Modelling casualties

March 22, 2018

Abstract

We present a method to generate a predictive model for numbers of fatalities and injuries resulting from road-traffic collisions, itemised by the specifics of the collision. We apply the methodology to data for England spanning the years 2005 to 2015. Our unit of risk is: number of casualties (injuries or fatalities), per km travelled by the casualty, per km travelled by the striker, per year. There is one rate per road type, per casualty severity, per casualty demographic group, per striker demographic group, per casualty travel mode, per striker travel mode, per year. We use three different road types, three levels of casualty severity, and seven travel modes. Our demographic groups consist of age and gender labels.

1 Introduction

Death and disability due to road-traffic crashes ... . We predict injury and fatality rates as number per distance travelled, as functions of the specifics of the crash. We are interested in the role of the striking-vehicle covariates, as well as the pattern of injury rates as they pertain to the casualties of crashes (**??**).

Our model of road-traffic injuries requires synthesis of data from many sources. The primary source is the Stats19 database. These are records of all road-traffic injuries reported to the police, where each entry (row) corresponds to one injury or fatality, which we refer to in general as "casualties". We use this source to count the total number of events of each type.

The records are grouped into "types" according to the specifications of the crash and resulting casualties. We refer to these specifics as predictors or covariates. We model casualty counts as functions of these predictors. The predictors we use are: casualty mode, casualty demographic group (casualty age and casualty gender), casualty severity, strike mode, striker demographic group (striker age and striker gender), road type, and year. The "striker" is defined as the driver of the largest other vehicle in the collision.

Do we, in parallel, present a model for no-other-vehicle crashes, which would include pedestrian falls from a different source, and has as predictors only casualty mode, casualty demographic group (casualty age and casualty gender), casualty severity, road type, and year? (**??**)

The casualty counts are to be understood in terms of exposure: what amount of exposure to risk led to how many casualties. We measure exposure in terms of distance travelled. We must therefore estimate the distance travelled by each demographic subgroup by each mode on each road type in each year. These data are not readily available, and must be constructed from numerous different sources.

In order to know both the number of people on the roads and the number of vehicles on the road, we must have a sense of vehicle occupancy. We achieve this by further subdividing the group travel distances as passengers or non-passengers.

The set of predictors, which form the indices of our data, is shown in Table 1. We describe the number of casualties, $I_{a_{\text{cas}},g_{\text{cas}},m_{\text{cas}},a_{\text{str}},g_{\text{str}},m_{\text{str}},c,t,y}$, as a smooth function of the predictors and the distance travelled per casualty, $A^{(\text{cas})}_{a,g,m,t,y}$, and per striker, $A^{(\text{str})}_{a,g,m,t,y}$. We use a regression model to relate the covariates to the casualty counts in a smooth manner that allows the sharing of support across the (sometimes sparse) counts of casualties.

The distances travelled ($A^{(\text{cas})}_{a,g,m,t,y}$ and $A^{(\text{str})}_{a,g,m,t,y}$) are unknown. $A_{a,g,m,t,y}$ denotes the total distance travelled by people in demographic group $\{a, g\}$, by mode m, on road type t, in year y. $A^{(\text{cas})}_{a,g,m,t,y}$ and $A^{(\text{str})}_{a,g,m,t,y}$ might differ (a) in terms of having different age-group categories a, and (b) due to the inclusion of passengers in the former but not the latter. We construct them from various data sources in a process described in Section 2.

A future development to our methodology would be to include a spatial component, as in **?**.

2 Exposure: distance travelled

To estimate distance travelled, we use a number of data sources, as no single one reports distances at the itemised level we require. This Section describes how we use the various sources.

Table 1: Indices used to denote model variables.

Index	Name	Values
a	Age groups	0–105
c	Casualty severity	Fatal
		Serious
		Slight
g	Gender	Female, male
m	Transport modes	Pedestrian, cyclist, car/taxi, motorbike, bus, HGV, LGV
r	Role	Casualty (cas)
		Striker (str)
s	Scenarios	(User specified)
t	Road type	Motorway
		A
		B, C, unclassified
y	Year	2005–2015
z	Passenger	No (0)
		Yes (1)

We assume traffic count data to provide an accurate representation of the total distance travelled by different modes on different road types (RTS data, Section 2.1). We use it to scale up survey data, which forms a baseline estimate for total travel on any road type, for most but not all of the categories (NTS data, Section 2.2). We allocate distances to three road types using a heuristic method so that in total the distance travelled on a road is a fair approximation to the count data (Section 2.2.2). For the subset of groups for which there are no relevant data from the survey we use license data (Section 2.3). Finally, we smooth the resulting estimates (Section 2.4.1).

We use the datasets, listed in Table 2, to construct the distance-travelled arrays (A) via a complete array, B, that includes all covariates: $B = B_{a,g,m,t,y,z}$. Then we define the casualty exposure as

$$A^{(\text{cas})}_{a,g,m,t,y} = \sum_{z=0,1} B_{a,g,m,t,y,z}$$

and the striker exposure as

$$A^{(\text{str})}_{a,g,m,t,y} = B_{a,g,m,t,y,z=0}.$$

Table 2: Datasets used to calculate $B_{a,g,m,t,y,z}$

Label	Description	Source	Years	Geographical region
$R_{m,t,y}$	Total miles driven on different road types by vehicle type	RTS	2005–2015	England
D_j	Distance travelled on trip j in NTS database	NTS	2005–2015	England
$L_{a,g,m}$	License-holder data from the DVLA[1]	DVLA	2015	GB

2.1 Vehicle count data

The backbone of our distance data is data from road traffic statistics (RTS) pertaining to road usage. RTS estimates the total distance travelled by each mode on each road type in each year using automatic traffic counters and manual counts. We denote this dataset $R_{m,t,y}$.

[1]Data for other years are missing and/or improbable. We use license type D (full) and D (auto) for bus drivers and C1E and C1 full and auto for HGV: `http://data.dft.gov.uk/driving-licence-data/Driving-Licence-Data-User-Guide%20%288%29%20June%202017%20version.docx`

The $R_{m,t,y}$ dataset provides our first constraint: that the "driver" component of B is constrained to equal the sum of all distance travelled, i.e.

$$\sum_{a,g,t} B_{a,g,m,t,y,z=0} = \sum_{t} R_{m,t,y}. \tag{1}$$

In what follows, when using datasets to populate B, we are subject to the constraint of Equation 1.

2.2 NTS data

Our primary source for estimated distances travelled is the National Travel Survey (NTS). This is an annual survey that includes a week-long travel diary, completed by ˜15,000 participants. It results in a trip-level dataset that forms the basis of our estimate of distance travelled in England.

Each trip in the dataset is taken by a specific individual and by a specified mode, so that the dataset tells us distances travelled categorised by five of the six covariates we require: age, gender, mode, year, and passenger status; i.e., each trip j in the dataset has by definition one of each $a(j)$, $g(j)$, $m(j)$, $y(j)$ and $z(j)$.

Each trip in the dataset might consist of more than one stage, and therefore more than one mode. The dataset lists the main mode alongside any additional distance travelled by bike or on foot. We add these additional distances as journeys in their own right, and subtract the distance from that travelled by the main mode. The total dataset increases from 3,114,437 entries to 3,161,698 entries.

Each person in the dataset has an assigned weight, according to the perceived representativeness of the person based on their demographic description. Each trip also has a weight, that accounts both for the person taking the trip, and the day on which the trip was made.[2]

In order to use these data alongside the Stats19 data, we must (a) distribute each trip made by each individual in the population to the three road types, and (b) fill in the gaps: The NTS data lack information on bus drivers and lorry drivers.[3] "Bus driver" is not a mode available to persons filling in the survey, and we find that lorry driving is under-reported, so we instead populate these data using information from the Driver and Vehicle Licensing Agency (DVLA, Section 2.3).

2.2.1 Matching NTS data to RTS data

We find that the NTS data estimate total travel in England to be less than that predicted by RTS (Figure 1). It is unclear why this should be the case. Assuming RTS data to be more reliable estimates of total distance travelled, and therefore a better indicator of exposure to events that might result in a record in the Stats19 database, we scale NTS trip weights by the following factor:

$$\rho_{m,y} = \frac{\sum_t R_{m,t,y}}{\sum_{\substack{j:m(j)=m,\\ y(j)=y,z(j)=0}} D_j}$$

where $m(j) = m$ and $y(j) = y$ give the mode and year of trip j, and $z(j) = 0$ restricts us to counting only drivers. I.e., for each mode and for each year, we sum the total distance travelled on all road types in the RTS dataset, and sum the distance travelled by all people on all trips from the NTS dataset, and take their ratio as the scaling factor.

[2] See `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/550854/nts-technical-report-2015.pdf` for details of NTS methodology

[3] We assume that there are no lorry passengers

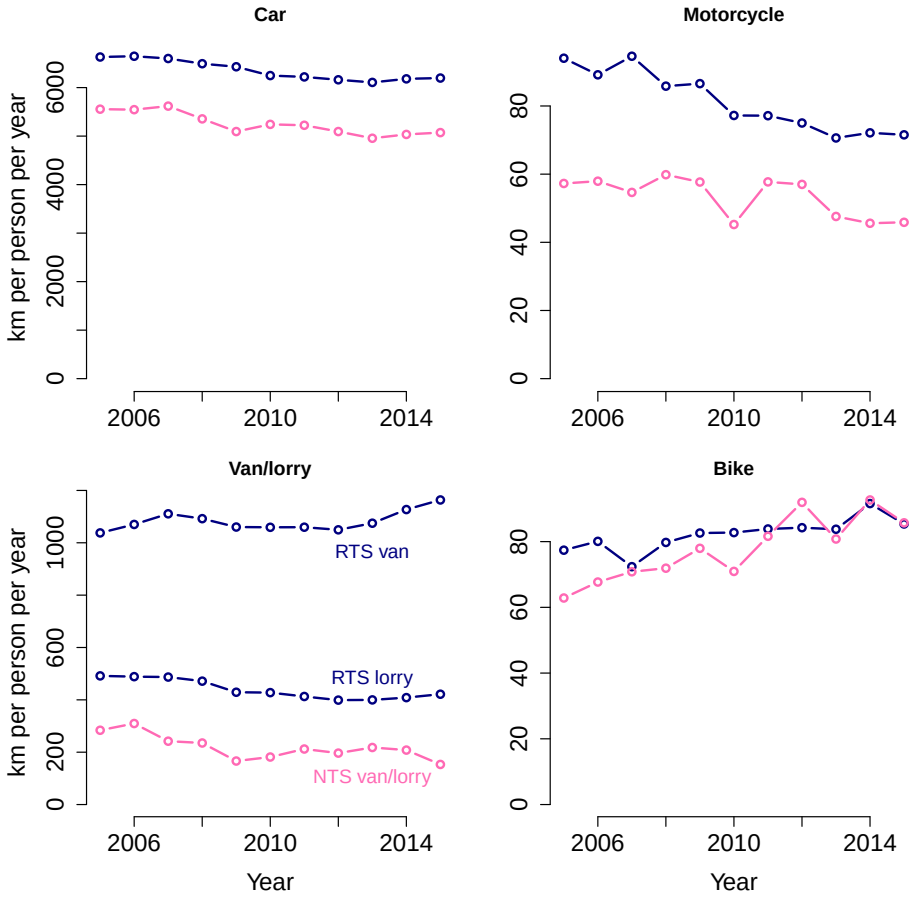

Figure 1: Estimated average distance travelled per person per year over time. For the RTS estimates, in navy, we use the total distance covered in England[4] divided by the total population in England[5]. For the NTS estimates, in pink, we sum the total distance travelled by mode[6], weighted according to the NTS-provided trip weights, and divide by the total weight of the participants of the survey.

Hence, we are assuming that the RTS data are accurate[7] and the NTS data an underestimate; further, we assume that the distances recorded are representative of true distances, and that the person weighting is representative, but the number of trips is not. I.e., we assume that there were more of the same trips, by the same people, that were not reported in the NTS travel diary. Thus we are effectively increasing the trip weight.

With this factor, we have an equivalent constraint to Equation 1, that all the listed journeys should add up to the total distance travelled per mode, per year:

$$\sum_t R_{m,t,y} = \rho_{m,y} \sum_{\substack{j:m(j)=m,\\ y(j)=y,\\ z(j)=0}} D_j, \tag{2}$$

which relates to our target array, B, disaggregated also by age and gender, as follows:

$$\sum_t B_{a,g,m,t,y,z=0} = \rho_{m,y} \sum_{\substack{j:a(j)=a,\\ g(j)=g,m(j)=m,\\ y(j)=y,z(j)=0}} D_j. \tag{3}$$

We assume this to be the true total distance travelled (as a non-passenger) by persons in demographic group $\{a, g\}$ by mode m (excluding bus and lorry) in year y.

[7] `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524848/annual-methodology-note.pdf`. Reasons that RTS might overestimate: a) the growth assumption for minor roads; b) if lots of lorry and van drivers reside in Wales and Scotland but cover a lot of ground in England, without a reciprocal effect.

2.2.2 Distributing NTS data to road types

We now seek some functions $f_{m,t}(D_j)$ that will divide each trip distance D_j, taken by mode $m(j)$, into three distances on the three road types, so that the distances add up to D_j. Then we can estimate the total distance travelled by a demographic group by a mode on a road type by summing all relevant distances as

$$B_{a,g,m,t,y,z=0} = \rho_{m,y} \sum_{\substack{j:a(j)=a,\\ g(j)=g,m(j)=m,\\ y(j)=y,z(j)=0}} f_{m(j),t}\left(D_j\right), \quad \text{subject to} \quad \sum_t f_{m(j),t}(D_j) = D_j.$$

We employ a crude division scheme for $f_{m,t}$. We define thresholds $U_{m,t}$ such that the first $U_{m,t=\mathrm{B}}$ km travelled on a journey by mode m are attributed to road type "B, C, or unclassified", and the next $U_{m,t=\mathrm{A}}$ are attributed to road type "A". Any remaining distance is spent on road type "motorway", with the proviso that the remaining distance exceeds $U_{m,t=\mathrm{M}}$ km. It is otherwise added to the total for "A".

More formally, D_j is decomposed to $D_{j,t}$ as follows:

$$f_{m(j),t=\mathrm{B}}(D_j) = \min\left(D_j, U_{m(j),t=\mathrm{B}}\right); \tag{4}$$

$$f_{m(j),t=\mathrm{A}}(D_j) = \begin{cases} 0 & \text{if } D_j \leq U_{m(j),t=\mathrm{B}} \\ D_j - U_{m(j),t=\mathrm{B}} & \text{if } U_{m(j),t=\mathrm{B}} < D_j \leq \sum_t U_{m(j),t} \\ U_{m(j),t=\mathrm{A}} & \text{if } D_j > \sum_t U_{m(j),t} \end{cases} \tag{5}$$

$$f_{m(j),t=\mathrm{M}}(D_j) = \begin{cases} 0 & \text{if } D_j \leq \sum_t U_{m(j),t} \\ D_j - \sum_{t\in\{B,A\}} U_{m(j),t} & \text{if } D_j > \sum_t U_{m(j),t} \end{cases} \tag{6}$$

We propose the following heuristic values for $U_{m,t}$ (note that we preclude any cyclist travel on motorways):

$U_{m,t}$	$t = \mathrm{B}$	$t = \mathrm{A}$	$t = \mathrm{M}$
$m =$ bike	9	80	$\max_j(D_j)$
$m =$ car	6.5	40	5
$m =$ motorcycle	12	60	5
$m =$ van	10	50	5

(7)

We use these functions to distribute all trips in the NTS dataset, regardless of passenger status. Pedestrian trips are distributed as cyclist trips, and taxi, minicab and bus trips are distributed as car trips.

Figure 2 shows how $f_{m,t}$ distributes car journeys differently from how they would be distributed were we to apportion each journey according to the same (mode-specific) proportions. We see that, as car journeys taken by males tend to be longer, more of the distance is attributed to motorways, whereas for females, where there are more short journeys, more of the distance travelled is attributed to B, C and unclassified roads.

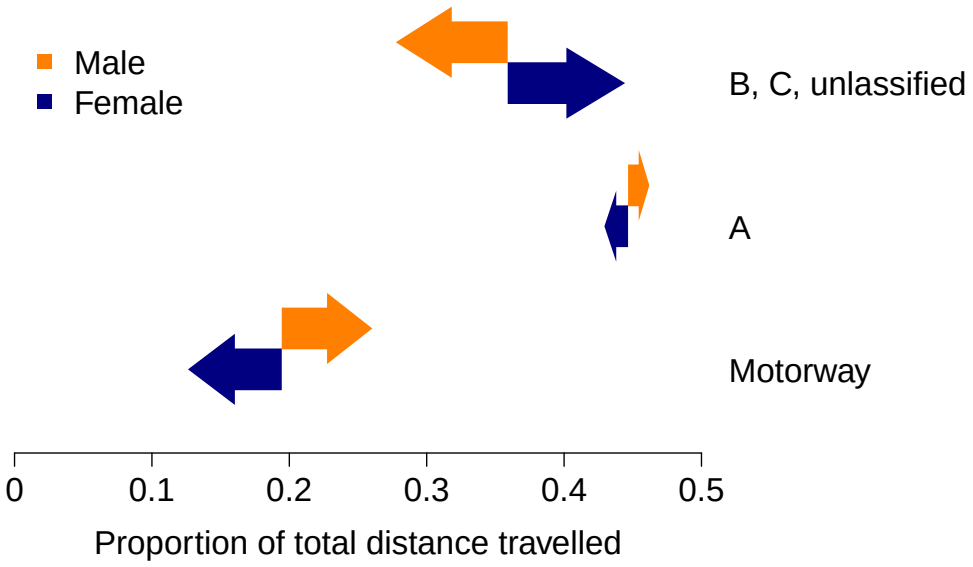

Figure 2: Proportion of total distance travelled by car by male and female 30–39 year olds on each road type: the ratio of car distance on road type m to car distance on all road types. At arrow origins, the proportion is the same as the total car–road proportions according to the RTS. At arrows points are the proportions when road distances are distributed according to the function $f_{m,t}$. Note that the distributions at the origins are the same for male and female, but differ at the points.

2.3 DVLA data

NTS data do not provide useful information on lorry travel (see Figure 1). We therefore assume all NTS "van/lorry" trips to be van trips, and we use a separate data source for lorry travel. We use the same method to estimate bus driver travel, i.e. the total distance travelled by buses (whereas the NTS data provide information on bus passengers).

The RTS-provided estimates for total bus and HGV travel are used directly to understand the total distances travelled by the vehicles. These distances are distributed among the population according to use license-holder data from the DVLA as follows:

$$B_{a,g,m\in\{\text{bus,HGV}\},t,y,z=0} = \frac{L_{a,g,m\in\{\text{bus,HGV}\}}}{\sum_{a,g} L_{a,g,m\in\{\text{bus,HGV}\}}} \cdot R_{m\in\{\text{bus,HGV}\},t,y}. \tag{8}$$

2.4 Resulting annual travel distances & smoothing

Summaries of the resulting annual travel distances are shown in Figures 3 and 4. Note the unstable behaviour of light goods distances in Figure 3 towards 2015; this pattern is not reflected in RTS data. It suggests that a better road-partitioning method is required for light goods vehicles.

Note also the noisiness in motorcycle data in Figure 4. There are numerous entries with a total distance of 0 in this set which we believe to be inaccurate representation of the true amount of travel. For example, if we have $B_{a,g,m,t,y,z} = 0$ for $a = 30$, $g =$ female, $m =$ motorcycle, $t =$ A road, $y = 2010$ and $z =$ driver, but we have $B_{a,g,m,t,y,z} > 0$ for the same $\{a, g, m, t, z\}$ and $y \in \{2008, 2009, 2011, 2012\}$, then we might assume that, in reality, $B_{a,g,m,t,y=2010,z} > 0$.

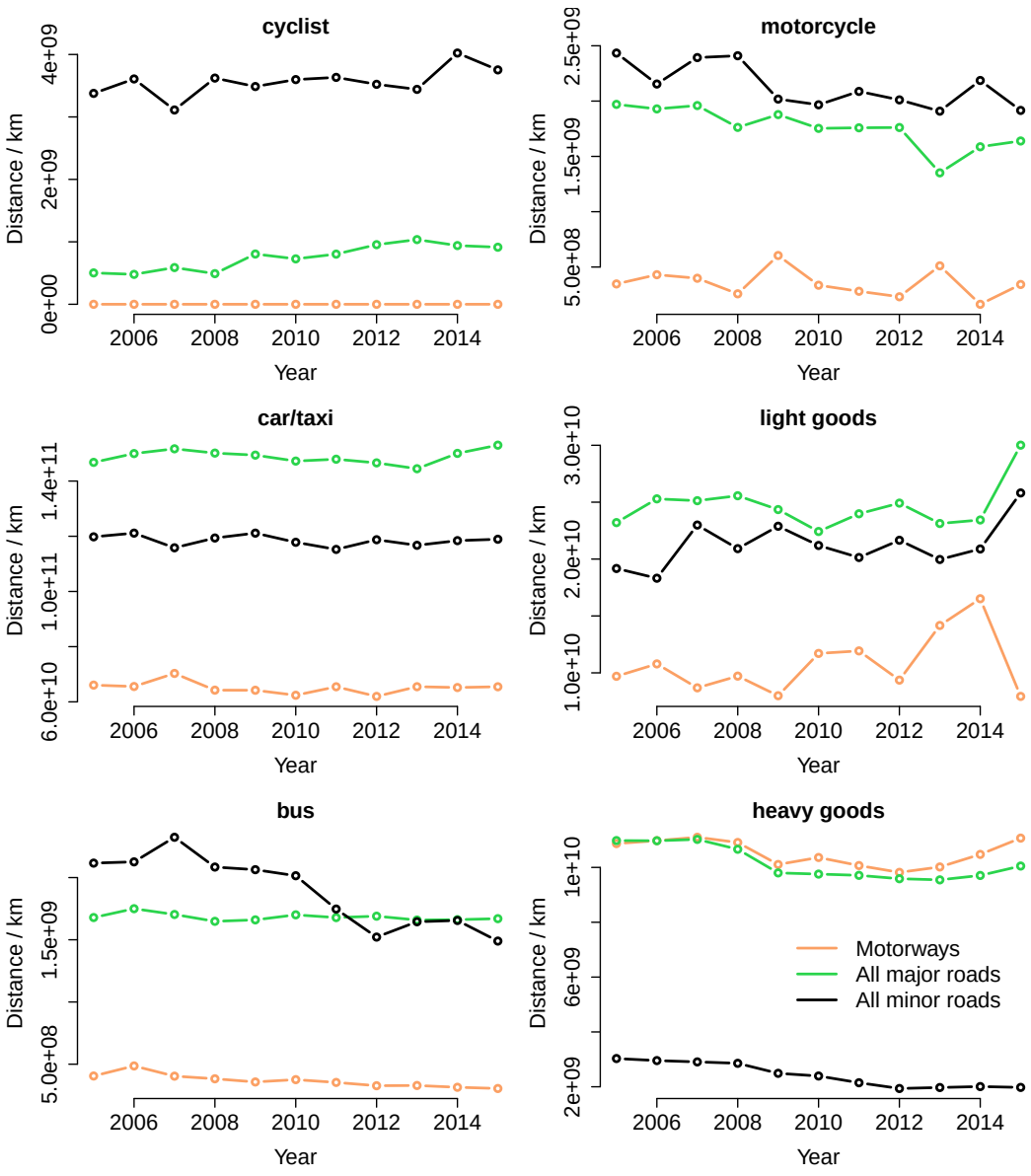

Figure 3: Estimated total distance travelled by mode and road type over time.

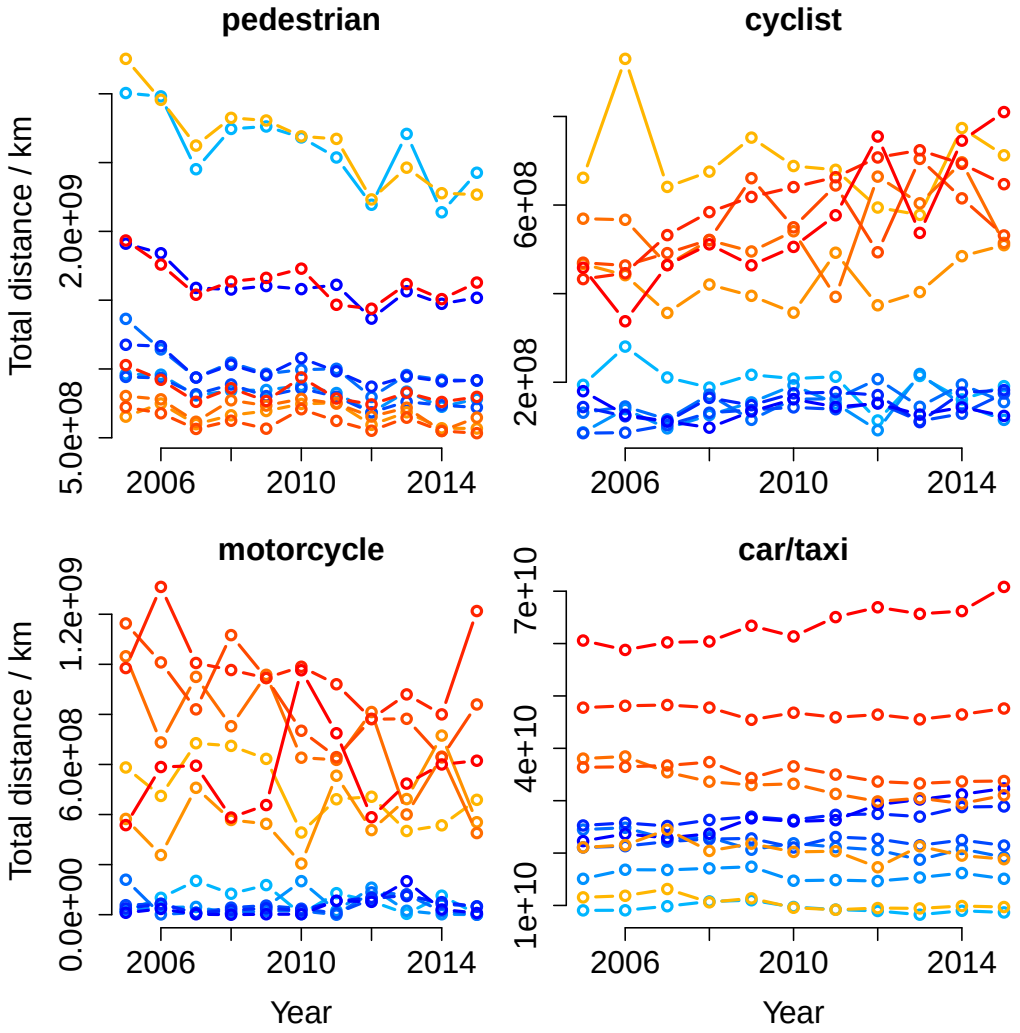

Figure 4: Estimated total distance driven by mode and demographic over time. Colours blue and red indicate female and male groups, respectively. Lighter shades indicate younger age groups. The median ages of the age groups are: 20, 27, 35, 42, 50, and 64.

2.4.1 Smoothing

To correct for errant zero entries, we smooth these data, shown for drivers and passengers in Figures 5 and 6. At the same time we also smooth out noisiness of similar provenance. In doing so we are estimating what we believe to be the true exposure of the population to casualty risk.

We use a logistic model to decide if a distance should be zero or non-zero, expecting, for example, that the distance driven by children is zero. We then apply a log-normal regression to all non-zero values. We then predict distances for all rows for which the probability of being non-zero is greater than 0.1.

The model for the distances, $B_{a,g,m,t,y,z}$, was

$$\log(\tilde{B}_{a,g,m,t,y,z}) \sim \mathcal{N}(\log(\mu_{a,g,m,t,y,z}), \sigma^2) \tag{9}$$

$$\log(\mu_{a,g,m,t,y,z}) = f(m : (\mathcal{S}_1(y) + g \times z + \mathcal{S}_1(y) : \mathcal{S}_8(a) + g : (\mathcal{S}_6(a) + t : z + t : \mathcal{S}_3(a) : z))) \tag{10}$$

where $\mathcal{S}_d$ is a smooth spline with d degrees of freedom.

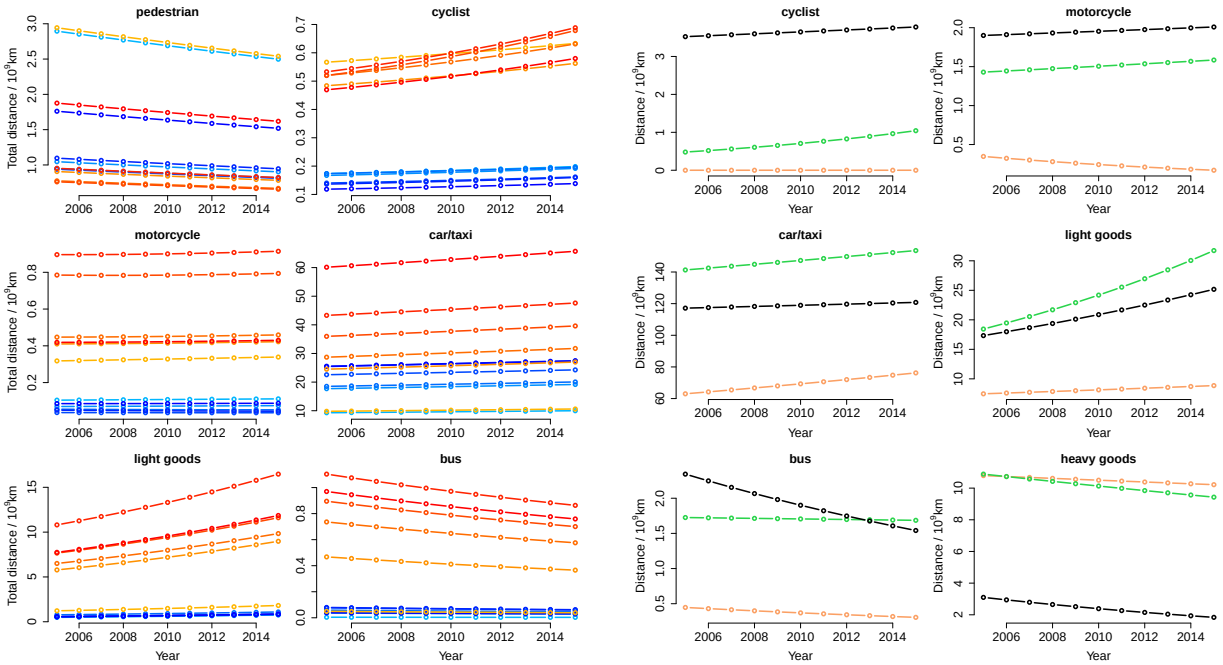

Figure 5: Smoothed data: drivers.

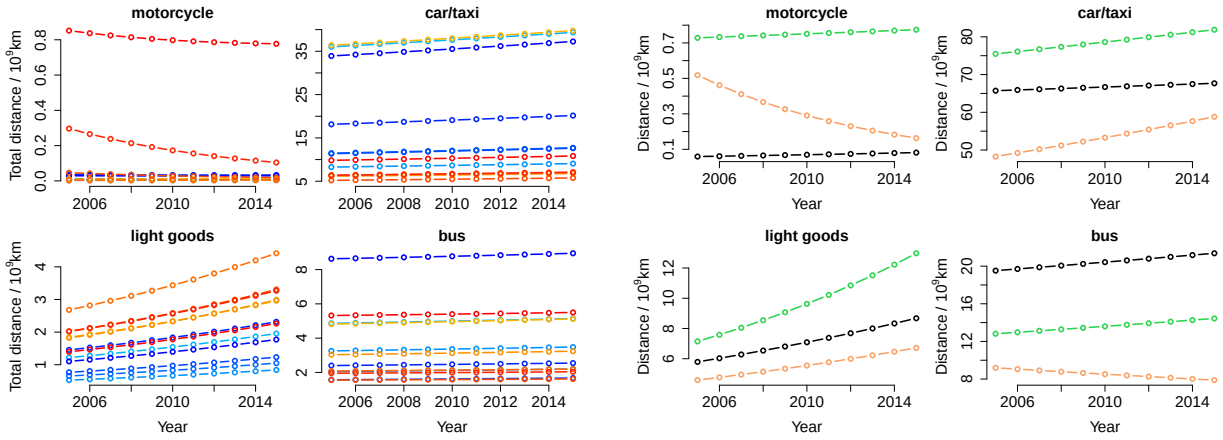

Figure 6: Smoothed data: passengers.

3 Casualties

The Stats19 database lists 2,137,625 road-traffic injuries and fatalities for the years 2005–2015. We process these data in order to group the casualties according to nine predictors: year, striker age, casualty age, road type, strike mode, casualty mode, casualty gender, casualty severity, striker gender. Note that the Stats19 database does not distinguish between casualties and strikers: we define these ourselves in processing the data, choosing the largest party as the striking mode and any other parties as casualty modes. Their marginal distributions are shown in Figure 7.

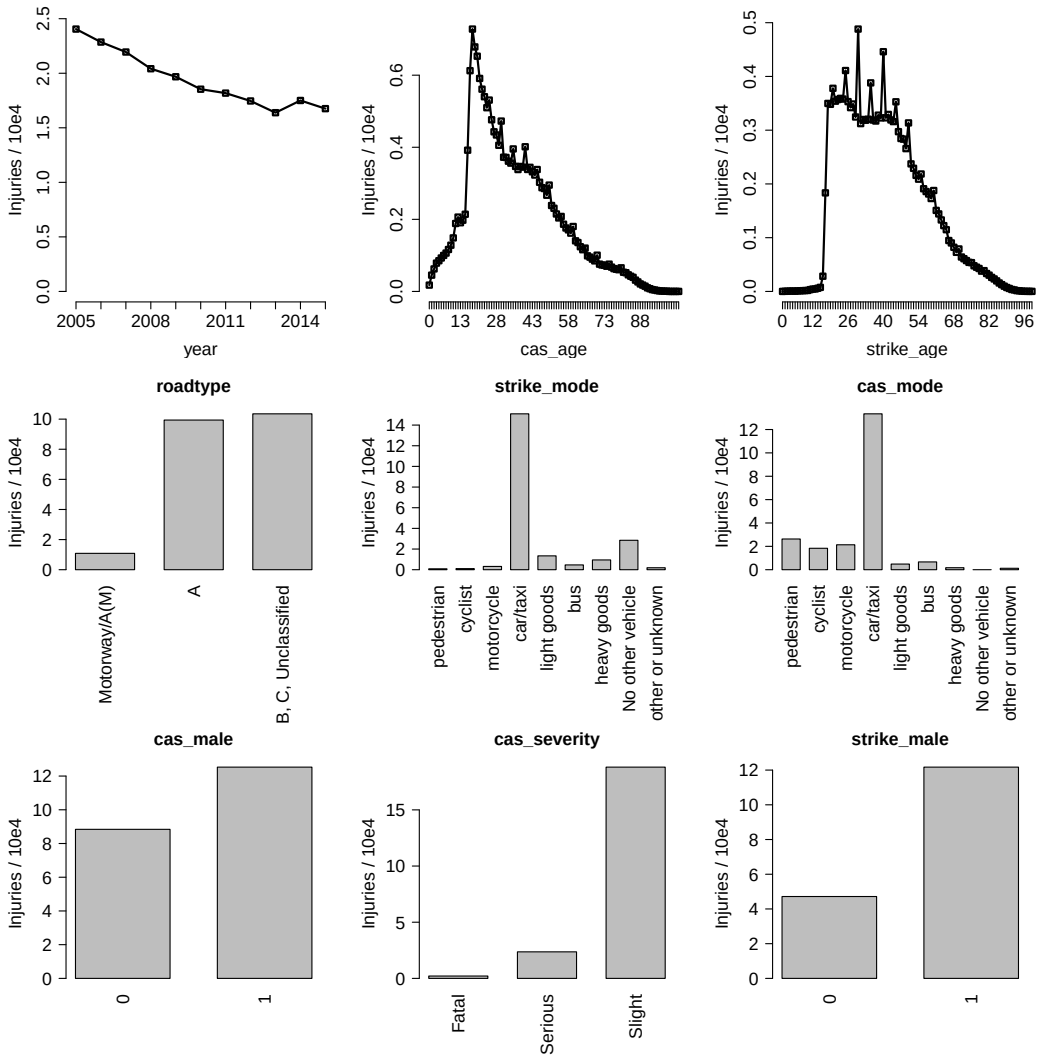

Figure 7: Marginal distributions of all casualties in the Stats19 database from 2005 to 2015.

Before proceeding, we omit any entries with missing data, which includes any entry with modes "no other vehicle" or "other or unknown". Then 1,473,533 entries remain. OR: we have a separate model for "no other vehicle" that includes data from other sources.

Finally, we select age groupings for both the casualty and striker ages, so that the number of age groups is reduced from 106 to 6? 10? for each. The reduced number of groups allows for greater computational efficiency. Age groups are chosen as the quantiles extracted from the data shown in Figure 7.

4 Regression modelling

We seek to build a model for estimating the number of casualties sustained by road users by fitting a negative binomial model to data from the Stats19 database. We make estimates for each combination of: casualty severity, road type, year, strike mode, casualty mode, striker demographic group, and casualty demographic group. The number of casualties in each group is assumed to follow a negative binomial distribution with its own rate, λ. The covariates determine the rate λ in fitting the model. Thus we say that λ is parametrised by the covariates.

We use regression to estimate the number of casualties per distance travelled in order to smooth the casualty data. Our disaggregation of the data results in many counts of zero. Smoothing allows the sharing of support across similar groups. Note that distance data were smoothed prior to regression.

4.1 Negative binomial equation

[EITHER]

We augment the Stats19 dataset by adding to each year-mode-road-demographic group their distance travelled ($A^{(\text{cas})}_{a,g,m,t,y}$ and $A^{(\text{str})}_{a',g',m',t,y}$), the safety-in-numbers exponents (β_1 and β_2), the AADF ($N^{(\text{cas})}_{m,t,y}$ and $N^{(\text{str})}_{m',t,y}$) and the distance travelled by AADF groups ($A^{(\text{cas})}_{m,t,y}$ and $A^{(\text{str})}_{m',t,y}$). Here, the age groups, a and a', are those defined in Section 3.

Then we have all components for the model defined as

$$I_{a,a',m,m',g,g',c,t,y} \sim \text{NB}(\lambda_{a,a',m,m',g,g',c,t,y}, \theta) \tag{11}$$

$$\log\left(\lambda_{a,a',m,m',g,g',c,t,y}\right) = \beta_0 + \log(A^{(\text{cas})}_{a,g,m,t,y}) + \beta_1 \log\left(N^{(\text{cas})}_{m',t,y}\right) - \log(A^{(\text{cas})}_{m',t,y}) + \log(A^{(\text{str})}_{a',g',m',t,y}) + \beta_2 \log\left(N^{(\text{str})}_{m',t,y}\right) - \log(A^{(\text{str})}_{m',t,y}) + \sum_{i=3}^{n} \beta_i X_i. \tag{12}$$

The parameters β_1 and β_2 are safety-in-numbers exponents. They represent the non-linearity in the relation between distance travelled and number of casualties incurred (**??**).

[OR]

We augment the Stats19 dataset by adding to each year-mode-road-demographic group their distance travelled ($A^{(\text{cas})}_{a,g,m,t,y}$ and $A^{(\text{str})}_{a',g',m',t,y}$). Here, the age groups, a and a', are those defined in Section 3.

Then we have all components for the model defined as

$$I_{a,a',m,m',g,g',c,t,y} \sim \text{NB}(\lambda_{a,a',m,m',g,g',c,t,y}, \theta) \tag{13}$$

$$\log\left(\lambda_{a,a',m,m',g,g',c,t,y}\right) = \beta_0 + \log(A^{(\text{cas})}_{a,g,m,t,y}) + \log(A^{(\text{str})}_{a',g',m',t,y}) + \sum_{i=3}^{n} \beta_i X_i. \tag{14}$$

We omit safety-in-numbers exponents on the basis that there are no published coefficients appropriate to a country-level dataset, and our dataset is insufficient to learn them directly (**??**). We show in supplementary material the model that results when we try to learn these as additional parameters (?).

[END]

The model matrix is represented by X and β are the coefficients to fit. In fitting, we remove from the dataset any entries for which $A^{(\text{cas})}_{a,g,m,t,y}$ or $A^{(\text{str})}_{a',g',m',t,y}$ is zero. This takes us from 698,544 unique combinations to 580,536.

4.2 Model building

All covariates we might use are listed at the beginning of Section 3. In addition to using these as predictors alone, we use interactions between them as covariates. To begin, we choose which pairs, triplets, and quadruplets of interacting covariates to include in parametrising the rate parameter λ.

To build the negative binomial regression model, we start from the main-effects model (which consists of nine covariates as predictors, and no interactions), and add one interaction at a time. The "year" covariate is modelled as a smooth spline with two knots, and both age variables are modelled as a smooth spline with five knots.

The algorithm is:

1. Set $\mathcal{M}$ as the nine main-effects model
2. **for** i in 1:N **do:**
   (a) Define the set of possible additional interactions $\mathscr{I}$
   (b) Calculate AIC for all models $\mathcal{M} + \mathcal{I}$ where $\mathcal{I} \in \mathscr{I}$
   (c) Set $\mathcal{M} \leftarrow \mathcal{M} + \mathcal{I}^*$, where $\mathcal{M} + \mathcal{I}^*$ is the model with minimal AIC
3. **end for**

5 Results for baseline

We plot some results for certain sections of the set in Figures 8–15. Predictions are summed over each subgroup and plotted for each gender and each road type. Raw estimates are faded in the background.

These plots highlight some problems with the distance data. E.g., there is a very high rate of females injured on motorcycles on motorways in crashes with no other vehicle; this might be explained by an underestimation of time spent on motorways.

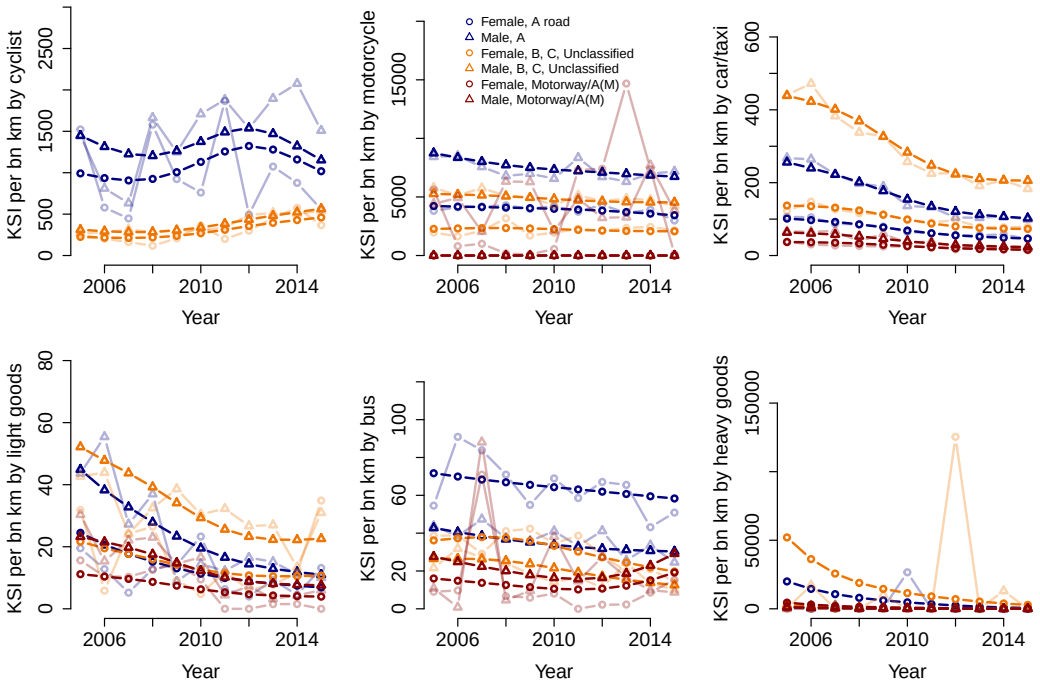

Figure 8: Sum of injuries caused by no other vehicle for each year.

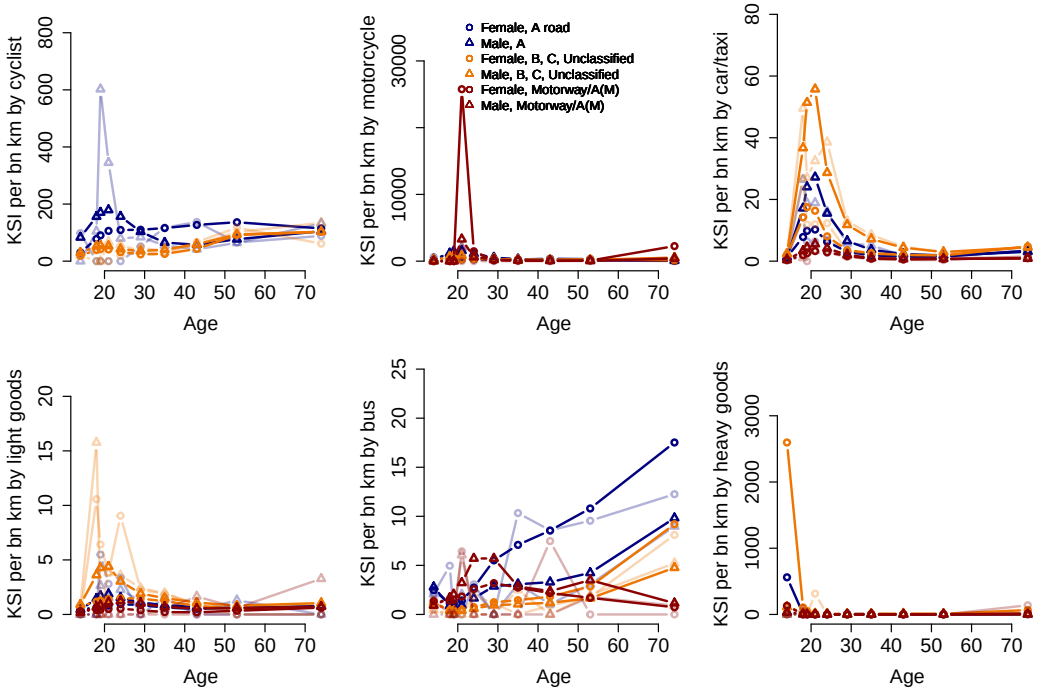

Figure 9: Sum of injuries caused by no other vehicle for each casualty age group in 2015.

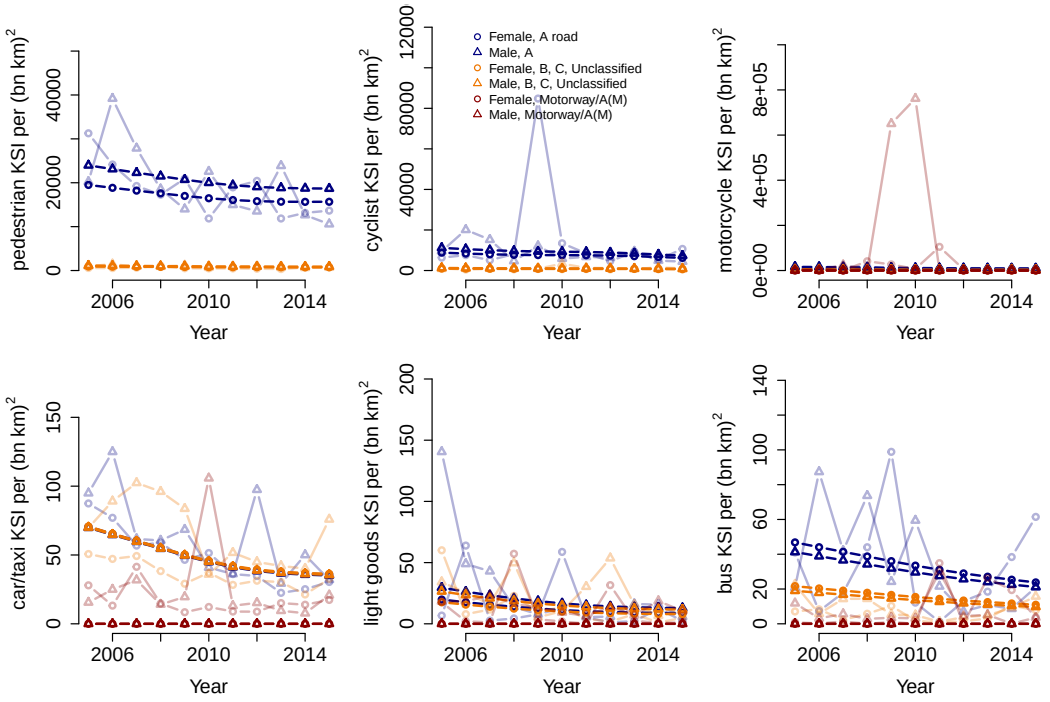

Figure 10: Sum of injuries caused by any vehicle for each year.

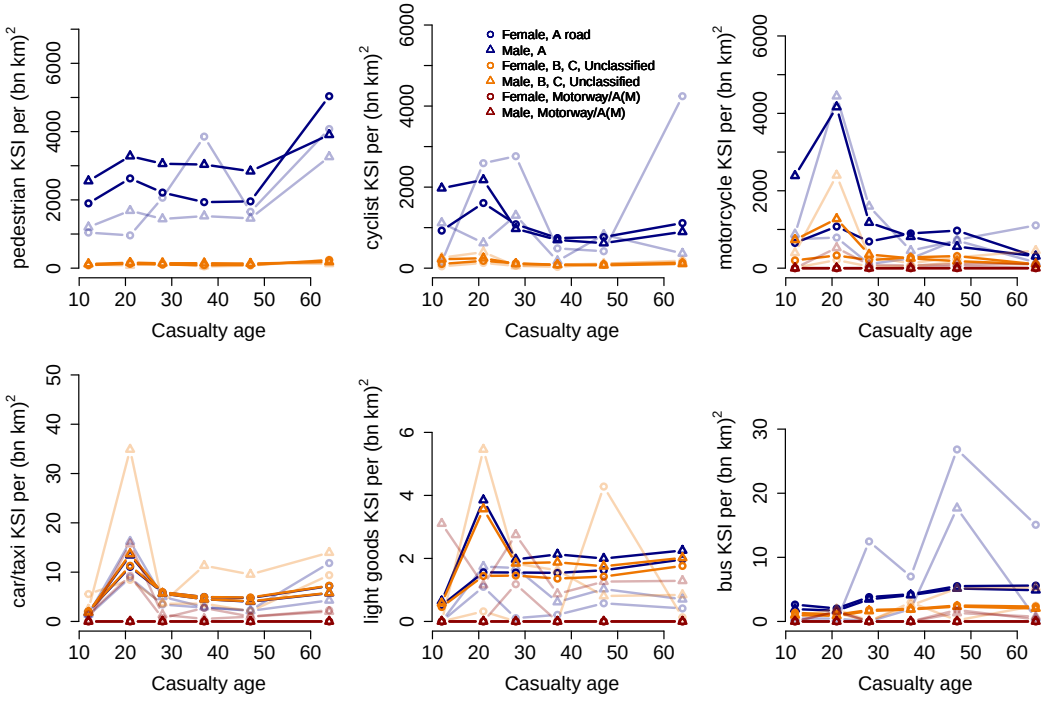

Figure 11: Sum of injuries caused by any vehicle for each casualty age group in 2015.

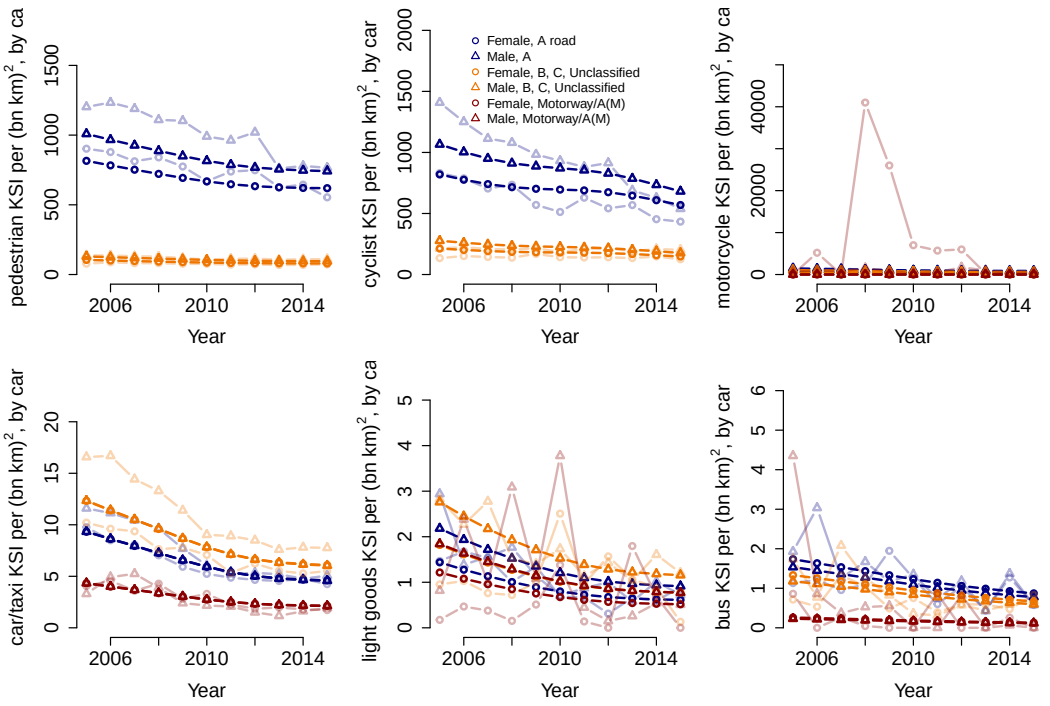

Figure 12: Sum of injuries caused by car for each year.

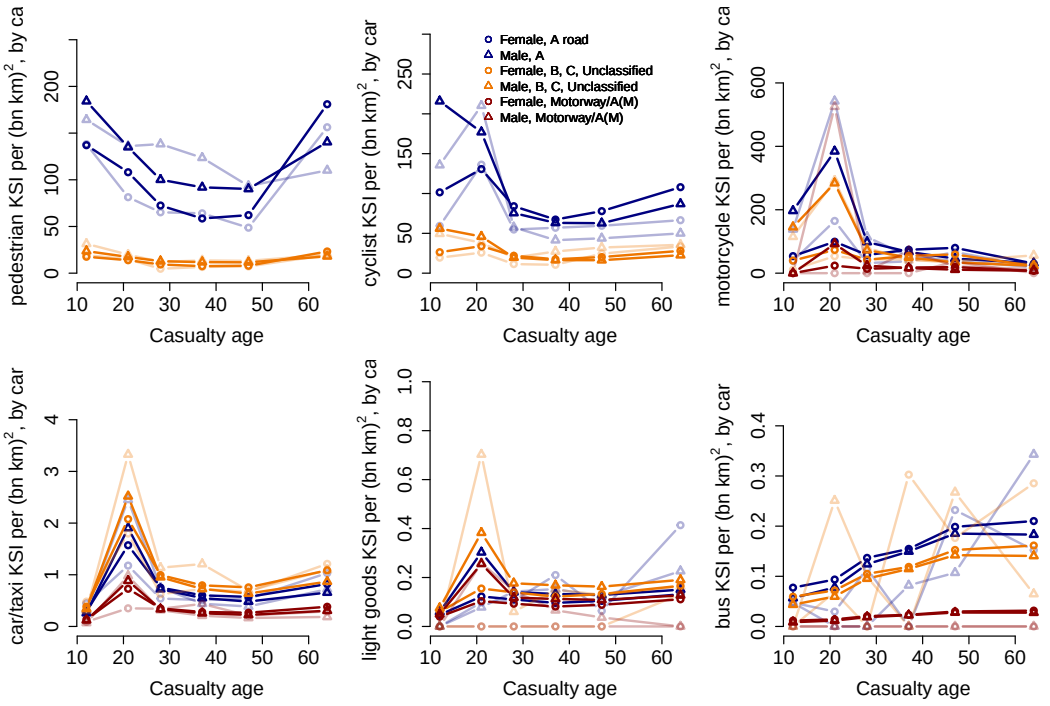

Figure 13: Sum of injuries caused by car for each casualty age group in 2015.

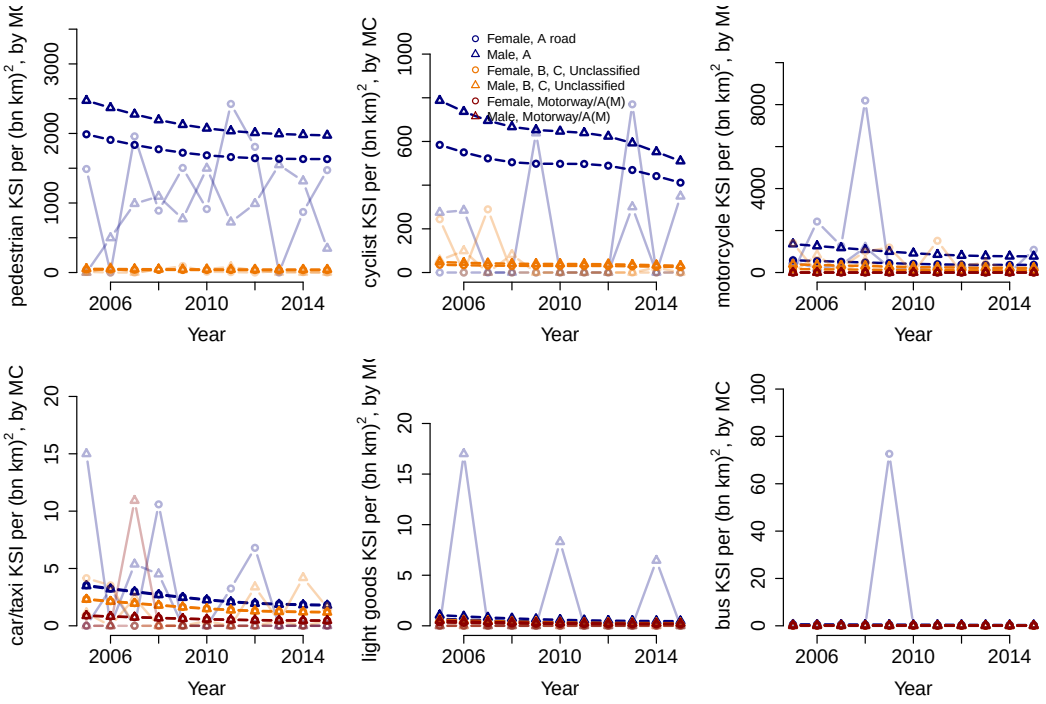

Figure 14: Sum of injuries caused by female motorcyclists for each year.

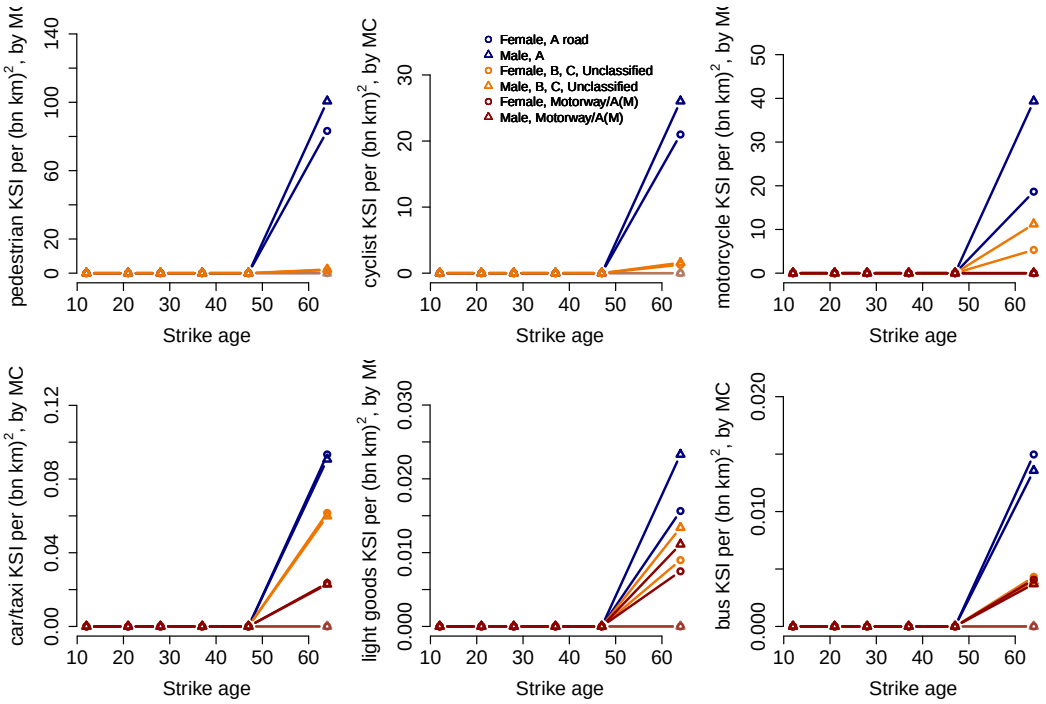

Figure 15: Sum of injuries caused by female motorcyclists for each striker age group in 2015.

6 Predictions for novel scenarios

We make predictions for the baseline by applying the baseline travel distances to the predictive model. These distances might be the same as the offset in the negative binomial model, or calculated in the same way. It could be a sum over the synthetic population.

To make predictions for supposed scenarios, we alter the distances travelled, $B_{a,g,m,t,y,z}$, to reflect the scenario-specific travel behaviour. We can either sum over the synthetic population, or simply re-calculate the baseline distances, e.g., to capture the change resulting from a journey j_{s_0} in the baseline to j_s in the scenario, with the subscript s for scenario, and $s = s_0$ the baseline:

$$B_{a(j),g(j),m(j_s),t(j_s),y(j),z(j),s} = B_{a(j),g(j),m(j_s),t(j_s),y(j),z(j),s_0} + D_{j_s,t(j_s)}$$

$$B_{a(j),g(j),m(j_{s_0}),t_{s_0},y(j),z(j),s} = B_{a(j),g(j),m(j_{s_0}),t_{s_0},y(j),z(j),s_0} - D_{j_{s_0},t_{s_0}}$$