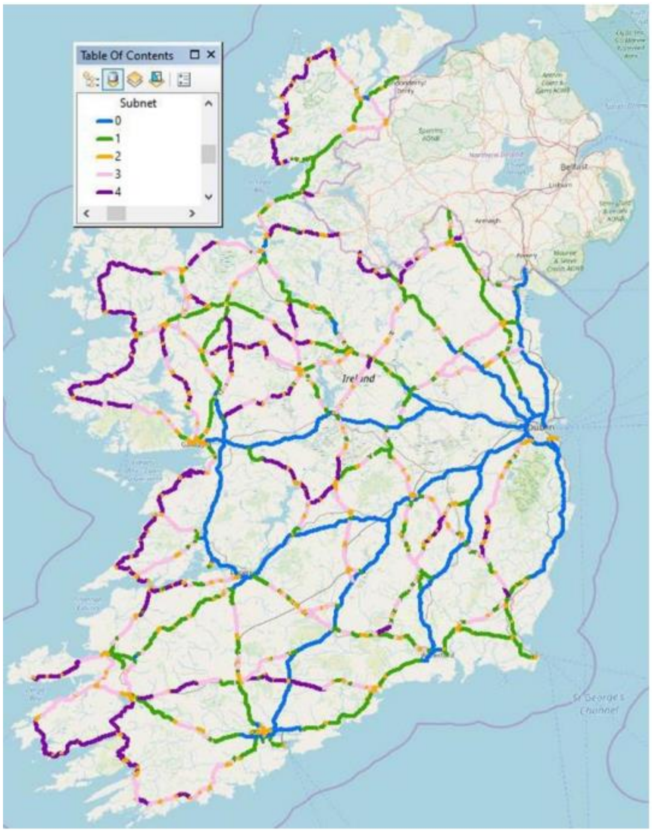


Figure 9: Subnetwork groups in the PMS asset inventory data

Finally, the VRS data locates the **safety barriers** on the network with GIS lines representing each barrier. Data on the features of these barriers (for example height and material) is also present, though the fields representing this extra information would need consolidating for use in the modelling. The 2014 data provided by local authorities (who manage the majority of the TII road network[12]) is less up-to-date than the 2020 MMaRC provided data. An updated local authority dataset will not be available until later this year.

[12]The TII road network is approximately 5,300km in length, of which c.4000km is managed by local authorities, c.750km operated by Motorways Maintenance and Renewals Contract (MMaRC) contractors and c.400km operated by PPP operators.

4.2 Data for modelling

This section describes the collision data in more detail, focusing on the features of the data that have implications for the modelling process. The nature of this data, particularly the distribution of collisions by severity and type, is critical to determining the modelling approach. The variables suitable for inclusion in the modelling are also identified and discussed.

4.2.1 Collision data

As outlined in Section 4.1.2 there are two sources of collision data – the raw collision data and the 1km aggregated collision rates data. From analysis, the raw collision dataset is more suitable for use in the modelling for several reasons:

1. The raw data can be utilised more flexibly using the location of individual collisions. For any approach to defining and generating the segments for modelling (e.g. segments homogeneous in length or traffic flow – see Section 4.4), collisions can more easily be assigned to these segments.
2. As a suitable network base layer exists (to which the other data sources can be linked), the collision rates data is not required as a base layer. This was an initial consideration due to the coverage of this dataset.
3. The raw collisions data has extra detail that is not present in the rates data, for example the collision type, and presents information at vehicle level.
4. At the time of writing the 2019 collision rates data is not yet available.

One of the disadvantages of the raw collisions data is that some of the reported collisions cannot be linked to the base layer by their co-ordinates (see Section 4.2.1.1). However, assuming there is no bias in the actual location or nature of these collisions (so that, for example, a relatively high proportion of collisions are not missing from a particular section), the remainder of the dataset should be suitable for the modelling exercise. The number of these collisions actually located on the national road network is also unclear. Therefore, the proposed dataset for use in the modelling is the raw collision data, which is the collision data described in the remainder of this report.

4.2.1.1 Collision Assignment

With a 10m buffer zone to the base layer, just under 80% of the collisions can be linked to the network (41,546 of 53,873). There are nearly 7,000 collisions with missing co-ordinates which cannot be linked with a buffer of any size (see Section 4.1.2). These can be linked to a particular route by the route number field, but not to a particular location on that route. The remaining collisions not linked with a 10m buffer have co-ordinates further from the base layer, and many are off the national road network, as shown in Figure 10. Therefore, consideration needs to be given to a linking method that ensures that the maximum number of collisions can be used accurately in the modelling, without modelling collisions not on the national road network.

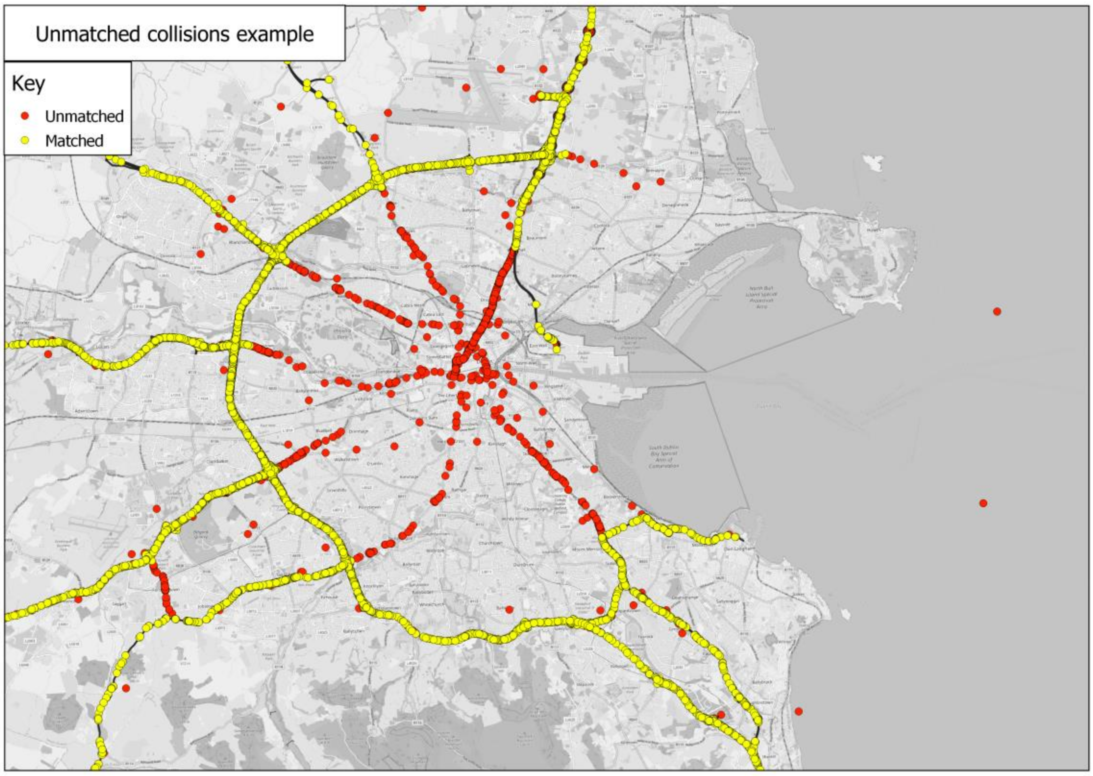


Figure 10: Collisions matched and unmatched with a 10m buffer to the network base layer around Dublin

4.2.1.2 Collisions by severity and road type

The number of collisions on the network between 2014 and 2019 **by severity** is shown in Table 5.

Table 5: Collisions by severity, 2014-19

Collision severity	Number of collisions	Approximate average number of collisions per km (Number/5300)
Fatal	349	0.1
Serious injury	1,114	0.2
Non-serious injury	6,178	1.2
Material damage only	46,232	8.7
All severities	**53,873**	**10.2**

There were 7,641 injury collisions on the network, equating to approximately 1.5 collisions per km over six years. This low collision density can be problematic for modelling (see Section 3.1.3). Including damage only collisions increases the number of collisions to 53,873 and the average number of collisions per km to 10.2.

The distribution of the 6,758 injury collisions that could be matched according to the 10m buffer is shown in Figure 11, with the network split into sections of at most 1km in length. The increase in density when damage only collisions are included is illustrated in Figure 12. As shown, including damage only collisions in the modelling would substantially reduce the number of segments with zero collisions, particularly in rural areas where collision numbers per km are on average much lower. For injury collision types, 53% of the 12,785 sections have zero collisions over six years. When damage only collisions are included this is reduced to 15% of sections.

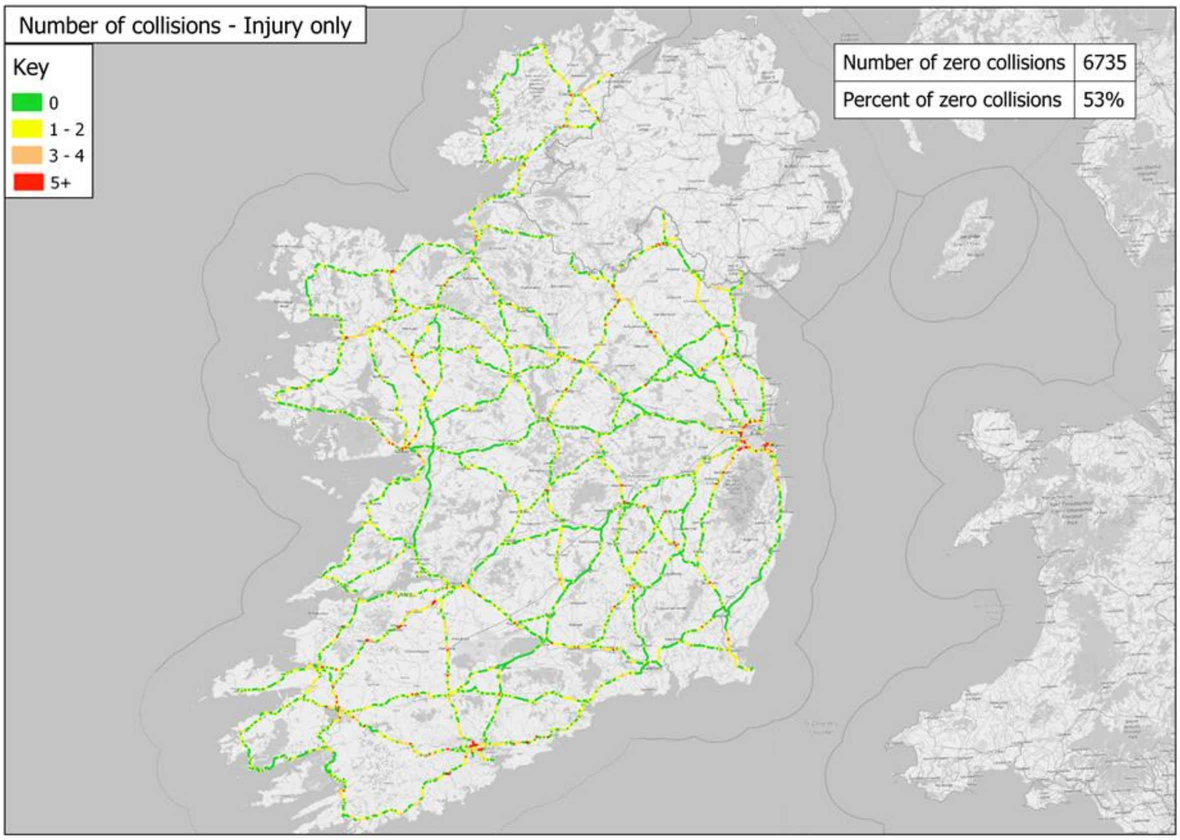

Number of collisions - Injury only
Key
🟩 0
🟨 1 - 2
🟧 3 - 4
🟥 5+
Number of zero collisions 6735
Percent of zero collisions 53%

Figure 11: Distribution of <u>injury only</u> collisions; network processed into sections of at most 1km in length (N=6,758)

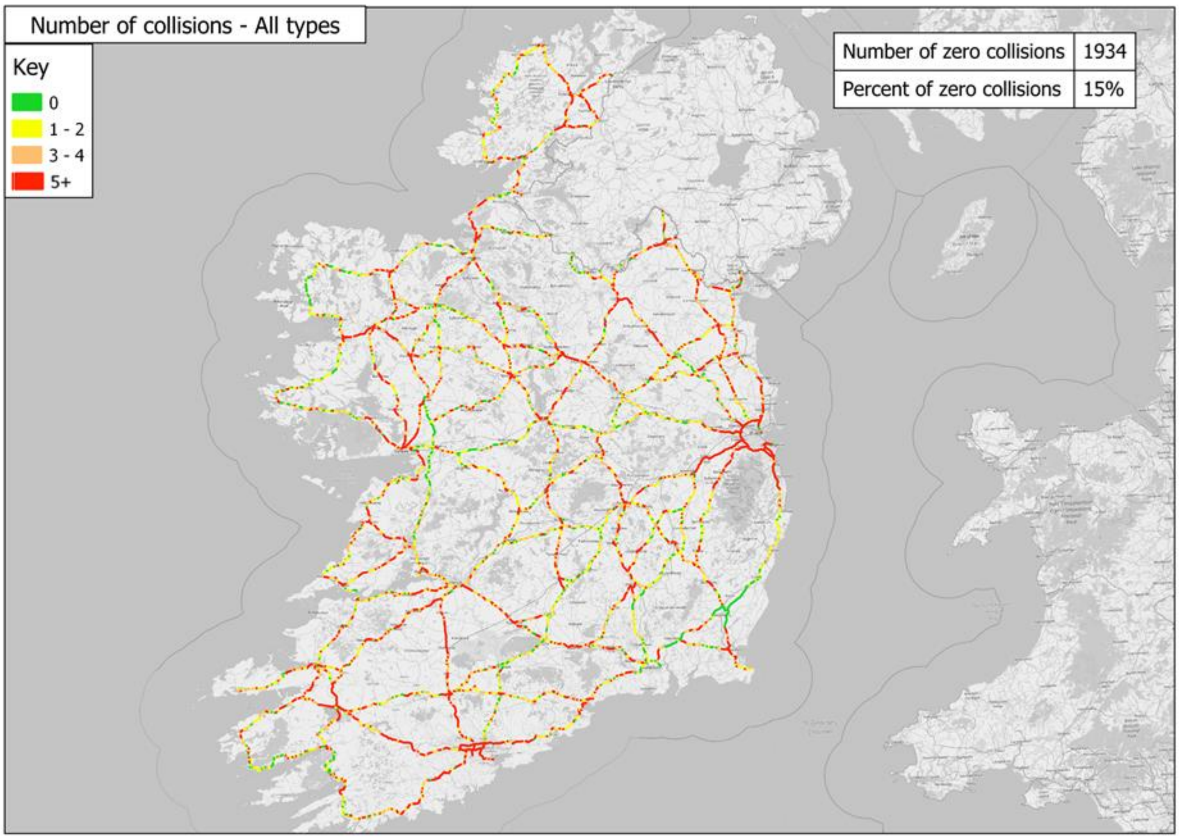

Number of collisions - All types
Key
🟩 0
🟨 1 - 2
🟧 3 - 4
🟥 5+
Number of zero collisions 1934
Percent of zero collisions 15%

Figure 12: Distribution of all collisions; network processed into sections of at most 1km in length (N=41,546)

Figure 13 to Figure 15 show the distribution of the matched collisions (including damage only) according to other section lengths: 100m, 500m and 2km.

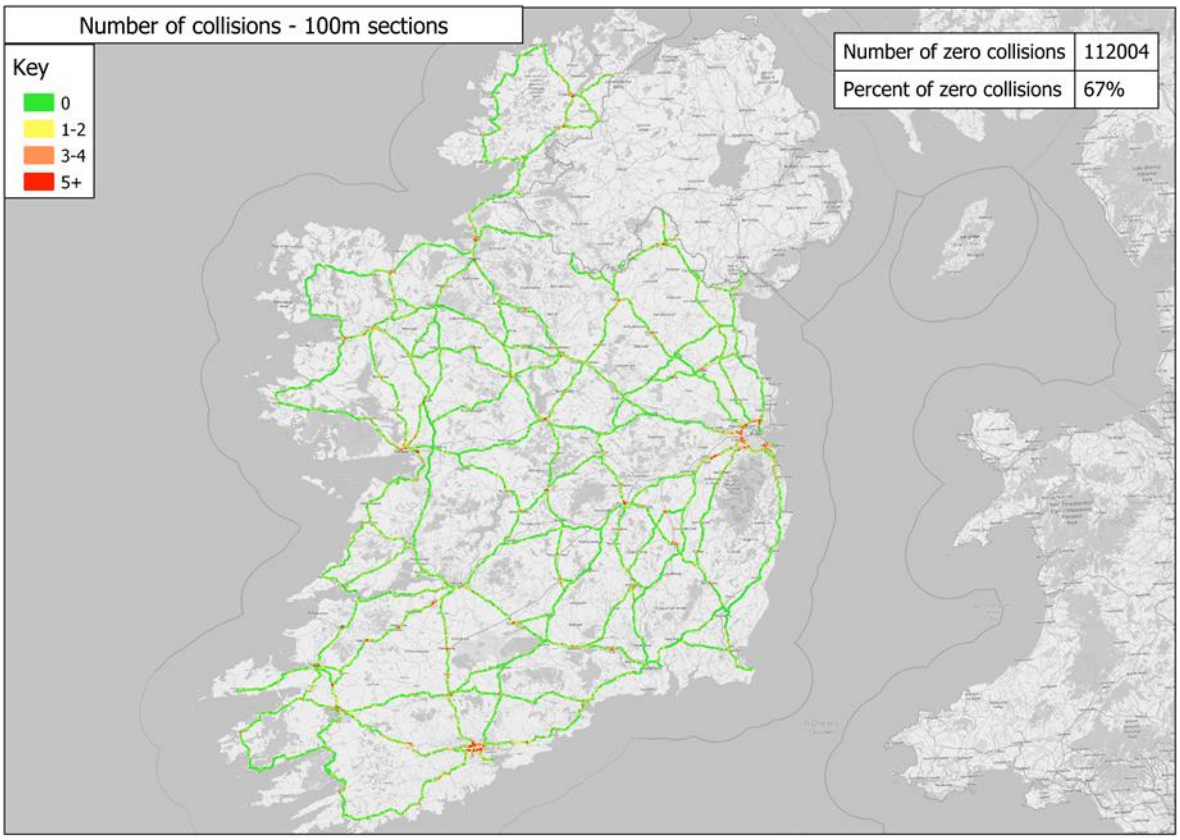

Number of collisions - 100m sections
Key
0
1-2
3-4
5+
Number of zero collisions
112004
Percent of zero collisions
67%

Figure 13: Distribution of collisions on 100m sections

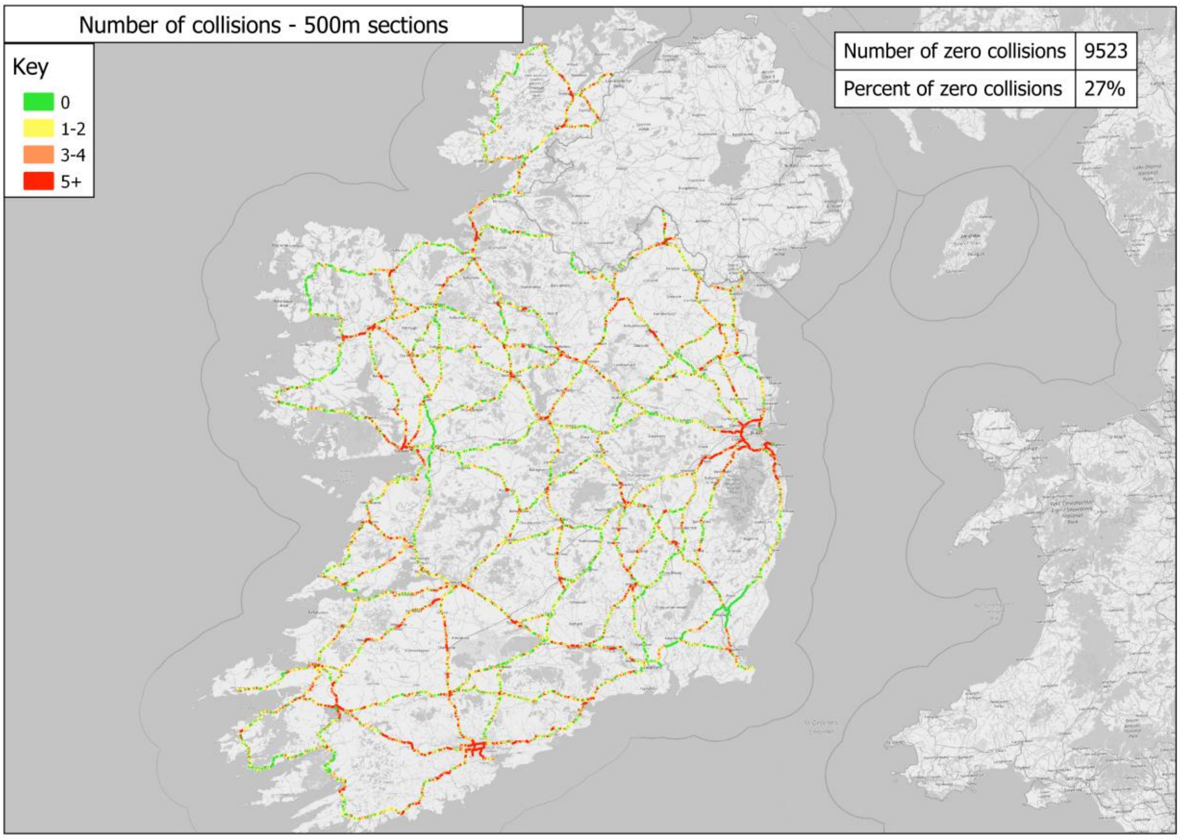

Number of collisions - 500m sections
Key
🟩 0
🟨 1-2
🟧 3-4
🟥 5+
Number of zero collisions 9523
Percent of zero collisions 27%

Figure 14: Distribution of collisions on 500m sections

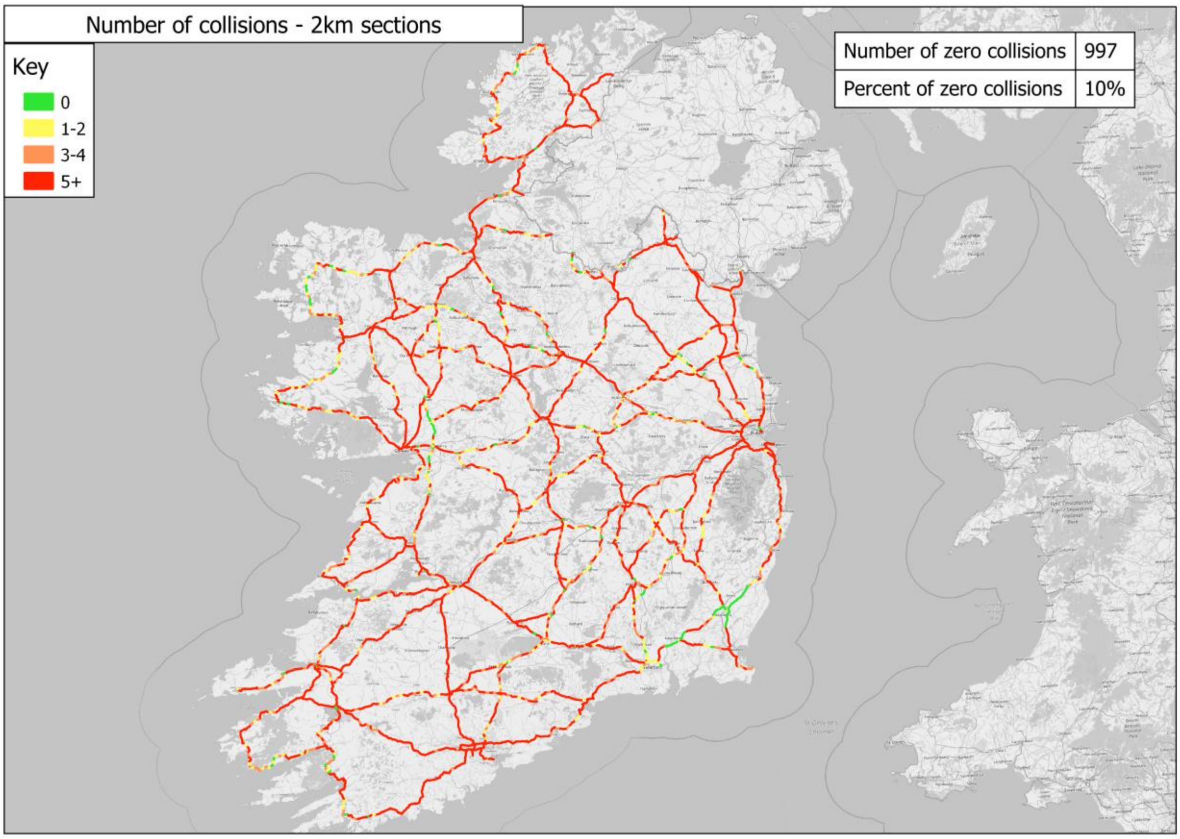

Number of collisions - 2km sections
Key
🟩 0
🟨 1-2
🟧 3-4
🟥 5+
Number of zero collisions | 997
Percent of zero collisions | 10%

Figure 15: Distribution of collisions on 2km sections

With section lengths of at most 100m, 67% of sections have zero collisions. For sections at most 500m this figure reduces to 27% and for sections at most 2km this figure reduces further to 10%. Consideration should be given to whether defining a minimum segment length is appropriate in the modelling, to avoid problems with zero inflation.

The percentage of the matched collisions on different **road types** is given in Table 6, according to the grouping in the TII GIS base layer, the road type dataset and the subnetwork classifications. The length of each road is given according to the length of the centreline, so that the two directions of dual carriageways and motorways are counted together. Legacy roads are comprised mostly of single carriageway sections.

Table 6: Proportion of network and collisions by road type; N (all) = 41,456, N (injury) = 6,758

Road type	% of network base layer by length	Number of collisions per km	% of all collisions on the network	% of injury collisions on the network	% of collisions on this road type that are injury	Average two-way AADT	Collisions per year per 10^8 veh-km
Mainline motorway	17.2%	6.0	14.6%	13.1%	14.5%	23,684	11.6
Mainline dual carriageway (non-motorway)	5.7%	17.1	13.7%	10.4%	12.3%	26,907	29.0
Mainline – single carriageway (non-legacy)	33.8%	8.4	40.2%	46.5%	18.7%	9,232	41.7
Legacy roads (subnet 3 and 4)	33.7%	4.3	20.5%	24.2%	19.1%	3,905	50.5
Link road	2.5%	22.0	7.6%	3.5%	7.4%	10,304	97.4
Roundabout	1.3%	2.6	0.5%	0.4%	13.2%	23,556	5.0
Ramp	5.8%	3.4	2.8%	1.8%	10.5%	4,550	34.5

Just over 40% of collisions on the network are on non-legacy single carriageway roads, and nearly half of all injury collisions, highlighting that these roads have a higher proportion of collisions that resulted in an injury. Around a fifth of collisions (and nearly a quarter without damage only) were on legacy roads, which have lower flows than all other road types. On both types of single carriageways the collision rate is high, indicating that these are high risk roads for road users.

The proportion of all collisions on non-motorway dual carriageways and link roads is substantially more than the proportion of the network comprising of these road types, giving a higher number of collisions per km. Motorways and non-motorway dual carriageways have the highest average AADT, with motorways having the lowest collision rate (not including roundabouts). In particular, the M50 had a much higher AADT than other motorways.

4.2.1.3 Other collision information

The field 'PCT_Desc' in the collision level data describes the type of each collision. A breakdown of the **collision types** by severity is shown in Figure 16. The most common collision type was 'Rear end', with more than a third of all collisions in this category. 'Side swipe' collisions were the second most common, making up just under a fifth of all collisions. A high proportion of 'Head-on' collisions and collisions involving a cyclist or pedestrian involved injury. The 'Other' category includes 24 different descriptions, each of which relates to fewer than 1,000 total collisions.

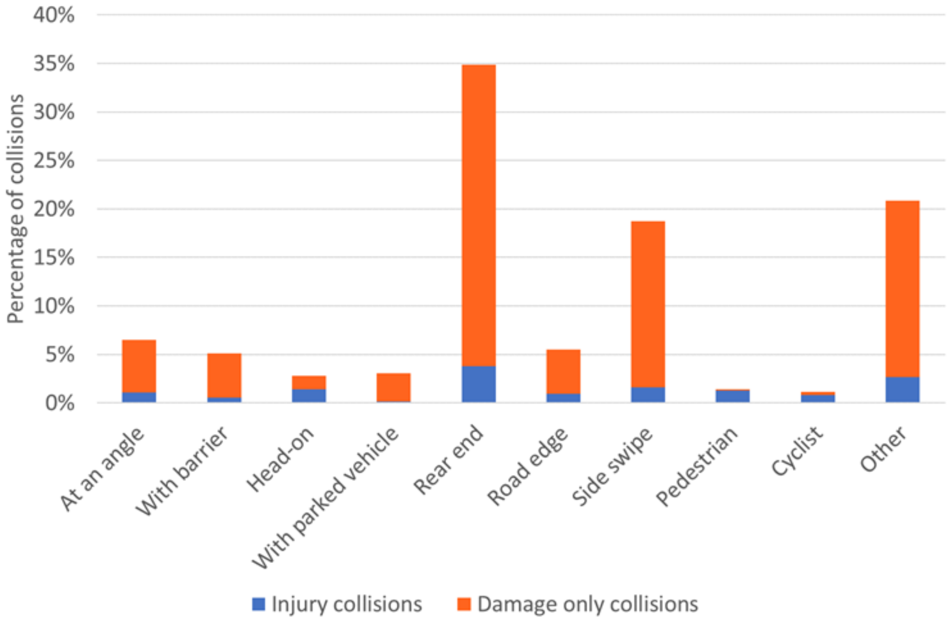


Figure 16: Distribution of collisions by type, 2014-19 (N=53,873)

The distribution of all collisions by **junction type** is shown in Figure 17. Around two thirds of all collisions did not occur at a junction and just over one fifth of all collisions occurred at roundabouts or T-junctions – the most common junction types.

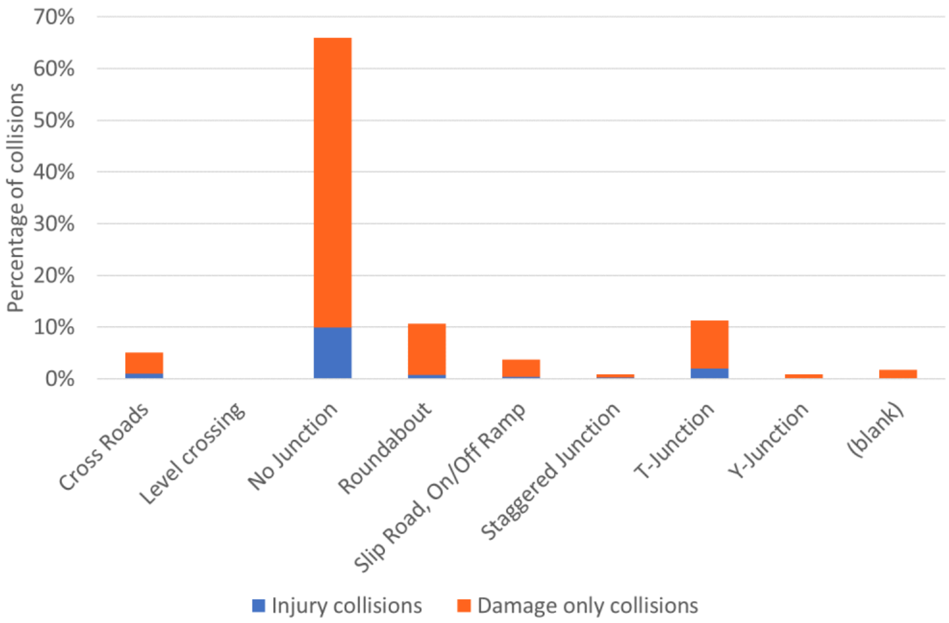


Figure 17: Distribution of collisions by junction type, 2014-19 (N=53,873)

The distribution of collisions by **weather** recorded is shown in Figure 18. The majority of collisions occurred in dry collisions and just over one fifth in wet conditions.

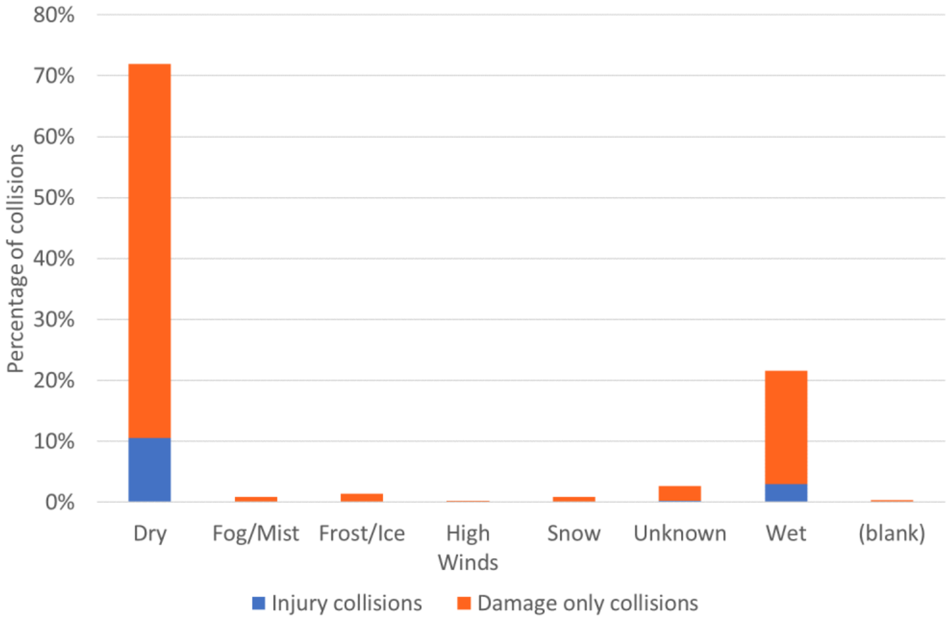


Figure 18: Distribution of collisions by weather (N=53,873)

The number of vehicles involved in the collisions was 98,587 (the same vehicle may be counted multiple times if involved in multiple collisions) at an average of 1.83 vehicles per collision. The distribution of collisions by **number of vehicles** involved is shown in Figure 19. Most collisions involved one or two vehicles and there were 57 collisions involving more than five vehicles. A very small percentage (0.2%) of the collisions were not linked to any vehicles in the data.

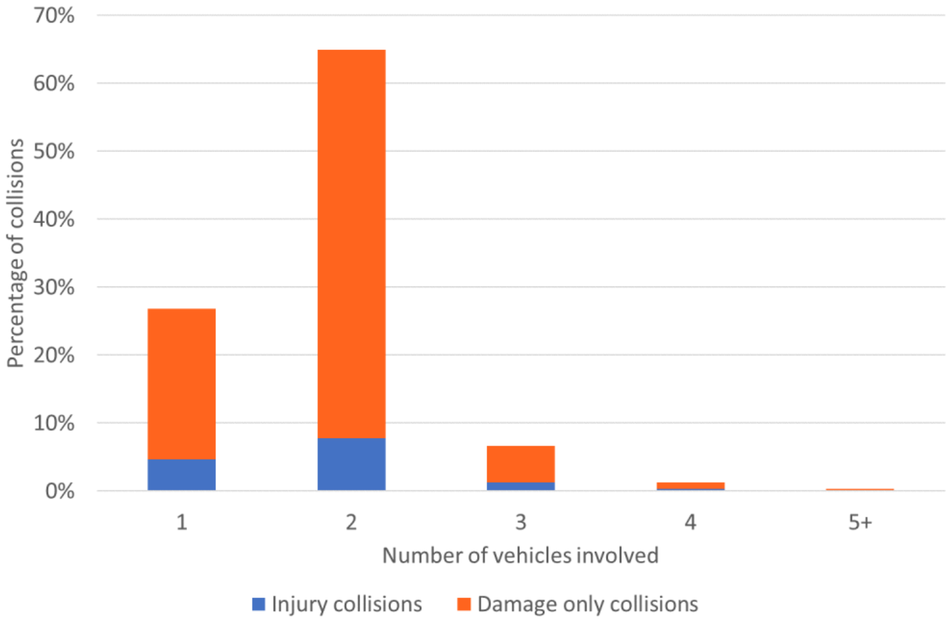


Figure 19: Distribution of collisions by number of vehicles involved, 2014-19 (N=53,746)

The number of collisions by recorded speed limit is given in Figure 20. The majority of collisions (for both damage only and injury severities) occurred on sections of road with a speed limit of 50km/h or 100km/h.

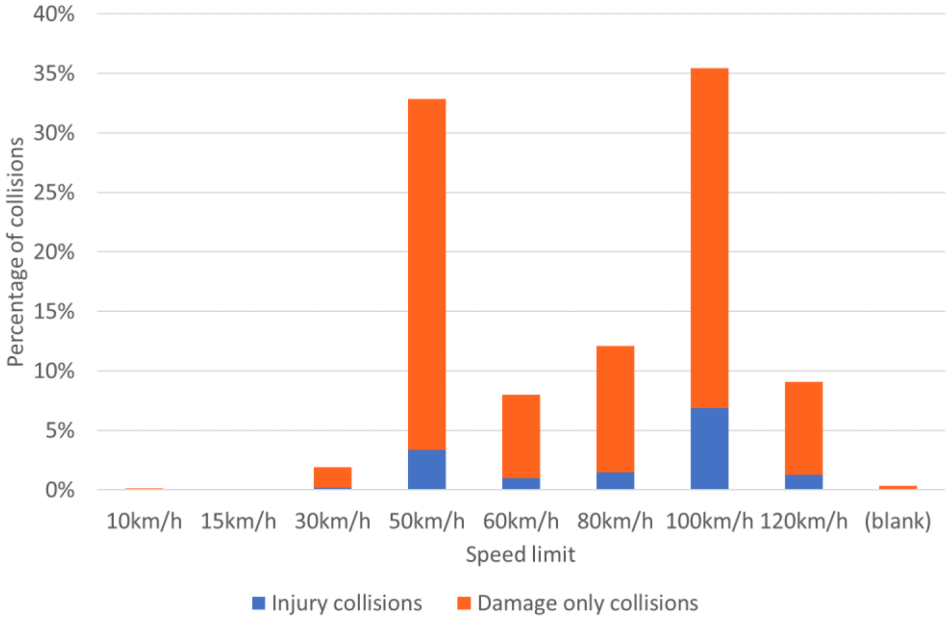


Figure 20: Distribution of collisions by speed limit (N=53,873)

4.2.1.4 Modelling both side of the carriageway

One critical consideration relating to the collision data is whether collisions can be assigned to one side of the carriageway. This would allow modelling of each side of the carriageway separately. From expert judgement and data exploration, the latitude and longitude co-ordinates in the data are not reliable enough to assign collisions to one side of the carriageway. On undivided roads assignment is particularly challenging as vehicles may cross the centreline during collisions (particularly head-on and loss of control collisions). From examination of the fields in the collision data, the free text 'ActionFrom' and 'ActionTo' fields give an indication of direction. However, the descriptions are inconsistent with many entries not location specific with text such as 'school' and 'work', so using these to compute direction of travel is not feasible. Therefore, in the absence of more information, modelling both sides of the carriageway separately will not be feasible from the collisions data.

4.2.2 Explanatory variables

This section summarises the important variables to the modelling process, the values or levels they can have and the suggested dataset(s) from which they are taken. As previously, the variables are discussed by category.

4.2.2.1 Network base layer variables

The most suitable network base layer to which other data can be linked is comprised mainly of the TII GIS base data, which has coverage as a GIS lines layer. This data can be merged with the road type data, traffic data and asset inventory data to give a more comprehensive lines covering of the network.

Table 7 outlines the variables included as part of this base layer.

Table 7: Base layer variables suitable for inclusion

Variable	Possible variable levels	Suggested dataset taken from	Comments on incorporating into modelling
Road section category	- Main line - Link road - Ramp - Roundabout	TII GIS dataset	Over 90% of the network is 'main line'. Each modelled segment will be one of these categories. Segments on link roads, ramps or roundabouts will likely be quite short.
Road type	7 levels based on dual or single and number of lanes: - Dual carriageway (1-3 lanes) - Single carriageway (1-4 lanes)	Road type dataset	Variable levels in this table are a possible enhanced grouping of those in the dataset (discussed below). Ideally modelled segments will have a constant number of lanes and carriageway type.

Consideration will need to be given to the most appropriate separation of the network into road categories or types, likely as a combination of the levels presented in the table above, both in defining the categories for separate models (if appropriate) and splitting up the network into segments. The levels in the road type data are more extensive than those suggested in the table above, including various descriptions such as 'Wide single' and '2 lane road'. Grouping by carriageway type (single or dual) and number of lanes can be achieved with verification from the lane width data.

4.2.2.2 Traffic variables

Table 8 outlines the traffic variables suitable for inclusion in the modelling.

Table 8: Traffic variables suitable for inclusion

Variable	Range of values	Suggested dataset taken from	Comments on incorporating into modelling
AADT (modelled) – given for light and heavy vehicles; one-way and two-way values both presented.	One-way flow values up to 75,000 vehicles per day.	Traffic data	Values presented for light and heavy vehicles, in both directions, gives flexibility for incorporating this variable into the modelling. Using total flow across all vehicle types and % of HGV's would be a possibility.

Variable	Range of values	Suggested dataset taken from	Comments on incorporating into modelling
Speed limit	Values from 0 to 120kph	Traffic data	Using the speed limit data from the traffic dataset is more convenient for data linking (one fewer dataset to be linked).
AM and inter-peak speeds (modelled) – given for light and heavy vehicles	Values from 0 to 120kph	Traffic data	Values for light and heavy vehicles can be treated separately or combined.

Ideally, modelled segments will not combine road sections with vastly different flows or speeds. However, if this is the case, weighted averages may be used to assign a single value for AADT and peak speeds to a modelled segment. Values could also be grouped (e.g. AADT from 50,000 to 75,000) so that each modelled segment is assigned a speed or flow range rather than a single figure. It is unlikely that both speed limit and peak speeds will be included in the model as there will almost certainly be a strong correlation between them.

4.2.2.3 Road geometry and condition variables

Table 9 outlines the road geometry and condition variables suitable for inclusion in the modelling.

Table 9: Road geometry and condition variables suitable for inclusion

Variable	Levels/Range of values	Suggested dataset taken from	Comments on incorporating into modelling
Gradient	Absolute values vary from 0 to 8.75 degrees (can be negative or positive)	PMS overall survey data	If modelling both sides of the carriageway combined, absolute values will be used.
Crossfall	Absolute values vary from 0 to 7.76 degrees (can be negative or positive)	PMS overall survey data	If modelling both sides of the carriageway combined, absolute values will be used.
Radius (curvature)	Absolute values from 0 to 10 km (can be negative or positive)	PMS overall survey data	This variable is used to assess curvature. If modelling both sides of the carriageway combined, absolute values will be used.
SCRIM value	Values from 0 to 1	PMS overall survey data	

Variable	Levels/Range of values	Suggested dataset taken from	Comments on incorporating into modelling
Minor junction	• Crossroads • Junction left • Junction neutral • Junction right • T-junction	Junctions data	Minor junctions could be incorporated in a number of ways. For example, the number of minor junctions in a segment can be counted to give a junction density or segments can be split up by junction locations.

As all the variables from the PMS overall survey dataset are presented at 10m intervals, this data needs to be aggregated when creating segments, for example with averages over all the points in a segment, or using the maximum or minimum value. It is important to ensure that modelled segments do not span 10m points with vastly different geometric values.

Curvature is given by the radius field as these parameters have an inverse relationship. The radius of a point represents the radius of a circle (in km) drawn from that point, according to the bend in the road. Therefore, the larger the radius value, the less curved the road.

Major junctions such as roundabouts can be counted along with minor junctions to give a total junction density value for a segment, if excluding major junctions from the modelling.

4.2.2.4 Roadside features variables

Table 10 outlines the roadside features variables suitable for inclusion in the modelling.

Table 10: Roadside features variables suitable for inclusion

Variable	Levels/Range of values	Suggested dataset taken from	Comments on incorporating into modelling
Safety barrier - location and material	<u>Location</u> of safety barriers is indicated by the presence of lines on the map <u>Material</u> • Concrete • Steel • Wooden • Wire • Other	VRS data	The location is the most crucial information from the modelling perspective. If modelling both sides of the carriageway together consideration will be given to incorporating this variable appropriately.

Variable	Levels/Range of values	Suggested dataset taken from	Comments on incorporating into modelling
Access density (split by building function)	Access density estimated by number of buildings (within 1km of the network) divided by segment length Function • Residential • Business • Commercial	GeoDirectory data	This GeoDirectory data is a proxy for access density in the absence of more detailed access information. There is flexibility in the exact data used in the modelling and the distance threshold to the modelled segment.
Urban or rural	Urban ‘yes’ or ‘no’	Lane width data	

The GeoDirectory data has the most convenient information for obtaining a proxy for access density. However, consideration needs to be given to the most suitable way of incorporating this data as an accurate measure of access density. The regional aggregation in the census data makes using this data difficult and the extra detail presented in the PRIME2 data is not relevant.

As data on hard shoulder locations is incomplete across the network and within road types, the impact of hard shoulders on collisions is not reliably quantifiable. Therefore, this variable can’t be incorporated into the modelling.

Through data exploration, the urban ‘yes’ or ‘no’ field in the lane width data was found to match the subnetwork classifications in the PMS asset inventory data (subnet ‘2’ = urban, else = rural) and this classification looks the most reliable. This breakdown into urban and rural is illustrated in Figure 21. A much higher proportion of the network is marked as rural and rural sections are generally much longer.

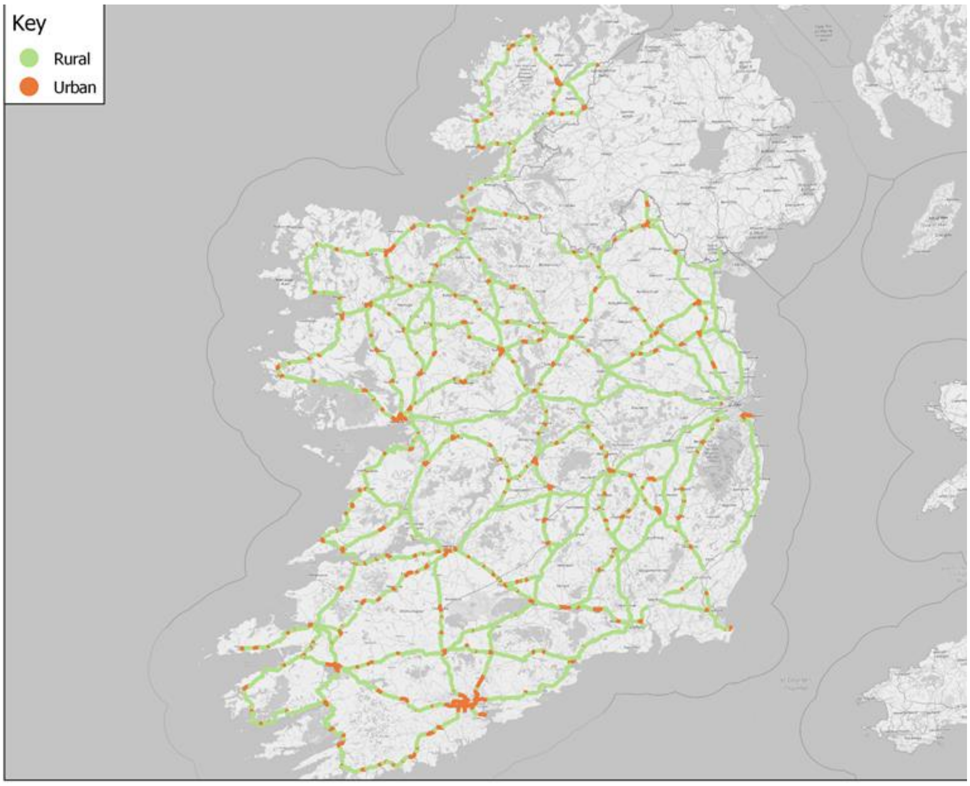


Figure 21: Urban and rural split in the lane width data

Ideally, modelled segments will be either entirely urban or entirely rural, however where that is not the case the more prominent of the two can be assigned to that segment. The same applies to the presence of a safety barrier. Alternatively, a '% of segment with safety barrier present' variable could be used. If modelling both sides of the carriageway combined, consideration will need to be given to segments where a barrier exists on one side of the carriageway only. The vast majority (approx. 90%) of safety barriers present on the network are steel so the material variable may not be useful for modelling.

4.3 Data linking

This section outlines the method for creating a joined georeferenced database of all variables to facilitate easy segment generation for modelling. The flow chart in Figure 22 summarises the method.

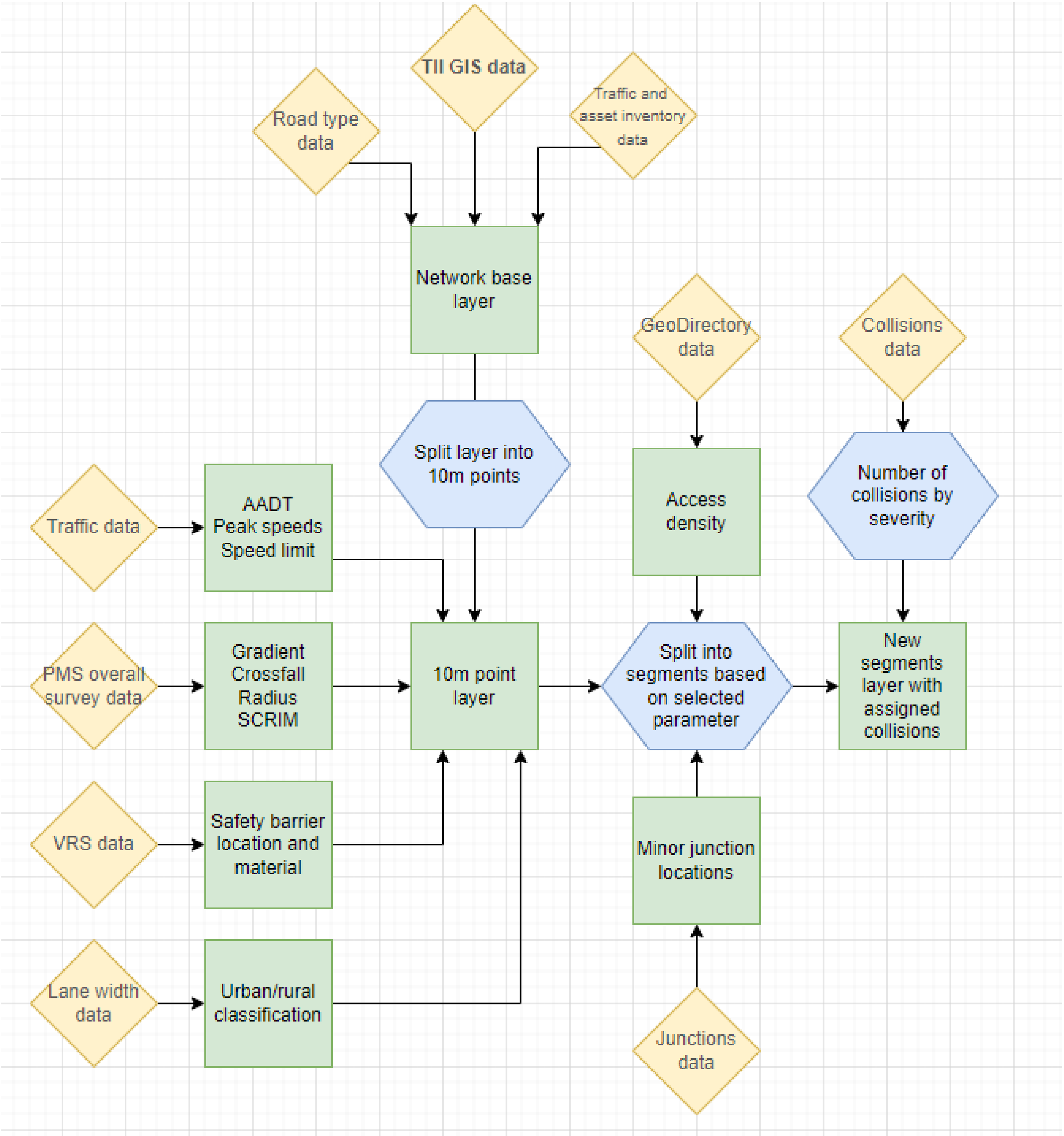


Figure 22: Method for linking all the data together and creating segments

Base Layer

First, the network base layer must be created, to which all other variables can be linked. To achieve this, the TII GIS dataset is combined with the traffic data, asset inventory data and road type data to give a GIS line covering of the network. The TII GIS data splits the network according to four section types: main lines, link roads, ramps and roundabouts. Main lines represent the vast majority of the length of the network. The road type dataset assigns each road section a number of lanes and carriageway type – dual or single.

Explanatory variables and section definition

The most granular data is the road geometry and condition data, with values specified every 10m. Therefore, the base layer is split into 10m points (aligning with those in the geometry and condition data), on the centreline of each road, to which the other variables can easily be linked using GPS location and matching algorithms. After data linking, each 10m point has:

- A section type (main line, link road, ramp, roundabout)
- A road type (e.g. 3 lane dual carriageway)
- AADT values, peak speeds and speed limit from the traffic data
- A gradient, crossfall, radius and SCRIM value from the PMS overall survey data
- A safety barrier 'yes' or 'no' and material from the VRS data
- An 'urban' or 'rural' classification

Where they exist, values for both sides of the carriageway can be assigned to each 10m point, for example AADT values and gradients for both directions.

The values at the 10m point level can then be used flexibly to split up the network into segments, for example by curvature or traffic flow (see Section 4.4). Once the segments have been created, values at 10m can be aggregated, for example by averaging or taking absolute values, to assign a single value for every explanatory variable in the model to each segment. If both sides of the carriageway are modelled together, direction specific values will be combined, such as traffic flow (by using two-way flow) and gradient (by taking an average of absolute values). GeoDirectory data on access points and junctions data can also be incorporated once segments have been created. By converting the line-based approaches to junctions (as in the original dataset) to single points, minor junctions are added on the section, and the number or density of these junctions in each section can be counted.

Collision data

Once all the explanatory variables have been linked, the collisions can be linked to the segments using their latitude and longitude co-ordinates by applying a buffer around the network and/or using the route name. This gives each segment a number of collisions by severity.

4.4 Collision distributions by segment definition

This section outlines possible methods for defining the segments for modelling (in accordance with the findings from the literature review) and assesses the resulting distribution of collisions. To demonstrate these methods, a 90km section of the road network comprised predominantly of the M1 is taken between Dublin and Newry, with suitable variation in the defining parameters. As previously, a 10m buffer is used to link the collisions to the network.

4.4.1 Segments by traffic flow

A possible method for defining segments according to traffic flow is as follows, using the 2019 traffic data:

- Each 10m point has a two-way AADT value from the creation of the joined database described in Section 4.3;
- The difference in two-way AADT between each 10m point and its adjacent points is calculated;
- Any two-way AADT differences above a certain threshold are extracted;
- The points that give these large differences are then used as section divides on the original base layer

Figure 23 shows a screenshot of the M1 around Dundalk split into segments defined by this method, with an AADT difference threshold of 5000.

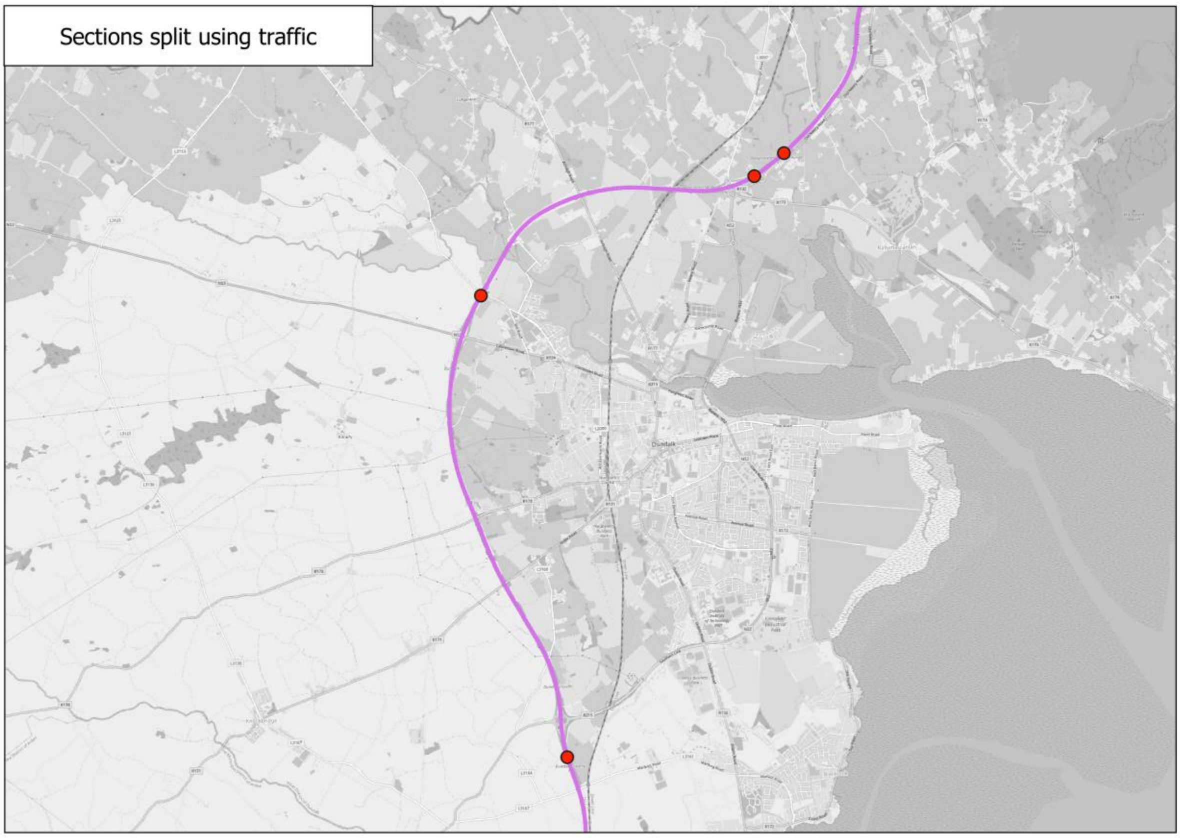


Figure 23: Split into segments around Dundalk by AADT with a threshold of 5000 vehicles per day

According to this method, the length of the segments varies greatly. The lengths of the defined segments and the number of collisions on each segment is given in Figure 24. There are 17 segments across the 90km road section, varying in length from 257m to 26,219m. There are no segments with zero collisions and the lowest number of collisions across all segments is 6.

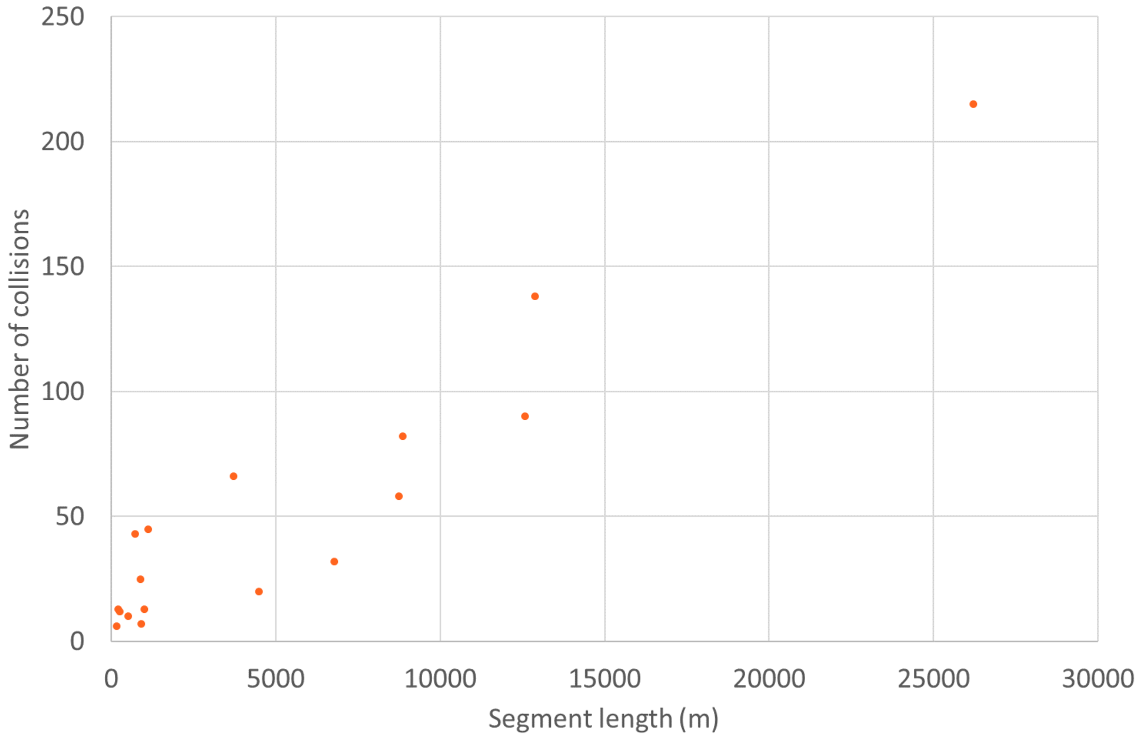


Figure 24: Number of collisions on segments and their lengths

4.4.2 Segments by curvature

A possible method for defining segments according to curvature is as follows:

- Each 10m point has a radius (curvature) value from the creation of the joined database described in Section 4.3;
- A rolling average is applied to smooth the data and avoid any potential spikes in curvature;
- The difference in curvature between adjacent points is then calculated;
- Any curvature differences above a certain threshold are extracted;
- The points giving these large differences are then used as section divides on the original base layer

Figure 25 shows a screenshot of the road network split into segments defined by this method, with a threshold radius value of 2.

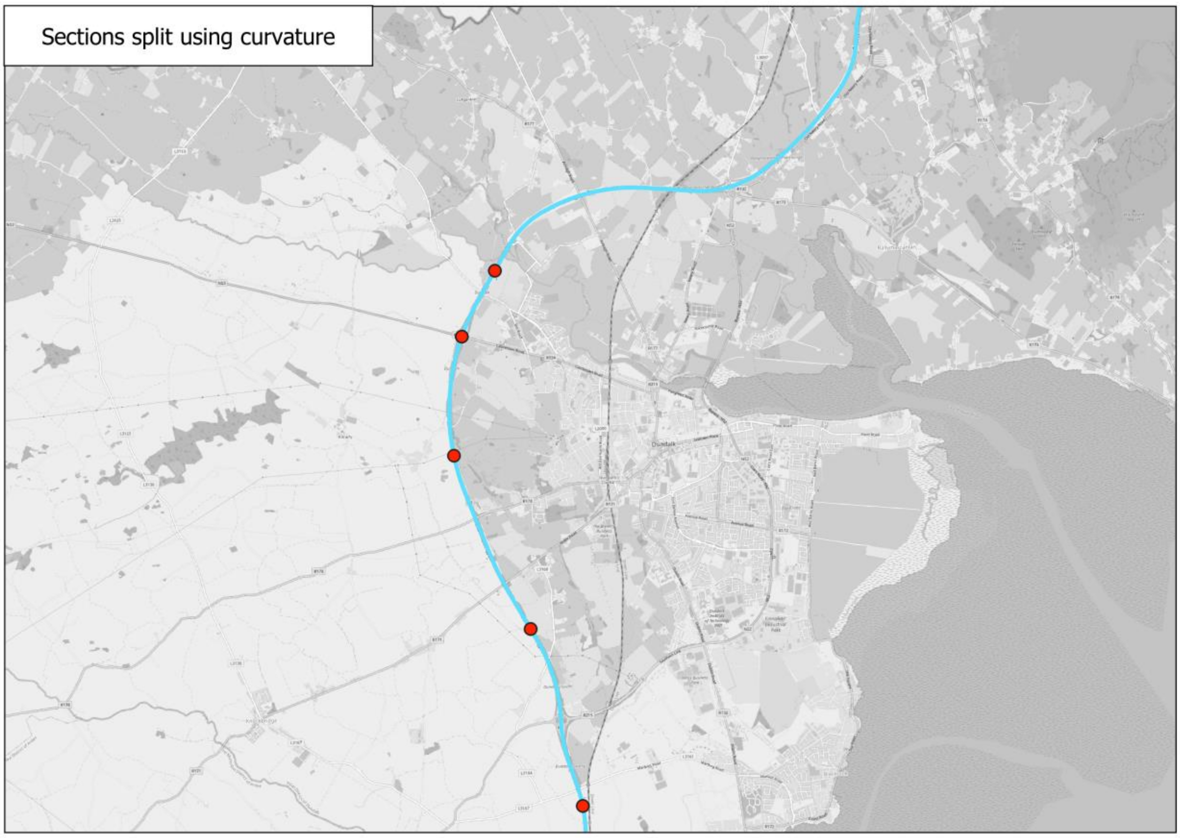


Figure 25: Split into segments around Dundalk by curvature

The lengths of the defined segments and the number of collisions on each segment is given in Figure 26. Of the 18 segments, only one has fewer than 10 collisions. As with defining segments by AADT, the lengths of the segments vary considerably; however, none of these segments defined by curvature have a length less than 1km.

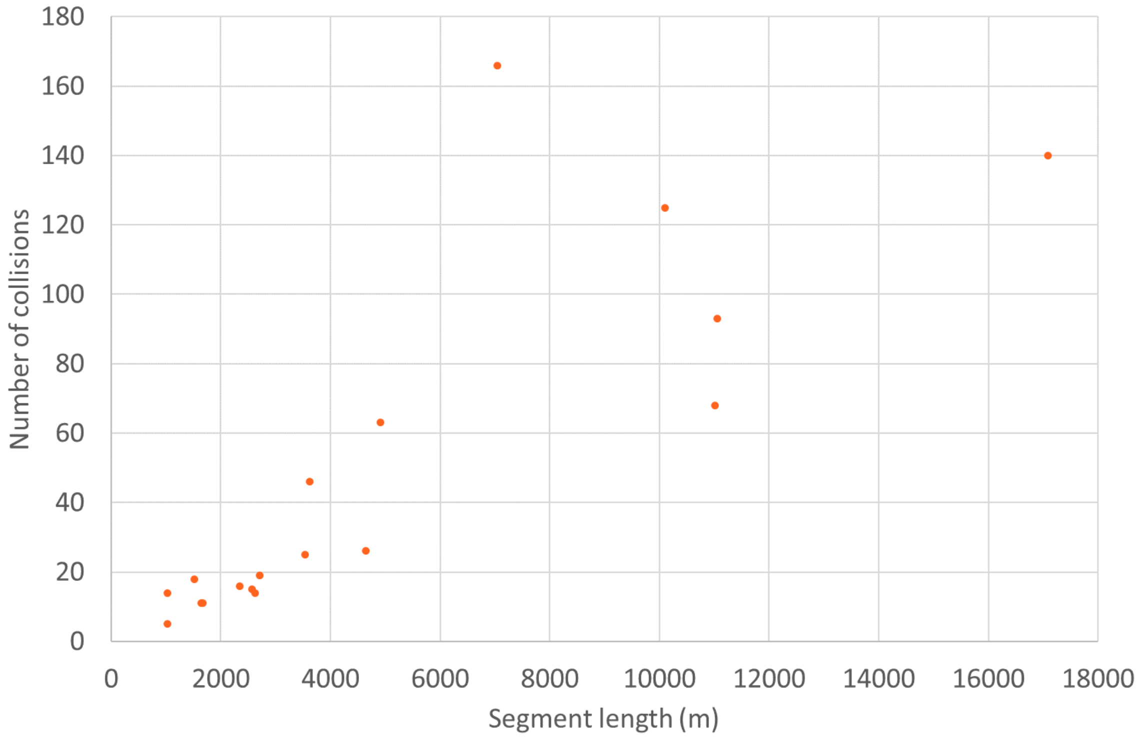


Figure 26: Number of collisions on segments and their lengths

4.4.3 Segments by length

A possible method for defining segments by length is as follows:

- A target length for segments is identified, *X* km;
- The base layer is split into (at most) *X* km segments. Road sections already under *X* km, such as roundabouts, ramps and link roads will remain as their original lengths.

With *X* = 1, Figure 27 shows a screenshot of the network split up into segments by this method.

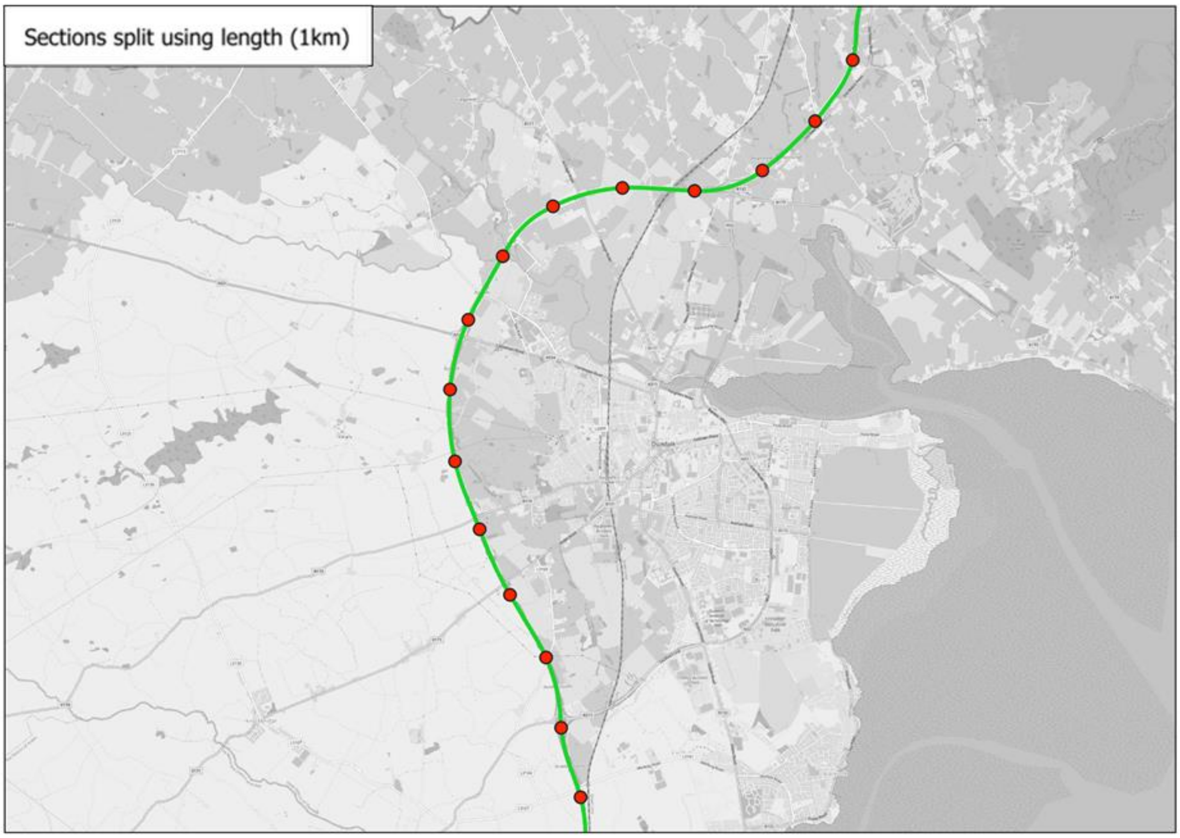


Figure 27: Split into segments around Dundalk by length (1km)

The distribution of the number of collisions on these segments is shown in Figure 28. None of the 90 segments have zero collisions; however, ten segments have three or fewer collisions.

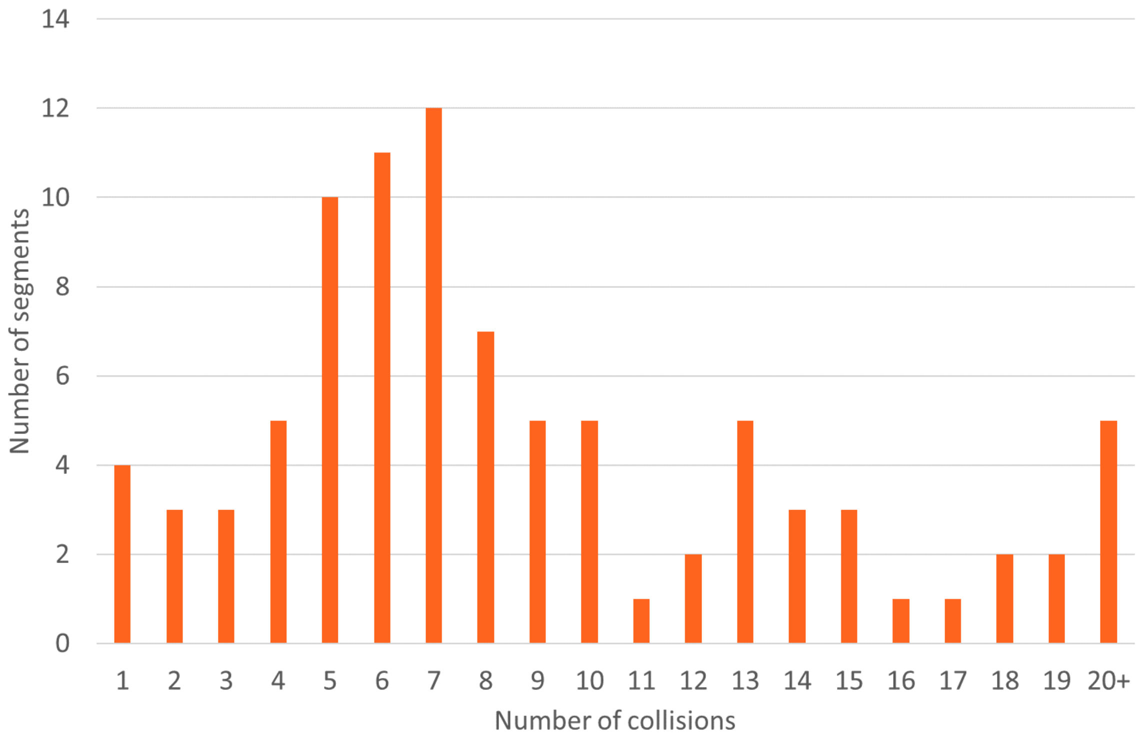


Figure 28: Number of segments with different numbers of collisions

4.4.4 Discussion

The examples above are for a motorway and dual carriageway route close to Dublin, which has a higher density of collisions. For more rural routes with lower flows these methods may need to be adapted to ensure appropriate segments are defined. Single carriageway roads, especially the legacy roads, will have more variation in curvature, and therefore segments based on curvature may be shorter (depending on the threshold used) and hence have fewer collisions on them. Figure 29 below illustrates part of a more rural section of the network split according to curvature, by an alternative method. This method uses the physical dimensions of the road line rather than the curvature data as this was deemed more accurate for this section:

1) The network is split into sections of 100m
2) The radius (curvature) is calculated for each 100m section using the physical dimensions of the line
3) A threshold of 0.5km is used to extract sections deemed to be curved
4) Adjacent 100m sections that are curved or straight are combined

The average segment length is 2,261m and the average number of collisions per section is 12. There are no segments with zero collisions.

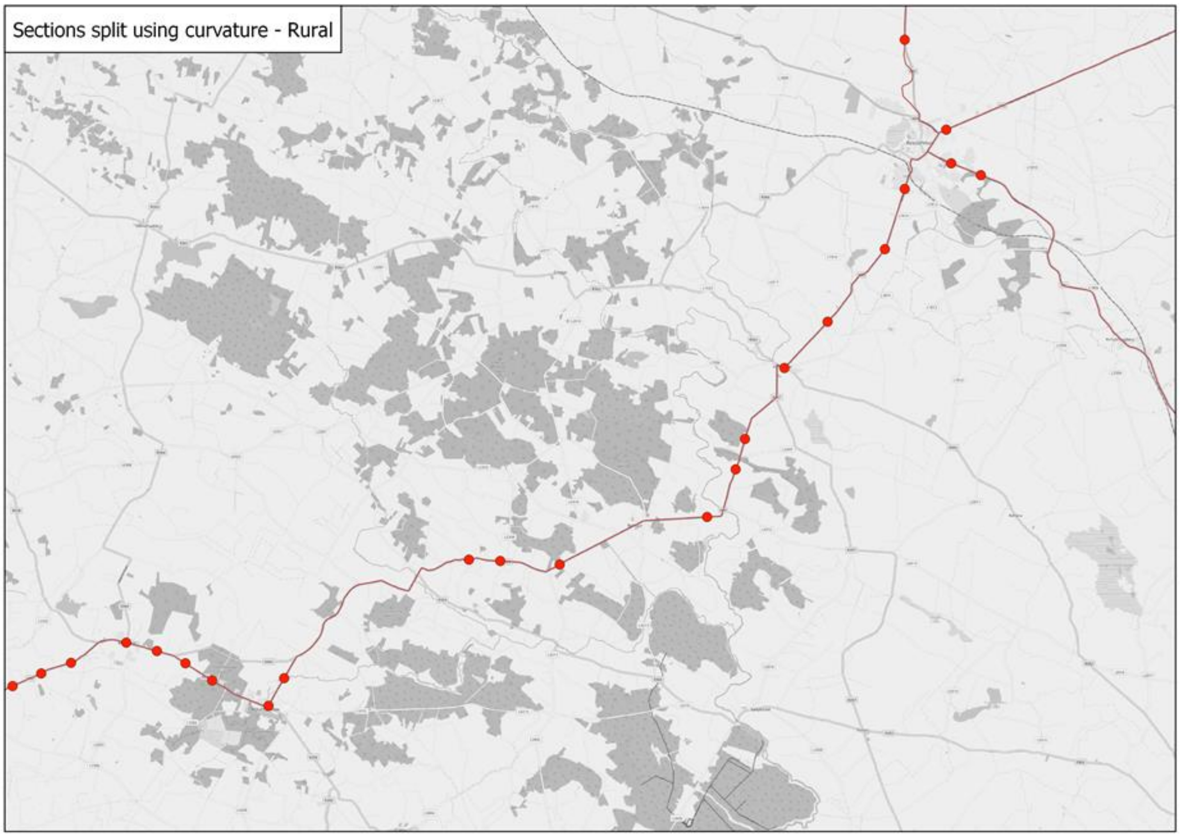


Figure 29: Split of a more rural section of the road network by curvature

More rural routes may also have smaller variations in traffic flow and therefore require a different AADT threshold for creating new segments. Using the same method as in the previous section, with a smaller threshold of 2,000 vehicles per day, the inconsistency in segment lengths on this rural road section is large (illustrated in Figure 30 below). There are multiple segments of more than 25km in length and multiple of less than 1km in length. As such, the number of collisions also varies greatly from 1 to 217.

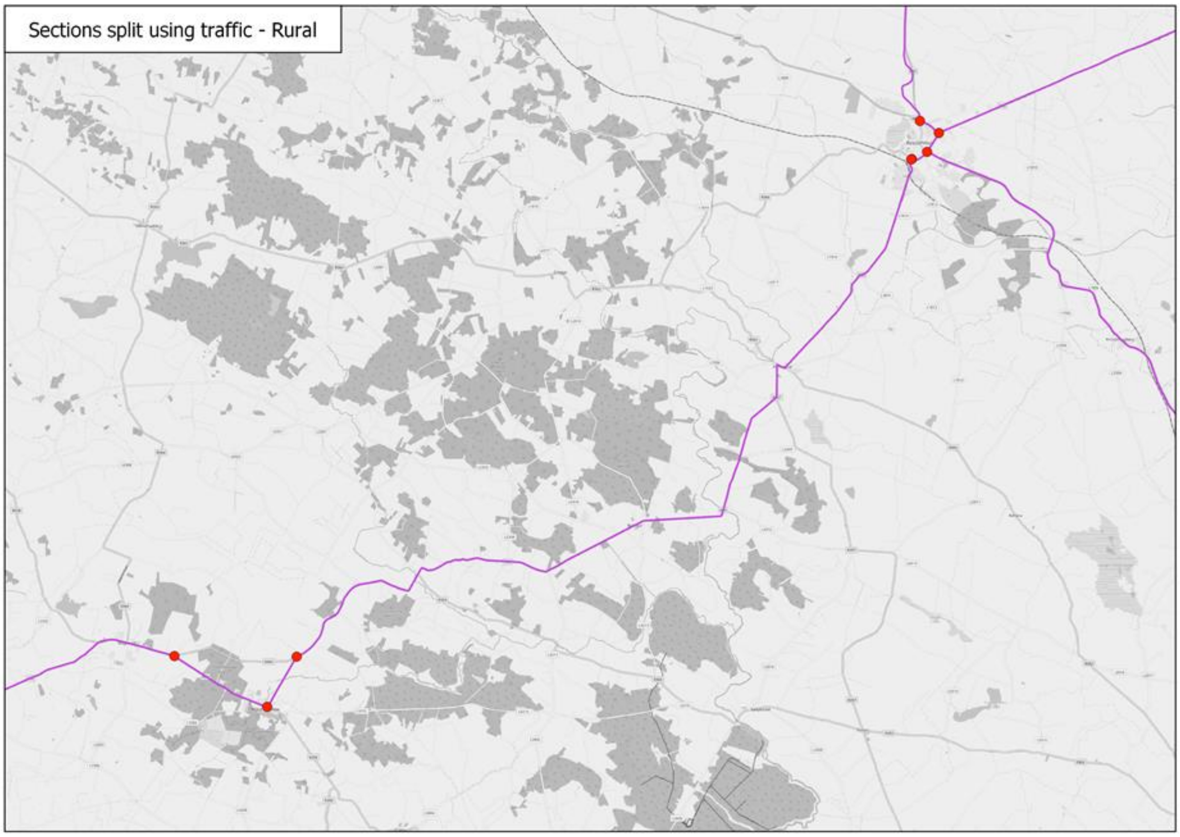


Figure 30: Split of a more rural section of the network by traffic flow

For segments of fixed length, the optimal length would have to be carefully considered. When segments are too short the number of zero collision segments will be high (see Section 4.2.1.2) and when segments are longer the variation in other parameters will increase. One option would be varying the length by region or road type to account for lower densities of collisions in rural areas and higher densities on dual carriageway roads.

An extension of these methods would be to use a combination of variables to define the segments, for example, traffic flow and curvature, though the thresholds may need to be wider when used in combination to reduce the number of very short sections with few or zero collisions. In combination with curvature and/or traffic flow, a minimum or maximum length threshold could be used to avoid modelling segments that are too short or too long.

Things to consider for model development

- Which years to use in the modelling
- The suggested collision data for use in the modelling is the raw data at vehicle and collision level:
  - Co-ordinates give flexibility for linking to segments and extra detail is present in this data, for example: collision type and junction type
  - The method for linking this data needs to ensure that all collisions on the network are accurately captured; A buffer of 10m is suggested; however, just over 20% of collisions are not linked in this way from their co-ordinates, mostly because legitimate co-ordinates are not present in the data
- Whether to include damage only collisions:
  - Incorporating damage only collisions increases the total number of collisions from 7,641 to 53,873 and the average number of collisions per km from 1.5 to 10.2; this reduces the likelihood of segments with zero collisions
- The number of zero collision segments, if determining segments by length:
  - With segments of length 100m, 67% of segments had zero collisions compared with 10% for a length of 2km
- Modelling each side of the carriageway separately will not be feasible due to lack of data to accurately assign collisions to separate carriageways
- The most appropriate separation of the network by road type:
  - This is likely to be the four main road types (motorways, dual, single and legacy single), with roundabouts, ramps and link roads excluded because they make up a relatively small amount of the network
- Aggregation of variables will be required for assigning values to segments:
  - Weighted averages may be used for variables such as AADT and speed
  - Ranges or groupings may also be applied and rolling averages can reduce noise in the data (for example with curvature)
  - Taking an average, minimum or maximum may be appropriate for geometric or road condition variables such as gradient and curvature
  - Density values are more appropriate for counting the number of junctions or considering the number access points
  - Most of the network is rural; urban sections are typically much shorter
  - For safety barriers a ‘% of segment covered’ variable is useful

Things to consider for model development - continued

- There may also be strong correlations between some variables (such as speed limit and peak speeds) which need to be investigated prior to building the model to avoid confounding factors
- With all variables assignment will depend on whether both directions are combined for the modelling; for example two-way AADT is more appropriate if modelling both directions combined
- When defining segments according to metrics such as curvature and AADT:
    - Different thresholds may be required for different road types or regions to ensure that segments are of appropriate length
    - Segment lengths (and therefore number of collisions) may vary greatly if using curvature or AADT; a minimum or maximum length threshold could be applied for greater consistency
    - A combination of variables could also be used to define segments, though thresholds may need to be wider to ensure segments are sufficiently long

5 Task 3: Development of methodological approach

Task 3 brings together the findings from Task 1 (literature review) and Task 2 (data assessment) and assesses the feasibility of developing APMs for the Irish national road network. The aim of this task is to makes recommendations on the best approach, given the methodological review and the data available, and to highlight the likely outcomes and limitations with this approach.

Section 5.1 gives an overview of the recommended method for developing the APMs; Section 5.2 outlines the data sources to be used and the potential road safety interventions which could be evaluated once the models have been developed; Section 5.3 presents a more detailed step-by-step methodology for developing the APM. Finally, Section 5.4 summarises the risks and limitations of the proposed approach.

5.1 Recommended methodological approach

5.1.1 Type of model developed

The literature review identified Generalised Linear Models (GLMs) to be the most common type of model used to develop APMs. However, another approach used in one study was Generalised Estimating Equations (GEE) which accounted for time trends. Whilst annual traffic and collision data are available for six years (2014-2019), road geometry and condition data are only available as a snapshot in time, from the PMS survey conducted in 2021 (see Table 4). This survey, therefore, does not enable any understanding of changes in road features over the period of interest and as a result, it is not possible to used time-trend based GEE models for this study.

Based on the findings from the literature review, it is recommended to use GLMs to develop the APMs for the national road network. Depending on distribution the response variable, the most appropriate distribution (Poisson or Negative Binomial) will be used to model the data. The method for identifying the appropriate distribution is discussed later in Section 5.3.2.

Furthermore, depending on the distribution of collision numbers across segments, a zero-inflated GLM may be considered. However, it is likely that using homogenous road segments will negate the need for a zero-inflated model, as seen in most studies included in the literature review. The process for this is described in Section 5.3.2.

5.1.2 Models for specific road types or crash types

All studies in the literature review modelled APMs for a specific road type. This is mainly due to differences in characteristics between different road types. Therefore, Task 2 (Table 6) explored the distribution of the network and collisions by various road types and identified four road types which covered a substantial part of the network and had sufficient sample sizes to develop an APM. The proposed models are shown in Figure 31.

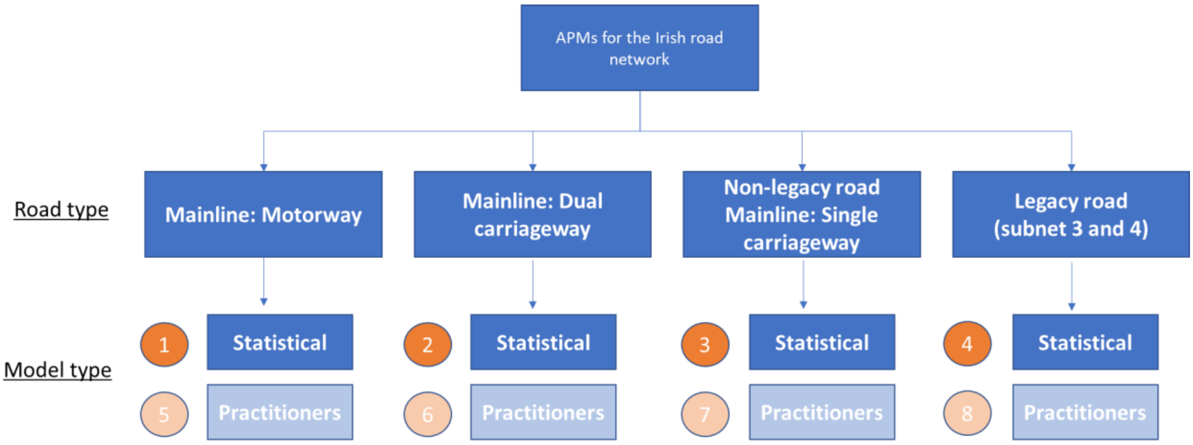


Figure 31: APMs for national road network

TRL propose developing **four APMs** for the national road network, split by road type. This would account for just over 90% of the overall road network length and close to 90% of collisions. Details on the sections excluded from these models (i.e. the remaining 10% of network length) are included in Section 5.4.1.

In addition to statistical models, where variables are chosen to be included based on their statistical significance, some papers in the literature review also developed practitioners' models which include variables of practical interest to the road safety authority. The variables in these models may not all achieve 95% confidence but are included based on an assessment of their relevance to the practitioner. Once the statistical models have been developed and their variables and predictive power evaluated, TRL will consult with Transport Ireland Infrastructure and discuss the possibility of developing an **additional four** practitioners' APMs, based on the same data as the statistical APMs, that will focus on variables of interest, irrespective of the significance level.

A number of papers in the literature review also modelled various collision types individually. Task 2 explored the distribution of collisions by various collision types and found that majority of the collisions were either rear end, 'other' or side swipe (Figure 16); the remaining types accounted for less than 10% of the collisions. Due to the small sample sizes for most of the collision types and the vagueness of the category 'other', individual APMs will not be developed for each collision type. However, the proportion of rear-end collisions will be included as an explanatory variable in the model if it improves the overall model fit.

5.1.3 Division of the network into segments

The literature review found that most studies created homogenous road segments on which to develop the APMs (see Section 3.1.3). The variables most commonly used to develop these homogenous road segments were AADT and curvature; other variables such as road width were also considered in some studies, but the use of these were less common. Based on these findings and an exploration of the best method for creating homogenous segments (see Section 4.4), TRL propose using AADT and curvature to develop homogenous road segments for each of the four road types being modelled. It must be noted that the road segments will combine data from each side of the carriageway, as it is not possible to assign collisions to a particular direction of travel.

TRL propose using a multi-stage approach to create homogenous road segments. These steps will be applied independently to each of the four road types:

1. Data pre-processing: Calculate average AADT across the six years to obtain a single value for each road section. Curvature data has been collected in 2021 as a snapshot in time so no aggregation would be required.
2. Homogenous road segments: Use a combination of AADT and curvature to identify homogenous road segments. A road segment can be considered as homogenous when the variables being used are within a given threshold within the segment. Any data outside of the threshold of either variable mark the end of one segment and beginning of the next. In order to do so, appropriate thresholds for AADT and curvature will be defined based on the road type and distribution of these variables on the national road network. This has been illustrated in Section 4.4 where different thresholds for AADT and curvature were applied to a section of rural road and motorway.
3. Minimum segment length: Many papers in the literature review defined a minimum segment length (ranging from 50m to 2km) to avoid having too many segments with no collisions. After the homogenous road segments have been developed, TRL will review the segment lengths (similar to the illustrations in Figure 24 and Figure 26) and, if appropriate, apply a minimum threshold. Segments lengths that are smaller than these thresholds will be combined with other road segments which have similar values for AADT and curvature.

   To get an idea of the minimum segment length that should be applied to the dataset, Task 2 (Section 4.2.1) explored the number of collisions on each segment when using various segment lengths. This analysis showed that 67% of the 100m sections had zero collisions (including damage-only), and this reduced to 27% for 500m sections, 15% for 1km sections and 10% for 2km sections. Given the large reduction in zero collisions between 100m and 500m, TRL expect that the minimum segment length will be between these two thresholds. However, the final choice will be determined for each road type independently in the next phase of the study.
4. Maximum segment length: Based on the outcome from step 2, TRL will also review if a maximum segment length threshold is required. If so, other explanatory variables (such as number of lanes or urban/rural location) may be used to identify the most appropriate method for splitting long road segments.

The threshold for the homogenous road segment, minimum and maximum length will be defined in the next phase of the study using the national road network data. It is not possible to define these using values from other countries or datasets (such as those from the literature) due to the inherent differences in traffic flow and road geometry between countries.

5.2 Proposed data sources

5.2.1 Response variable in the models (number of collisions)

5.2.1.1 Duration of collision data

Task 2 (Section 4.1.2 and 4.2.1) identified the raw collision data from 2014 to 2019 as the most appropriate data source to be used for the collision modelling. This aligns with the recommendation from the literature review (Section 3.1.1) to use around five years of collision for the statistical model, avoiding the influence of changes in collisions due to long-term trends or policy influences. All of the papers in the literature review modelled collision data and not casualty data, as collisions can be influenced by interventions and have a direct impact on casualty numbers.

Therefore, it is proposed that six years of collision data will be combined and the total number of collisions in between 2014 and 2019 on each road segment will be the response variable in the statistical model. This would also avoid any impacts of the COVID-19 pandemic on collision numbers.

5.2.1.2 Collision severity

The literature review highlighted that the collision severities included in the modelling varied across papers: different approaches were taken to the inclusion of damage-only collisions depending on the data availability. Task 2 identified that the majority of the collisions on the network were material damage only (roughly 85% of all collisions) and the remaining 15% were injury-based collisions (Table 5). Furthermore, when the network was split into 1km sections, around 53% of the sections had zero collisions over the six years when looking at injury collisions only; whereas this reduced to 15% when damage-only collisions were included.

Based on this, it is recommended to include damage-only collisions in the APMs to increase the sample size for the statistical model. However, it must be noted that including damage-only collisions has its own pros and cons. While the main advantage is that it provides a much large sample size to develop the statistical model, the disadvantage is that there may be under-reporting of damage-only collisions and there may be variations in reporting methods by region or police force.

5.2.2 Variables in the base models

For the analysis in Task 2, traffic data from the National Transport Model was used from 2015 to 2019 for each direction of the carriageway; this covers the entire road network. In order to align with the collision data, traffic data from 2014 will also be included in the next phase of this study.

Table 11 summarises the variables that will be included in the base model and the potential form of the variable in the model. The base model is also known as Safety Performance Function (SPF) and has been used by other studies in literature.

Table 11: Variables included in the base model

	Variable	Variable form in the model
Base model	AADT	Average AADT across the six years for the overall traffic will be included in the model. Various functional forms (power, exponential etc.) will be explored.
	Road segment length	This will be derived from the homogenous road segments (in Section 5.1.3)
	Number of lanes[13]	This will be included to account for the fact that different roads will have different number of lanes, and this is likely to impact capacity of the road and the number of lane change/overtaking collisions (and thus the overall collision risk).

For the purposes of model development, we propose using six years of traffic data in the following manner:

- Combining AADT from each side of the carriageway. This is mainly because location information from the collision data is not accurate enough to determine which side of the carriageway the collision occurred on.
- Combining AADT from six years to obtain a single value for AADT across six years. In order to do so, AADT will be converted back to traffic in veh-km, averaged across the six years, and converted back to single average AADT value across the six years.
- Although AADT is available for heavy and light vehicles separately, we propose using the overall AADT for modelling purposes and including proportion of HGV traffic as a separate variable in the model. This would account for both effect of overall traffic on collisions, and the impact of HGVs which are known to create speed differentials due to speed limiters and are known to affect collision severity.

5.2.3 Explanatory variables tested in the models

The literature review identified a number of explanatory variables that were commonly used in APMs. These were lane dimensions, shoulder dimensions, median dimensions, curvature related variables and gradient related variables. Task 2 explored the road geometry and features variables in greater detail and Table 12 presents a summary of variables that will be suitable for testing in the APMs. Note that not all of these variables may end up being included in the final model, each will be assessed based on the contribution it offers to explaining collision risk (see Section 5.3 for information on how the models will be developed).

[13] TRL will review if there is correlation between AADT and the number of lanes. If multicollinearity issues are identified, then it may be necessary to apply appropriate transformations will be applied to the variables to deal with this, or consider the necessity of including both variables in the model.

Table 12: Potential variables to be included in the APMs

	Variable	Variable form in the model
Speed	Modelled AM peak and inter-peak speed	Average speed by vehicle type (light or heavy) across the six years will be included as a continuous variable in the model. Further exploratory analysis is required by road type to understand the likely range of values of these variables within each model, and the appropriate variable form (continuous or categorical).
Road geometry and condition	Gradient	The absolute maximum gradient across both carriageways will be used.
	Crossfall	Average of the absolute value (for both sides of the carriageway) will be used.
	Radius (curvature)	The tightest point (or minimum value) of a curve will be used based on the absolute value for each side of the carriageway.
	SCRIM value	Proportion of road segment with SCRIM value less than X, where X varies depending on the type of road being modelled.
	Junction density (major/minor, or by junction type)	Number of major or minor junctions in the road segment (or number of crossroads, T-junctions etc.) will be included as a continuous variable. The most appropriate form will be decided based on the variability present in each dataset.
Roadside features	Safety barrier: location (carriageway sides)	Proportion of road segment with a safety barrier present on both sides of the carriageway.
	Safety barrier: location (centre line)	Proportion of road segment with a safety barrier present on the centreline.
	Safety barrier: material	Included as a categorical variable
	Access density	The number of buildings within the segment will be used as a proxy to determine access density.
	Urban or rural	Included as a categorical variable.
Other variables that impact collisions	Proportion of traffic which are heavy vehicles	This will be averaged across the six years and included as a continuous variable (the reasons for which are explained in Section 5.2.2).
	Proportion of rear-end collisions	This will be estimated as a percentage of overall collisions across the six years and included as a continuous variable.

After the base model has been developed, speed, road geometry, road features and other variables will be individually tested for inclusion in the model.

The traffic data comprises of information on AM peak and inter-peak speeds split by light and heavy vehicles, and speed limit. As speed limit on the road does not represent the speed a driver chooses to drive at, especially in free-flow traffic, including this variable in the model will not accurately represent the speed distribution on the roads. It is also difficult to develop any interventions around speed limits when there is little information on what speed drivers would choose to drive at on those roads (under free-flow conditions). As a result, we propose to use the modelled AM peak and inter-peak speeds split by vehicle type (light or heavy) in the APM.

For the geometry variables, e.g. gradient, absolute values are necessary since combining data from two sides of the carriageway one with an ascent and one with a descent (as would be the case with a hill), the average value would not accurately represent the gradient of this segment. Therefore, the absolute maximum gradient across both carriageways will be used as a better representation of the hilliness of each segment. A similar argument applies to crossfall and curvature values.

In the case of SCRIM coefficient, the Design Manual for Roads and Bridges (DMRB[14]) provides guidelines on investigatory levels for SCRIM coefficients based on road type and traffic levels. Therefore, an increased risk of skidding will be defined as all values below 0.35 for motorways, 0.4 for dual carriageways and 0.45 for single carriageways with heavy traffic. The variable used for this study will be estimated by calculating the number of 10m sections within the homogenous road segment with SCRIM value below the acceptable threshold and converting it into a 0 to 1 value. This value can be interpreted as the proportion of road segment with SCRIM coefficient below the threshold.

Safety barrier is established from the VRS dataset which includes lines on the map where they are present: either side of the road and/or the centre line. Therefore, two variables will be included in the model: the percentage of the homogenous road segment with safety barrier present on both sides of the carriageway and the percentage of road segment with a safety barrier present on centre line[15]. For instance, a single carriageway with a safety barrier present on one side but not the other will be valued as 50%.

Access density will be estimated as a proxy from the number of buildings within a certain distance (to be determined) of each segment from the GeoDirectory Data.

5.2.4 Development of crash modification factors

The output from the APMs will be used to develop CMFs (crash modification factors) to be used by road safety practitioners to evaluate the potential impact on collision numbers of installing particular road safety interventions on the national road network. Based on the variables to be tested for inclusion in the models, this section presents examples of some of the potential interventions that could be evaluated following the development of the APMs and subsequent CMFs. These have been summarised in Table 13.

[14] https://www.standardsforhighways.co.uk/prod/attachments/50d43081-9726-41e8-9835-9cd55760ad9e?inline=true

[15] A small proportion of the network (71km, approximately 1% of length) is a 2+1 road. These are all classified as single carriageways under the road type definition and thus will be captured within the network for that model. Particular consideration will be given to how these are defined and included in the model, and whether the particular safety benefits these roads provide can be understood using the GLMs produced.

Table 13: Potential interventions linked with the variables in the APM

	Variable	Potential Intervention(s)
Base model	AADT	N/A
	Road segment length	N/A
	Number of lanes	Adding additional lanes (for example, conversion to a 2+1 road or adding climbing lanes) could have an impact on collision numbers as it influences congestion, overtaking and reduces the number of head-on collisions (iRAP, 2022). However, it could have a converse impact on VRU casualties as it increases the distance a VRU would need to cover if crossing the road.
Speed	Modelled speed	Interventions associated with changing the speed limit or applying stricter enforcement (e.g. average speed cameras).
Road geometry and condition	Gradient	Changes such as reducing the gradient, increasing the radius of a crest or minimizing vertical acceleration changes could result in reduced risk of head-on or overtaking collisions[16].
	Crossfall	N/A
	Radius (curvature)	Interventions including increasing bend radius, providing transition bends, removing compound bends or providing better warning signs could reduce the risk of head-on or run-off-road collisions[17].
	SCRIM value	Interventions associated with skid resistance such as resurfacing the road or using road warning signs to indicate slippery roads could be applied.
	Minor/Major junction	Closing minor junctions to reduce the conflict points
Roadside features	Safety barrier: location (carriageway sides and centreline)	Installing safety barriers where they are not already present can reduce run off the road crashes or head on crashes (in the case of a centreline barrier – this is a key feature of the 2+1 road).
	Safety barrier: material	Changes to the safety barrier material. However, there are pros and cons associated with different materials: for example, concrete barriers could require less maintenance but could cause more severe outcomes in the event of a collision.
	Access density	Replacing multiple access points with a single point of access could reduce the number of potential conflict points[18].

[16] https://toolkit.irap.org/safer-road-treatments/realignment-vertical/?id=24

[17] https://toolkit.irap.org/safer-road-treatments/realignment-horizontal/?id=23

[18] https://toolkit.irap.org/safer-road-treatments/restrict-combine-direct-access-points/?id=26

	Variable	Potential Intervention(s)
	Urban or rural	N/A
Other variables that impact collisions	Proportion of HGV traffic	Variable included to improve model fit. However, it might link to interventions around barrier type (roads with higher HGV traffic may need barriers with higher containment).
	Proportion of rear-end collisions	These collision types are more likely on roads with high traffic or junctions. Interventions to reduce rear end collisions could include improved in-vehicle technology or queue protection systems which reduce speed limits in responses to slow moving traffic, queues or congestion.

In addition to the interventions outlined above, it may also be possible to use the models to understand the effect of upgrading carriageways (e.g. converting a undivided single carriageway with higher AADT to a dual carriageway). The variables in each of the final road type models may vary, but they should enable practitioners to gain a qualitative understanding of the change safety if the road were upgraded.

While Table 13 outlines some of the interventions that it might be possible to evaluate following development of the APMs (the full list for each model will depend on the variables included in the final model), due to data availability there are a number of interventions that it will not be possible to evaluate using these APMs:

- Design of junctions: The model will include junction density as a variable but cannot capture changes in collisions due to upgrading a junction or changing the design of a junction.
- Vulnerable Road Users: Due to lack of exposure data for VRUs, it has not been possible to include specific variables which relate to the risk of these road users in the model; as a result, interventions relating to safety of VRUs cannot be evaluated from this model.
- Hard shoulders: Complete data on the presence or absence of a hard shoulder is not available for any of the road types (data is most complete for motorways and some is available for dual carriageways); thus, this variable cannot be included in the models. As a result, it will not be possible to evaluate the impact of installation of a hard shoulder or hard strip where these are not already present.
- Actual speed: The model will include modelled speed (derived from link type, speed limit and traffic flow) but there is no information on the actual distribution of speeds on the roads. Therefore, there will be a limit on the robustness of the evaluation of interventions that could be applied.
- Road surface conditions, lining and signing and lighting conditions: Interventions associated with these conditions cannot be developed using the proposed APM. These conditions generally vary over time, and the road features/geometry data available has been taken as a snapshot in time rather than measuring changes over time.

- Traffic calming measures: Interventions relating to local traffic calming measures such as installing high-friction surfacing on bends may not be possible to evaluate using the APM due to lack of information around the current measures on each road segment.

5.3 Recommended procedure for APM development

This section discusses the steps that will be followed to develop the APMs for each road type. It must be noted that due to inherent differences between the four road types being modelled, the process will be repeated four times as there might be differences in the distribution of the explanatory variables being included. In other words, if there is limited variability in one of the explanatory variables for a particular road type (e.g. all legacy roads are located in rural areas) then it might not be included in the list of variables tested for inclusion in the APM for that road type.

5.3.1 Exploratory analysis (prior to model development)

Prior to developing the statistical model, all explanatory variables being included will be checked for multicollinearity (correlation between explanatory variables). Having two or more highly correlated variables in the same statistical model can cause misleading model result, unstable model coefficients and wider confidence intervals. Multicollinearity will be checked by estimating the Pearson correlation coefficient between each pair of variables. The Pearson correlation coefficient value ranges between -1 and 1 where 0 indicates no linear relationship, -1 indicates a strong negative correlation and 1 indicates strong positive correlation. Generally, a value between 0 and 0.3 indicates a weak relationship, 0.3 and 0.5 is a moderate correlation and greater than 0.5 is strong relationship. Exploratory variables that show a strong correlation will be taken into consideration during model development and only one of those variables will be included.

The relationship between the collision data and each exploratory variable will be explored visually using appropriate graphs (such as scatter plots). This visual representation will capture any general trends, variability in the exploratory variables and give an indication of whether non-linear variable forms (such as power or exponential form) would be most appropriate.

5.3.2 Model development

5.3.2.1 Base model

Generalised Linear Models (GLMs) with the appropriate distributions (Poisson or Negative Binomial) will be used to develop the APMs. The base model will be developed using the following variables: AADT, road segment length and number of lanes as these are the key variables identified from the literature review that affect collision risk on all road types.

It is important that the statistical model yields logical results, i.e., if AADT is zero then the number of collisions on that segment should also be zero. Therefore, a number of functional forms will be tested (such as power or exponential form). The literature review found that the power form was most commonly used for AADT and either power or exponential forms were used for segment lengths. Therefore, we propose testing both functional forms and using various model evaluation techniques to assess the model fit; the most appropriate choice will depend on the distribution of the Irish dataset. The steps outlined to evaluate model fit are discussed in detail in Section 5.3.3.

It is also necessary to determine whether the number of collisions follows a Poisson or Negative Binomial distribution. The literature review found that most recent studies assume that the collision data are over-dispersed, and the mean is not equal to the variance; a Negative Binomial is a better fit compared to Poisson in these instances. The likelihood ratio test will be used to determine which distribution betters fits the Irish collision dataset. If the p-value from the test is statistically significant, then we can conclude that the Negative Binomial distribution offers a better fit to this dataset.

In order to determine if a zero-inflated Poisson or Negative Binomial distribution is required, a Vuong test will be applied to the data. If the test statistic is significant, then a zero-inflated model will be applied to the dataset.

5.3.2.2 Adding explanatory variables

After the base model has been developed, the next step will be to evaluate the model fit with each of the additional road geometry and features variables outlined in Table 12 added to the model. The variables will be added one by one using the forward selection technique[19] in order to avoid any overfitting. This process was followed by some of the papers in the literature review.

5.3.3 Variable and model fit

5.3.3.1 Variable selection

Variable selection is a crucial process in developing an accurate model. If too many variables are included in the model, the model could be over-fitted and perform poorly when used for predictive purposes. Therefore, a number of statistical tests and techniques will be used to assess variable fit:

- P-value: As each variable is added to the model, the p-value from the test statistic will be used to determine if the variable has a significant impact on the response variable (number of collisions). Variables that are statistically significant will be considered for inclusion in the final model.
- CURE plots: Cumulative Residual (CURE) plots will also be used to examine the goodness-of-fit of each variable. The residuals[20] are estimated using each variable and generally models whose CURE plots are within 2 x standard deviation limits are considered to be unbiased and a good fit.

[19] This technique adds variables one by one to the model and at each step of the model development, the variable that offers the single best improvement to the model is retained. The process is then repeated until no further variables are deemed to improve the model fit.

[20] The difference between the observed value and the value predicted by the model.

These techniques will be applied to test each variable, including the functional forms of the variables in the base model, and the explanatory variables.

5.3.3.2 Model goodness-of-fit

The overall goodness-of-fit of the model can be assessed using a number of measurements:

- Adjusted R-squared value: The R-squared value is a statistical measure that explains the amount of variation in the response variable that is explained by the independent (or explanatory) variables. The adjusted R-squared value is a variation of this measure which accounts for the number of variables in the model. Generally, the value increases when a new explanatory variable is added to the model that improves model fit. This will be used to assess general model fit.
- AIC and BIC: Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be used to assess model fit by comparing different models. Both measures use the log-likelihood of the model and generally a lower value indicates better model fit. The literature review identified that AIC was most commonly used in papers, however, AIC is known to penalise complex models less than the BIC. This means that AIC is more likely to pick complex models whereas BIC is likes likely to do so. Therefore, we propose using both AIC and BIC to compare models, especially as the model complexity increases, to ensure that the most appropriate model is selected.
- CURE plots: CURE plots can also be used to assess general model fit in addition to variable selection. The process followed is the same as for variable selection.
- Likelihood ratio test: This measure compares the proposed model to the more complex model. The deviance value shows if the more complex model is significantly better at capturing data than the simpler model. If the resulting p-value is significant, then the complex model is preferred.

TRL will use all the goodness-of-fit measures above to evaluate both variable and model fit and select the best fitted model as the final APM.

5.3.4 Model prediction

5.3.4.1 Cross-Validation

After the final model has been developed, the predictive performance of the model needs to be assessed to understand how well the model performs when used on a new dataset. Although not discussed in detail in any of the papers in the literature review, generally, it is advised to split the data into a 'train' and 'test' set, build the model on the train set and check the predictive performance on the test set. This process ensures that the predictive performance of the model is tested on a dataset that has not been seen by the model before.

We propose using re-sampling techniques such as K-fold cross validation where K is the number of sets (standard practice is 10) the data is divided into. One of the sets is randomly sampled as the test set and the remaining K-1 sets are combined to form the train set on which the model is built. This process is followed K times to obtain K values of the evaluation metric which can then be summarised. Using cross-validation has multiple advantages as it reduces biases in results, computation time and provides much more information about model performance.

5.3.4.2 Model evaluation metrics

The literature review identified a small number of papers that looked at predictive performance of their APMs. Two evaluation metrics were used in the literature: Mean Absolute Deviance (MAD) and Mean Squared Prediction Error (MSPE). MAD is estimated by subtracting the actual collision values from the predicted values, converting it to an absolute error and calculating the average. While this metric is the easiest to explain, the main drawback of this measure is that it averages out the error across the entire database which does not give an accurate result. Therefore, MAD is less accurate for outliers but better for 'normal' observations. The main advantage of the second metric MPSE is that it is more sensitive to large outliers compared to MAD; however, it might be less accurate for 'normal observations'. Generally lower error values indicate better model fit.

TRL propose using both metrics to evaluate the APMs as each metric offers information that complements the other.

5.4 Limitations and risks with proposed approach

Developing an accurate Accident Prediction Model (APM) is a challenging task. Although the literature review identified a number of APMs, there was limited information on model validation and prediction accuracy. TRL propose using similar statistical models (GLMs) to those that were used to build APMs in other countries; however, there is a possibility that the APMs developed might not predict collisions as accurately as expected.

Almost all the papers identified in the literature developed APMs for very specific road types or junctions. None of the papers developed an APM for the entire road network, as was the stated aim for this project. TRL has considered the challenges this poses and propose developing four APMs split by road type, this means that some of the roads and junctions will not be modelled in this study (namely roundabouts, link roads and ramps). Section 5.4.1 below gives details of all of the proposed exclusions from the model.

It is possible that some of the variables of interest to TII are not statistically significant in the final model. TRL have proposed discussing the possibility of developing so-called 'practitioners' models' (the variables included in these models might not necessarily achieve statistical significance but are of practical importance to practitioners). However, developing these models will have additional cost implications.

The potential interventions that could be evaluated based on the variables considered for inclusion in the APM have been highlighted in Section 5.2.4. This section also covers the types of interventions that it will not be possible to evaluate due to the nature of the data being used to develop these models. It is possible that TII have specific interventions in mind for the APMs and must consider the limitations outlined within that section. Alternatively, TRL could investigate the implementations of APMs calibrated using CMFs derived from other sources; the practicalities and limitations of this approach would need to be considered further before this approach is adopted.

5.4.1 *Exclusions from the models*

The following will be excluded from the analysis:

1. Roundabouts, link roads and ramps: Combined, these account for less than 5% of the road network length and have less than 10% of all collisions. Due to the small sample size, any collisions on roundabouts, link roads or ramps will be excluded from the modelling.
2. Junctions (major and minor junctions): Around 65% of collisions did not occur at a junction. The remaining junction types individually account for less than 10% of the collisions (see Figure 17). Aside from junction type, there is also relatively little data available to robustly model junctions (e.g. there is no data on pedestrian flows or the geometry of the junctions themselves. As a result, it is not possible to model junctions separately from the mainline models proposed, and it may not be possible to robustly understand the impact of junction type in the models. Two approaches will be trialled: including a count of the number of major and minor junctions (or junction density) on each segment as an exploratory variable, or including variables for 'count of T-junctions', 'count of crossroads' instead. Due to issues with multicollinearity, it is unlikely both options will be included in any given model. The most appropriate approach will be determined for each model separately, based on the variability observed in the data once segments have been created.
3. Missing data: Segments with missing information on a large number of exploratory variables will be excluded from the model, as this could reduce the accuracy of model prediction.
4. Outliers: If particular road segments are identified as outliers (e.g. have significantly higher traffic than all other segments of that type), then these may need to be excluded from the modelling, or incorporated through a separate variable, to avoid them influencing model fit. For example, the client has identified road segments around the Dublin ring road are often very different from other segments of this type.

5.4.2 *Variables to be considered in future model iterations*

In addition to the exclusions above, there are a number of variables which are not presently available for the modelling but could be beneficial to consider in future iterations of these models.

For example, additional variables that account for socio-economic factors and weather effects may improve the model fit. One way of doing so would be by following the methodology developed by Turner, Singh, & Nates (2012) and discussed in Section 3.1.4 of the literature review: variables such as 85th percentile speed, proportion of collisions in wet weather, under-reporting of collisions and proportion of alcohol-related collisions were compared by region and similar areas combined into bigger regions. These regions were included as an explanatory variable in the model. The main advantage of this is that it combines the effect of a number of variables and potentially reduces any issues of multicollinearity[21]. However, due to lack of available data this will not be included in the APMs proposed here but should be considered in future iterations of the model.

The literature review also identified roadside hazard ratings to be a significant variable in many APMs. These ratings can be estimated from road safety inspection data. TII are presently collecting these data, but it will not be available in time for the development of the APMs proposed here.

[21] Multicollinearity is an issue in statistically models where multiple explanatory variables are highly correlated to each other. This results in less reliable statistical inferences as the modelling will be unable to assign variance clearly to specific variables, all variables included in the model should therefore have low correlation (be independent of each other).

References

Ambros, J., & Sedonik, J. (2016). A Feasibility Study for Developing a Transferable Accident Prediction Model for Czech Regions. *Transport Research Procedia*.

Ambros, J., Havranek, P., Valentova, V., Krivankova, Z., & Streigler, R. (2016). Identification of Hazardous Locations in Regional Road Network – Comparison of Reactive and Proactive Approaches. *Transportation Research Procedia*.

Cafiso, S., & D'Agostino, C. (2012). Safety Performance Function for Motorways using Generalized Estimation Equations. *Procedia - Social and Behavioural Sciences*.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., & Persaud, B. (2010). Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis & Prevention*.

CEDR. (2013). *PRACT- Predicting Road Accidents- a Transferable methodology across Europe.*

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2011). Extended prediction models for crashes at roundabouts. *Safety Science*.

Elvik, R., Høye, A., Vaa, T., & & Sørensen, M. (2009). *The Handbook of Road Safety Measures.* Oslo: The Emerald Group Publishing Limited.

EUR-Lex. (2019, 11 26). *DIRECTIVE (EU) 2019/1936 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 October 2019 amending Directive 2008/96/EC on road infrastructure safety management*. Official Journal of the European Union. doi:http://data.europa.eu/eli/dir/2008/96/2019-12-16

Garach, L., de Ona, J., Lopez, G., & Baena, L. (2016). Development of safety performance functions for Spanish two-lane rural highways on flat terrain. *Accident Analysis & Prevention*.

Geedipaly, S. R., Lord, D., & Dhalavala, S. S. (2012). The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis & Prevention*.

Gross, F., Persaud, B., & Lyon, C. (2010). *A guide to developing quality Crash Modification Factors.*

Hauer, E. (2007). *Observational Before–After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety.* UK: Emerald.

iRAP. (2022). *Overtaking Lane and 2+1 Road*. Retrieved from IRAP Road Safety Toolkit: https://toolkit.irap.org/safer-road-treatments/overtaking-lane-and-2-plus-1-road/

La Torre, F., Domenichini, L., Meocci, M., Graham, D., Karathodorou, N., Richter, T., . . . Laiou, A. (2016). Development of a Transnational Accident Prediction Model. *Transport Research Procedia*.

Labi, S. (2011). Efficacies of roadway safety improvements across functional subclasses of rural two-lane highways. *Journal of Safety Science*.

Maher, M. J., & Summersgill, I. (1996). A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention*.

OECD. (2012). *Sharing Road Safety: Developing an International Framework for Crash Modification Functions.* Paris: OECD Publishing. doi:https://doi.org/10.1787/9789282103760-en.

Pei, X., Sze, N. N., Wong, S. C., & Yao, D. (2016). Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong. *Accident Analysis & Prevention*.

Pickering, D., Hall, R. D., & Grimmer, M. (1986). *Accidents at rural T-junctions.* Crowthorne: TRL.

Summersgill, I. (2000). *The availability of accident predictive models for inter-urban roads.* Crowthorne: TRL.

Summersgill, I., & Layfield, R. E. (1996). *Non-junction accidents on urban single-carriageway roads.* Crowthorne: TRL.

Taylor, M. C., Baruya, B., & Kennedy, J. V. (2002). *The relationship between speed and accidents on rural single-carriageway roads.* Crowthorne: TRL.

Turner, S., Singh, R., & Nates, G. (2012). *The next generation of rural road crash prediction models: final report.* Wellington: NZ Transport Agency.

Vieira Gomes, S., Geedipally, S., & Lord, D. (2012). Estimating the safety performance of urban intersections in Lisbon, Portugal. *Safety Science*.

Vogt, A., & Bared, J. G. (1998). *Accident models for two-lane rural roads: Segments and intersections.* Federal Highway Agency.

Walmsley, D. A., & Summersgill, I. (1998). *The relationship between road layout and accidents on modern rural trunk roads.* Crowthorne: TRL.

Walmsley, D. A., Summersgill, I., & Binch, C. (1998). *Accidents on modern rural single-carriageway trunk roads.* Crowthorne: TRL.

Walmsley, D. A., Summersgill, I., & Payne, A. (1998). *Accidents on modern rural dual-carriageway trunk roads.* Crowthorne: TRL.

Walmsley, D. A., Summersgill, I., & Payne, A. (1998a). *Accidents on modern rural dual-carriageway trunk roads. TRL report 335.* Crowthorne: TRL.

Yannis, G., Dragomanovits, A., Laiou, A., Richter, T., Ruhl, S., La Torre, F., . . . Li, H. (2016). Use of accident prediction models in road safety management – an international inquiry. *Transport Research Procedia*, 4257-4266.

Appendix A Acronyms

AADT	Annual Average Daily Traffic
AIC	Akaike’s Information Criterion
APM	Accident Prediction Model
BIC	Bayesian Information Criterion
CCR	Curvature Change Rate
CMF	Crash Modification Factors
CRF	Crash Reduction Factor
CURE plots	Cumulative Residual plots
DIC	Deviance Information Criterion
EB	Empirical Bayes
GEE	Generalised Estimating Equations
GIS	Geographic Information System
GLM	Generalised Linear Modelling
GPS	Global Positioning System
HGV	Heavy Goods Vehicle
LCMS	Laser Crack Measurement System
MAD	Mean Absolute Deviance
MMaRC	Motorways Maintenance and Renewals Contract
MSPE	Mean Squared Prediction Error
NTpM	National transport model
OSi	Ordnance Survey Ireland
PCA	Principal Component Analysis
PMS	Pavement Management Survey
PRACT	Predicting Road Accidents – a Transferable methodology across Europe
RSH	Roadside hazard rating
RSI	Road Safety Inspections
RSP	Road Surface Profiler
SCRIM	Sideway-force Coefficient Routine Investigation Machine (Skid resistance)
SPFs	Safety Performance Functions
TII	Transport Ireland Infrastructure
TRID	Transport Research International Documentation
TRL	Transport Research Laboratory
VRS	Vehicle Restraint Systems

Appendix B Literature review methodology

A range of approaches were applied to identify the papers and reports that have been reviewed for this task:

1. A systematic search of four academic sources of published papers and available reports was completed using a set of developed key words (see B.1.1 for more detail). This approach was applied to materials published in the last ten years in order to identify the most recent developments and applications of APMs.
2. The team identified a number of older but essential papers and reports, these included some of those listed in the client's scope for this project.
3. A search of the TRL archives of both published and unpublished Client Projects Reports (CPRs) for papers that contained methodology and results relevant to this project's aims. Specifically, those related to a significant programme of APM development for the Highways Agency (now National Highways) from the mid-1980s up until the early 2000s[22].

B.1.1 Methodology for the systematic search

A string of search terms was developed to identify literature relevant to these aims. These terms were applied to four online literature databases or APM model repositories, these being:

- Google Scholar
- Transport Research International Documentation (TRID) database
- Science Direct database
- Predicting Road Accidents – a Transferable methodology across Europe (PRACT) repository[23]

The search terms applied are shown in Table 14.

[22] These papers are considered relevant to the Irish context because the UK road system is comparable to that in Ireland due to the geographical proximity, there is also similar (although not identical) economic and behavioural backgrounds to road use across the different jurisdictions. Some reports generated for the Highways Agency APM programme were not formally published by TRL. This was primarily because their content tended to be highly technical and was not considered to be of general interest to a wide audience.

[23] https://www.pract-repository.eu/

Table 14: Search term for literature review

"crash prediction model" OR "incident prediction model" OR "collision prediction model" OR "accident prediction model”	OR	"crash predictive model" OR "incident predictive model" OR "collision predictive model" OR "accident predictive model" OR "predictive accident model"	AND	road* OR “road geometry”	AND	safety

B.1.2 *Papers identified and reviewed*

108 sources of literature were identified in total out of about 700 initial results, from the formal search and the other approaches. This initial long list of potentially relevant papers contained a mixture of academic research, studies and reports by various organisations in the UK and other countries. The papers were then assessed based on relevance and quality and 29 papers/books were included and referenced in the final literature review.

A large number of papers were excluded from the main review as they did not answer the key research questions identified for Task 1. Some papers did not contain detailed information on the variables used for APM development or details about the model itself. A large number of papers developed simpler SPFs (using only traffic and road segment information) rather than more complex APMs (with information on road geometric features). Furthermore, a number of papers discussed theoretical approaches around developing and implementing APMs rather than the practical methods of model development.

Collision Prediction Model for the Irish National Road Network

This report presents the results of Phase 1 of a two-phase project to develop Accident Prediction Models for Transport Ireland Infrastructure. The aim of these models is to assist engineers to better manage the safety of physical road features across its trunk network. Phase 1 reviewed the statistical approaches used by others to develop APMs, reviewed the data available in Ireland to develop these models and made recommendations on how these models could be developed and applied. Phase 2 developed the models and associated practitioners' tools for Ireland.

Other titles from this subject area

PPR2031	TII268 Lot1 Collision Prediction Model for the Irish National Road Network – Phase 2 Report. C Wallbank, N Harpham & J Fletcher. 2023

TRL
Crowthorne House, Nine Mile Ride,
Wokingham, Berkshire, RG40 3GA,
United Kingdom
T: +44 (0) 1344 773131
F: +44 (0) 1344 770356
E: enquiries@trl.co.uk
W: www.trl.co.uk

DOI: 10.58446/vpva4051

PPR2030