The injury model for the ITHIM

October 27, 2017

We seek to build a model for injuries by regressing a Poisson model onto data from the Stats19 database. The database provides information about the details of each injury, which allows the injuries to be grouped. The number of injuries in each group is assumed to follow a Poisson distribution with its own rate, λ. The details, which we call covariates, are what we relate to the rate λ in fitting the model. Thus we say that λ is parametrised by the covariates.

All covariates we might use are listed in Table 1. In addition to using these as predictors alone, we can use interactions between them as covariates in their own right. Our first aim, then, is to choose which covariates to include alone, and which pairs, triplets, etc. of interacting covariates to include in parametrising the rate parameter λ.

In the first instance (Section 2) we consider only the first eight covariates of Table 1, for which the total number of combinations is 57600, as is the total number of possible coefficients in the model. In later sections we investigate inclusion of smooth splines for year and age (Section 3), and then moving to district level (Section 4).

In the end we have a principled method for building a model (in terms of which predictors to use, and how), and can make predictions for different scenarios.

Table 1: Covariates used in the injury model. The modes each have eight categories: pedestrian, pedal bike, car/taxi, motorbike, heavy goods, light goods, unknown/other. The severity has three levels: slight, severe and fatal. The road type has three categories: M, A, and other. There are two categories for each gender. There are five categories for each age group. For strikers, they are: 0–20, 20–40, 40–60, 60–80, 80+. For casualties, they are: 0–16, 16–25, 25–60, 60–80, 80+. The years span 2005–2015. We include ages 0–104.

Covariate	Label	Levels
Casualty mode	cas_mode	8
Casualty severity	cas_severity	3
Road type	roadtype	3
Striker mode	strike_mode	8
Casualty gender	cas_male	2
Striker gender	strike_male	2
Striker age group	strike_age_band	5
Casualty age group	cas_age_band	5
Year	year	11
District	district	153
Casualty age	cas_age	105
Striker age	strike_age	105

1 Descriptive statistics

There are 1,494,789 injuries recorded in the data set, having excluded incomplete entries and those with 'no other vehicle'. Of these, 81% had car/taxi strikers, 63% had car/taxi casualties, 69% had strikers aged 25–60, 41% had casualties aged 40–60, 88% were slight injuries, 71% had male strikers, 58% had male casualties, and 0.05% occurred on motorways.

Figures 1 and 2 show the relationships between the individual variables and the number of injuries. Multivariate interactions are explored in Appendix A.

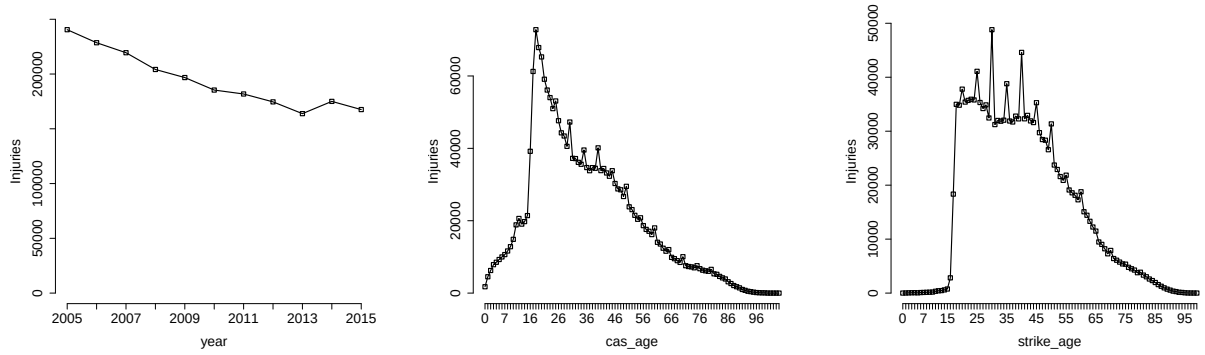

Figure 1: Line plots of injuries in terms of age and year.

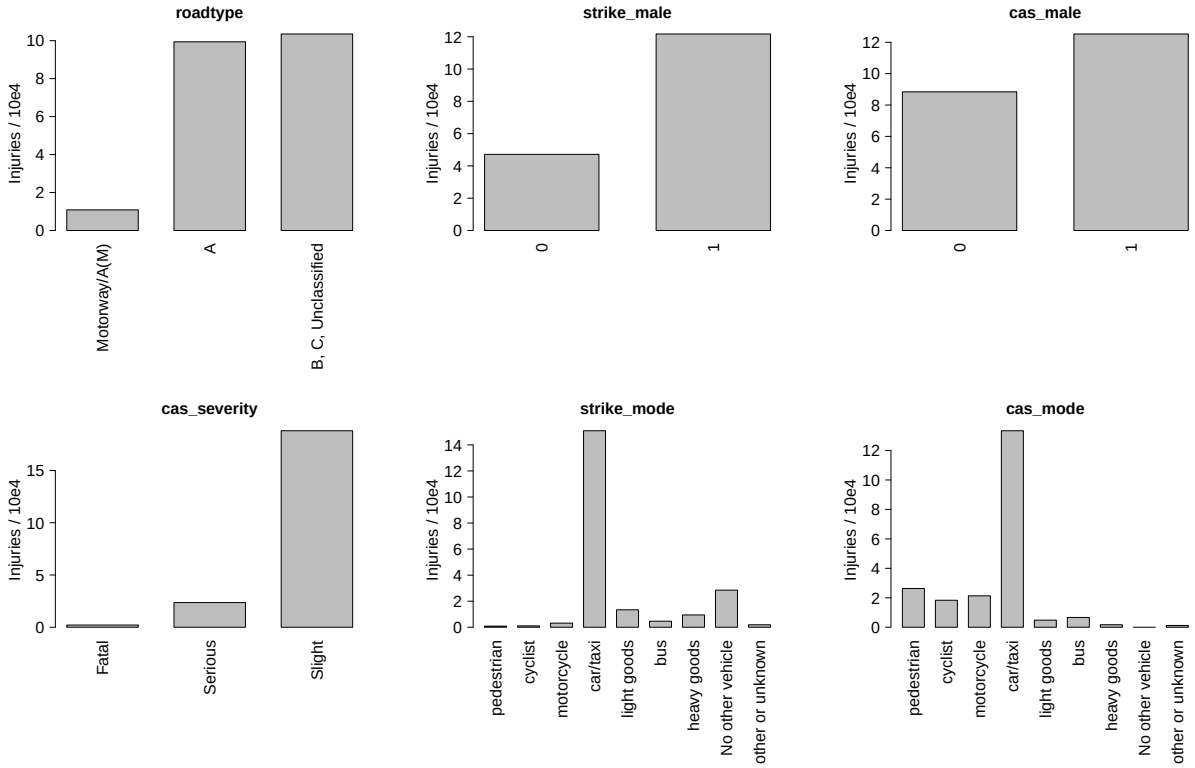

Figure 2: Bar plots of injuries in terms of road type, gender, severity, and mode.

2 Regressions

2.1 Method

To build the Poisson regression model, motivated by the approach outlined in Appendix B, I will start from the main-effects model (which consists of eight covariates as predictors, and no interactions), and add one interaction at a time. The algorithm is:

1. Set $\mathcal{M}$ as the eight 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**

2.2 Results: eight covariates

Using all eight covariates, we build a model with 87 factors, of which 28 are two-way, 50 are three-way, and nine are four-way. The factors are listed in order of inclusion in Table 6, and their AIC values are plotted in Figure 3.

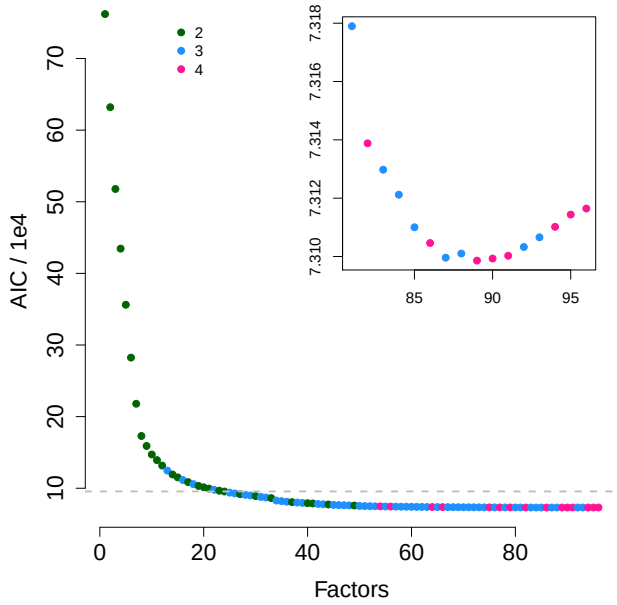

Figure 3: AIC values for the first 56 factors to be added to the greedy model. Each point is an additional interaction in the model. Points are added to the model from left to right. Green indicates a two-variable interaction; blue, a three-variable interaction. Grey line: James' model. Inset: factors 34 and onward.

2.3 Predictions

Our model allows us to predict the baseline number of injuries for England. The prediction ought to reflect the data originally supplied, and ought also to correct somewhat for noise via the assumptions implicit in the model.

For example, suppose we observe no 60-year-old male pedestrian fatalities. Suppose we observe many 59- and 61-year-old male pedestrian fatalities, as well as 60-year-old female pedestrian fatalities. Then our model ought to predict a number of 60-year-old male pedestrian fatalities, which will conflict with our data but agree with our intuition.

2.3.1 Scenario prediction model

The next step is to incorporate mode shift in order to predict outcomes for different scenarios. We use the 'offset' feature in the regression model, so that the predictions generated incorporate relative travel times and no post-processing is required. Specifically, we predict injuries with the following rates:

$$\lambda(i) = \lambda_0(i) \cdot a_v(i)^{b_v(i)} \cdot c_v(i) \cdot a_s(i)^{b_s(i)} \cdot c_s(i). \tag{1}$$

where $\lambda(i)$ is the rate and $\lambda_0(i)$ the base rate for group i. The base rate depends only on the eight covariates included (and their interactions). This equation shows how, to predict an outcome for a new scenario, we simply scale the result from the baseline case by the change in exposure. However, the change in exposure is not simply the total amount of time spent travelling: it must itself be scaled to account for "safety in numbers", the observation that collisions do not increase linearly with travel time.

Safety-in-numbers factors are represeted by $b_x(i)$, and are taken from Analytica. $a_x(i)$ are group relative times, and $c_x(i)$ are relative times. For the baseline, all $a_x(i) = c_x(i) = 1 \,\forall i$. For scenarios, these values relate to the ratio of travel times in the scenario to travel times in the baseline.

Using Analytica notation, $a_x(i) = \hat{A}_{m,s}$, and $c_x(i) = \check{A}_{a,g,m,s}$. Therefore, to calculate the two values for each group, we need to supply $A_{a,g,m,s=\text{baseline}}$:

$$A_{a,g,m,s} = N_{a,g} \cdot T_{a,g,m,s} \tag{2}$$

$$\tilde{A}_{m,s} = \frac{\sum_{a,g} A_{a,g,m,s}}{|a||g|} \tag{3}$$

$$\hat{A}_{m,s} = \frac{\tilde{A}_{m,s}}{\tilde{A}_{m,s=\text{baseline}}} \tag{4}$$

$$\check{A}_{a,g,m,s} = \frac{A_{a,g,m,s}}{A_{a,g,m,s=\text{baseline}} \cdot \hat{A}_{m,s}} \tag{5}$$

where $N_{a,g}$ is the population fractions of age groups (a) and genders (g) and $T_{a,g,m,s}$ are travel times, indexed also by mode (m) and scenario (s).

For simplicity, in our implementation, we make the following approximations:

- Striker times are the same as casualty times (i.e. $a_s(i) = a_v(i)$ and $c_s(i) = c_v(i)$).
- Travel data are constructed using average trip times[1] and average number of trips per mode, per gender, per age group[2]. These data do not include HGVs; LGVs are grouped with cars; taxis are separate; motorbikes are grouped with private vehicles (including school buses). The object $A_{a,g,m,s=\text{baseline}}$ is constructed using the following assumptions:
  - 'Intercity bus' trip-duration data can be used to represent HGVs;

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/632812/nts0311.ods
[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/632847/nts0601.ods

  - 'Car/van' trip-number data can be used to represent cars, ignoring taxi data;
  - 'Other private vehicle' trip-number data can be used to represent motorbikes (which should be corrected for school-age and retirement-age people);
  - 'Car/van' trip-number data can be used to represent LGVs, after scaling down according to road milage for each vehicle (see Table 2);
  - 'Car/van' trip-number data can be used to represent HGVs, after scaling down according to road milage for each vehicle (see Table 2), and accounting for 99% of HGV drivers being male[3];
  - The category of 'other or unknown' mode is currently the same as LGV.
- Travel data do not have a road-type variable. A new object $A_{a,g,m,s=\text{baseline},t}$ (where the index t is road type) is imputed from $A_{a,g,m,s=\text{baseline}}$ using the number of injuries as an indicator of frequency of road use. See Table 2 for an evaluation.
- We present predictions without uncertainty, though these are easily accessible from the R glm function.

Table 2: Total travel time is divided up into road types, of which there are three: motorway, A, and other. We use Department for Transport (DfT) data to compare to our approximation for three road users for which they supply data: cars, light goods vehicles and heavy goods vehicles. We compare these proportions with those resulting from summing over all subgroups in our table. In our implementation, we use injury data as a proxy for each subgroup's usage of each road type. Clearly, this is inherently biased, as we will overestimate presence on the more dangerous roads. (This could perhaps be alleviated to some extent by using the results of the regression model.) The DfT data used consist of the number of miles travelled[4] divided by the average speed[5]. The road relations are shown in Figure 4.

Mode	Source	Motorway	A road	Other road
Car	*Data*	0.10	0.44	0.46
	Model	0.07	0.50	0.43
LGV	*Data*	0.10	0.42	0.48
	Model	0.10	0.53	0.37
HGV	*Data*	0.32	0.50	0.17
	Model	0.22	0.55	0.23

[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/627597/domestic-road-freight-statistics-2016.pdf

[4] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/644453/tra2503.ods

[5] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/623257/spe0112.ods

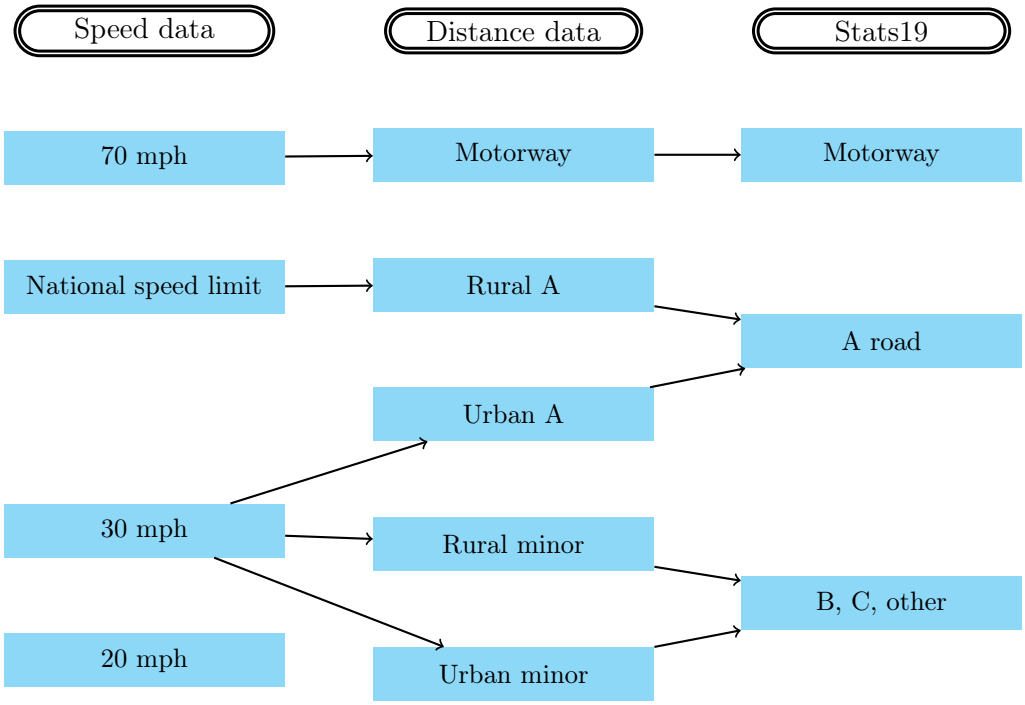

Figure 4: Relating road types in three data sets. The speed dataset has average speeds for the three vehicle types for roads of four types: 70, 30 and 20 mph speed limits, and national speed limit. These we map onto the distance dataset, which has five road types. (HGVs do not have the 'minor' road type; we interpret the sum of these as the difference between the total and the sum of the other three.) These map naturally onto the Stats19 database, which three road types. What is the room for improvement in Table 2? Where the model value exceeds the data value, it might be that we have overestimated the speed of the vehicle, e.g. HGVs on minor roads (30 mph).

2.3.2 Scenario prediction results

Results are given in Table 3. We show Scenario 2, pedestrians hit by motorcyclists, in more detail in Figure 5. We see that as the total travel time increases, so does the number of injuries. However, the total number of male victims decreases, as they are afforded protection by female pedestrians.

Note the unlikely prediction in Scenario 4 that the number of pedestrian injuries due to motorbikes decreases when all travel of 0–16 year olds doubles. This is a result of the mis-specification of motorbike travel time for 0–17 year olds; no motorbike-specific data were used, only that amalgamated with other private transport, e.g. school buses. This means that the model predicts a large increase in motorbikes with little accompanying increase in injuries, as few injuries were observed caused by young motorcyclists. (The effect is not seen in cyclists as the contribution to the casualty safety-in-numbers base is so small: 1.10 for cyclists, vs. 1.18 for pedestrians.)

Table 3: Expected number of injuries using offset in the Poisson regression, covering an 11-year period. For scenario 1, we double travel time for pedestrians and cyclists on A, B and C roads, and halve it for cars and taxis on all roads. In scenario 2, we double all travel time by female travellers. In scenario 3, we reduce 20–40-year-old male car driver travel times by a factor of ten. In Scenario 4, we increase all travel of 0–16 year olds by a factor of 2.

Scenario	Pedestrians hit by motorcycles	Cyclists hit by motorcycles	Cars/taxis colliding with motorcycles
Baseline	8161.000	1962.000	134390.000
Scenario 1	10767.865	2774.273	83734.180
Scenario 2	8996.834	2070.128	154722.99
Scenario 3	8161.000	1962.000	132001.27
Scenario 4	7995.307	2026.240	136076.94

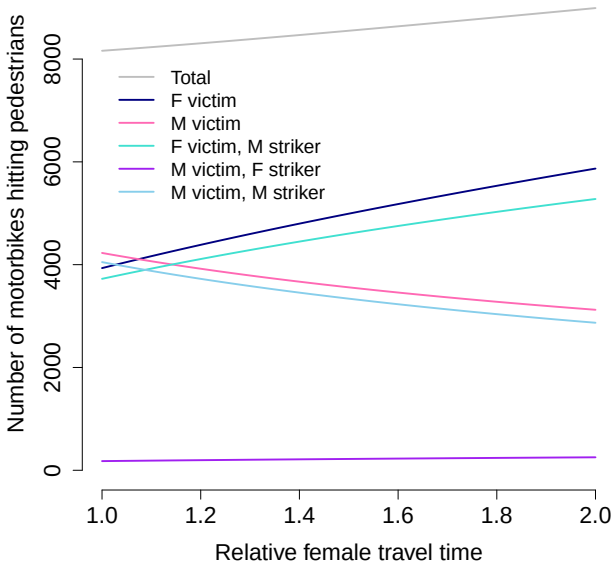

Figure 5: Number of pedestrians hit by motorcyclists as relative travel time of all female people in the population changes, broken down into groups by gender of victim and striker.

2.4 Conclusion to regression section

The model-building approach finds a natural limit, where addition of new interactions ceases to improve the AIC. For the seven-covariate model, there is a minimum at 48 factors (Figure 29, inset). For the eight-covariate model, there is a minimum at 87 factors (Figure 3, inset).

Using these models, we are able to make predictions for the number of injuries we expect in new scenarios, having made some assumptions. The prediction mechanism will be improved through the following adjustments:

1. Improving (use of) relevant travel-time data for England;
2. Reporting the uncertainty of the prediction;
3. Improving the partition of travel times to road types.
4. Treating victim times and striker times differently, so that, for example, increased car and bus passengers increase the travel time of potential victims, but not the travel times of potential strikers or the group relative times of those modes. E.g. additional 'casualty_mode' options: car passenger, van passengerÂÿbus passenger; and the equivalent strike modes specifically mean the drivers.

3 Regression using smooth splines

The value in using splines is that they leverage relatedness between adjacent categories. For example, using the year alone as a predictor, we find a fit with an AIC of 1796 (2 d.f.) using a categorical model, but 727 (4.8 d.f.) using a spline with five knots.

We test the spline against a categorical model with the same extent of disaggregation. Such a comparison is shown in Figure 6. On the left, we see that, if ages are the only factors, there is no benefit to using a spline rather than categorical data, at any level of disaggregation. However, introducing one interacting factor (casualty mode), we see an improvement in using splines upon categorical.

The relationship between disaggregation, knot number and goodness-of-fit seems not to be straightforward. It will be worth testing these in a later stage, when other elements of the model are confirmed.

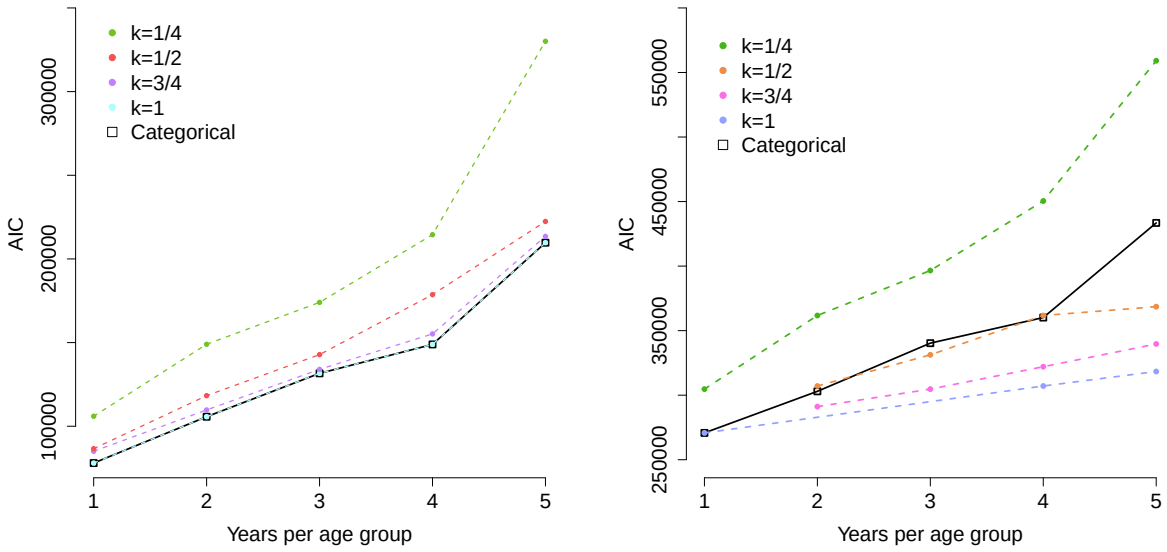

Figure 6: Evaluation of splines. AICs where both ages are splines. On the x axis, the number of years in each age group (1 year is full disaggregation). In the black, the same data are treated as categorical. The number of knots used for the spline is a fraction k of the total number of possible knots. Left: ages-only model. Right: casualty age interacts with casualty mode.

3.1 Model building

We use the `mgcv` package to build a spline for ages, beginning with casualty age, still binning over striker age. We use eight knots. The spline is shown in Figure 7. The model, built as before, is shown in Figure 8. Model factors are listed in Table 8.

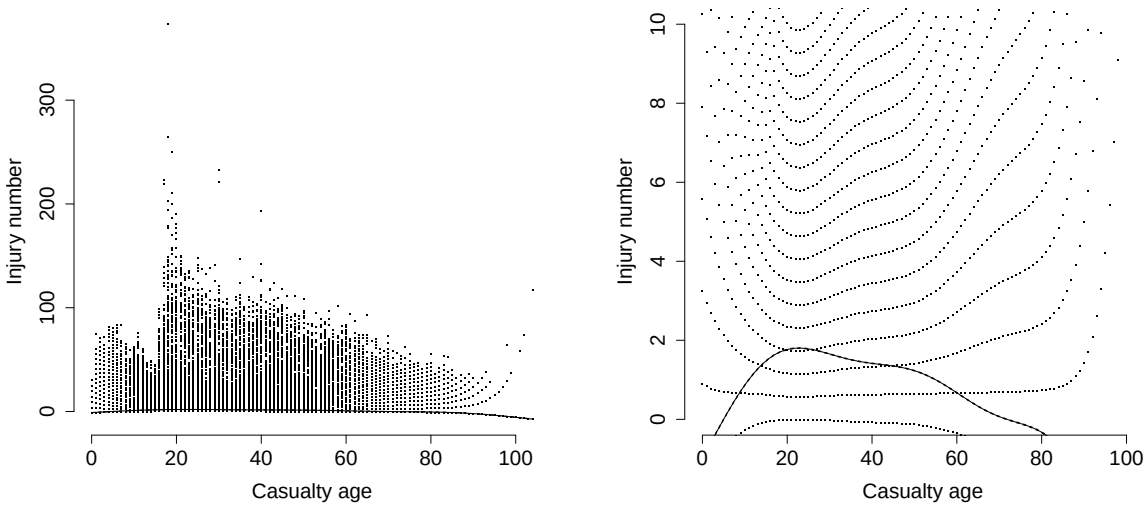

Figure 7: Fit splines to age for casualties, shown in full (left) and zooming in on the regressed line (right), using eight knots. Dots show residuals.

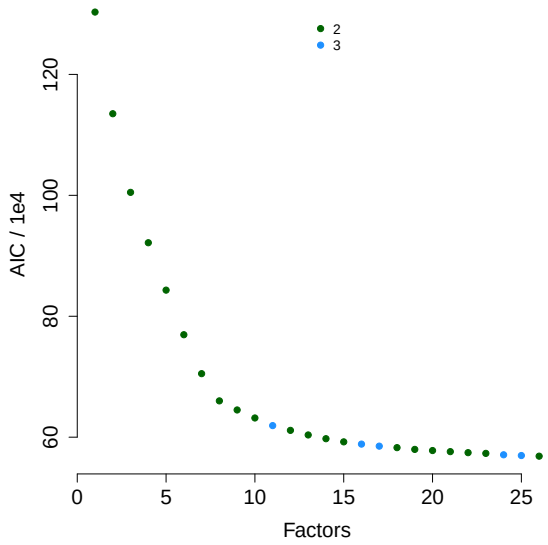

Figure 8: Model building using spline for casualty age.

How are interactions ranked in the eight-categorical model and the casualty spline age model? They are very similar in terms of order. It seems that inclusion of the spline promotes interactions involving casualty age and casualty mode. (We might expect to see the same effect in strikers.) Thus, as an approximation, we might use the result of the eight-covariate model building for the spline model.

We can compare the two models, then, up to the first eight interactions (which are common to both) on the disaggregated dataset. The AIC values are 1,527,246 (166 d.f.) for the categorical model and 660,072 (214 d.f.) for the spline model.

Table 4: Comparison of built models, where one models uses a spline for age and the other categories.

Iteration	Spline rank	Categorical rank
cas_mode, cas_age	1	3
cas_male, cas_mode	2	1
strike_mode, strike_male	3	2
cas_severity, cas_mode	4	4
roadtype, cas_mode	5	5
strike_mode, strike_age_band	6	6
strike_mode, cas_mode	7	7
roadtype, strike_mode	8	8
roadtype, cas_age	9	9
cas_severity, cas_age	10	10
roadtype, cas_mode, cas_age	11	13
strike_mode, cas_severity	12	11
strike_age_band, cas_mode	13	12
strike_age_band, cas_age	14	14
strike_mode, cas_age	15	15
strike_mode, cas_mode, cas_age	16	18
roadtype, strike_mode, cas_mode	17	16
strike_male, strike_age_band	18	19
roadtype, strike_age_band	19	17
roadtype, strike_male	20	20
cas_male, cas_severity	21	21
strike_male, cas_age	22	23
roadtype, cas_severity	23	24
cas_severity, cas_mode, cas_age	24	22
strike_mode, strike_age_band, cas_age	25	26
cas_male, cas_age	26	33

4 Injuries by district

4.1 Data & model

In order to make area-specific predictions, we include the location data from the Stats19 database. These label each injury with one of 153 different place codes such as 'E09000007',[6] yielding 8,812,800 cells in total. We refer to these as districts. In the first instance we use district as a categorical variable. We repeat the model-building exercise as before to see what pattern emerges for district as a predictor.

4.2 Results

Table 5: Building the regression model interaction by interaction for nine covariates. Ages are categorical. The new covariate is district. (Early iterations are each taking 36 hours.)

Iteration	Covariates	AIC
1	cas_male, cas_mode	2352826
2	strike_age_band, cas_age_band	2347866
3	strike_mode, strike_male	2217826
4	district, cas_age_band	2187371

Interestingly, with the additional district covariate, the relationship between age bands has jumped up 12 places (relative to Table 6).

4.3 Offset

We imagine now that we have some travel-pattern data, for example [distance travelled for cycles(?)] for each district.[7] These could be entered into the offset values a_x and c_x. Then scenarios would be implemented via adjustments relative to these values. Or we include the data as a covariate, rather than an offset. What would this mean for a_x and c_x? (The offset means we model rates, rather than counts.)

4.4 Conditional autoregressive models

To leverage spatial dependency of injury rates between districts, we can use a conditional autoregressive model. We define the rate $\lambda(i)$ to depend on a value that follows a conditional autoregressive (CAR) specification, i.e. it is parametrised by a function of its direct neighbours.

[6] They are E06 (Unitary Authorities), E08 (Metropolitan Districts), E09 (London Boroughs), E10 (Counties) and EHEATHROW; http://statistics.data.gov.uk/.

[7] Mode share: covariate. Time/distance: offset.

A Multiway interactions

The following plots illustrate the relationships between two, three or four variables and the number of injuries. Darker colours indicate more injuries. Where applicable, variables are subset by gender, shown in red, blue, and purple, for male individuals or pairs, female individuals or pairs, and mixed pairs, respectively.

In the following subsections, we test various possible independence models between the covariates. We denote the eight covariates as x_i, $i = 1, ..., 8$. We are interested in the relationships between the variables and, in particular, if there are ways to factorise the distribution $p(x_1, ..., x_8)$.

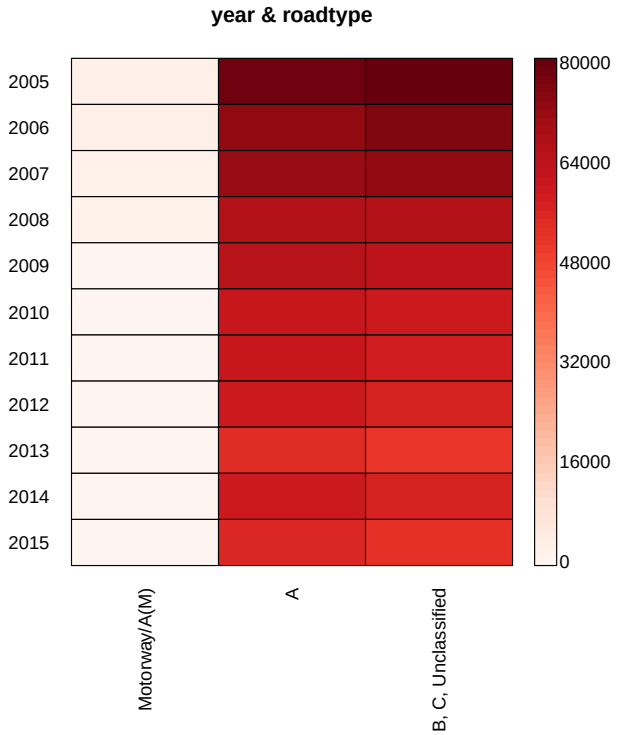

Figure 9: Number of injuries per year and road type.

A.1 Marginal independence

First we test marginal independence between all covariates. In terms of our notation, marginal independence for covariates i and j implies that

$$p(x_i, x_j) = p(x_i) \cdot p(x_j).$$

We use the standard χ^2 test for categorical data and ordinal χ^2 test using the `coin R` package for ordinal data. Of the eight covariates, four are categorical, and four ordinal: road type, severity, and the two age bandings.

With two categorical variables, we have a generalised Pearson χ^2 test. With one ordinal variable, we have an asymptotic generalised Pearson χ^2 test. With two ordinal variables, we have an asymptotic linear-by-linear association test.

With a significance threshold of 0.001, we find no reason to reject marginal independence of casualty gender from both road type ($p = 0.0775$) and striker age band ($p = 0.170$). We have reason to reject all 34 other two-way marginal independence assumptions ($p < 1e^{-36}$). See Figure 25.

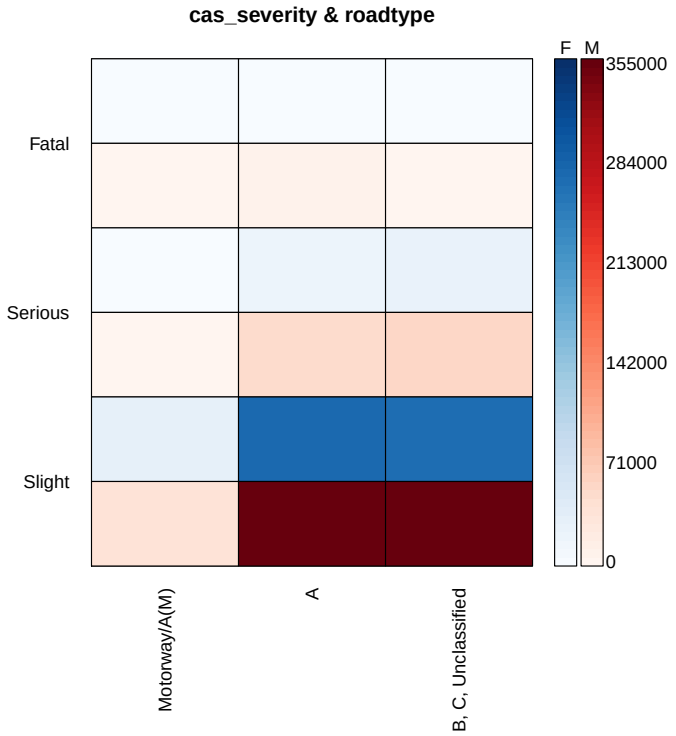

Figure 10: Number of injuries per casualty severity and road type.

A.2 Joint (marginal) independence

Considering a three-way marginal, joint independence asks whether two variables are jointly independent of the third. In our notation, we write

$$p(x_i, x_j, x_k) = p(x_k) \cdot p(x_i, x_j)$$

to say that variable k is jointly independent of variables i and j.

A.2.1 Assuming all covariates are categorical

Assuming all covariates are categorical, we can calculate joint (marginal) interactions between a single variable and n-way marginals up to $n = 7$. Many of these tests fail, particularly at higher orders due to low cell counts. All tests that succeed find no reason to accept the hypothesis of joint (marginal) independence.

A.2.2 Using ordering information

We use the ordinal χ^2 test as before, with intuitive application to the single variable. We consider only three-way joint marginals. With ordering applied only to the single variable, we find no significant results. With ordering applied also to the joint marginal, we find some significant results.

To apply ordering to a variable, weights must be supplied that imply the ordering. We use a monotonic, linear weight. When one of the joint variables is ordinal, the set of factors is weighted ("scored") according to the values for the ordinal variable. Multiple combinations of factors will therefore have the same weight.

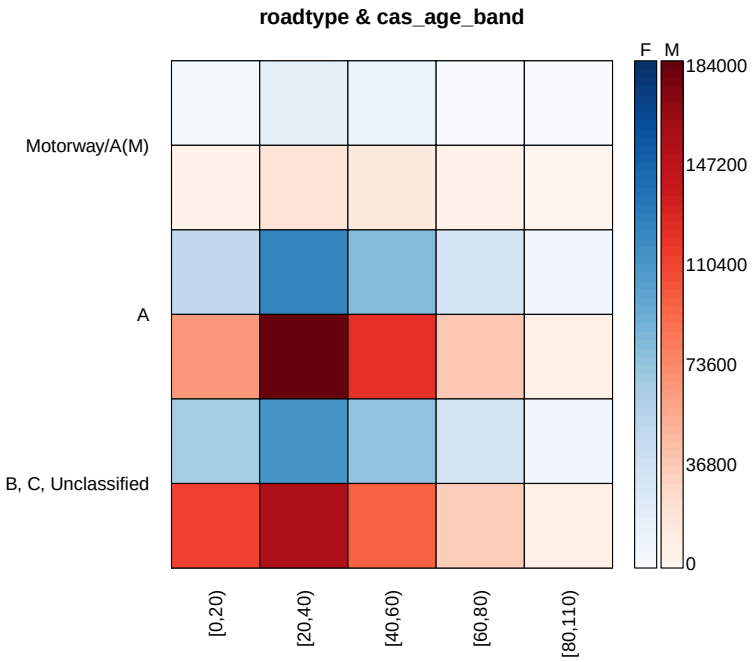

Figure 11: Number of injuries per casualty age and road type.

The ordinal χ^2 test is not easily applied to two joint ordinal variables (as it is unclear how to order *Motorway & fatal*, *Motorway & severe*, *Motorway & slight*, *A road & fatal*, *A road & severe*, *A road & slight*, etc). Therefore, combinations that would require such ordering were treated as categorical. (Note that this prevents an ordinal test of joint independence of striker age band and road type from casualty gender.)

The following hypotheses for joint independence are not rejected, using the same threshold as before:

- Road type and casualty mode are jointly independent of casualty gender ($p = 0.0775$);
- Road type and strike mode are jointly independent of casualty gender ($p = 0.0775$);
- Road type and strike gender are jointly independent of casualty gender ($p = 0.0775$);
- Striker age band and casualty mode are jointly independent of casualty gender ($p = 0.170$);
- Striker age band and strike mode are jointly independent of casualty gender ($p = 0.170$);
- Striker age band and striker gender are jointly independent of casualty gender ($p = 0.170$).

A.3 Conditional (marginal) independence

We test the notion of conditional (marginal) independence via log ratios. If variables i and j are conditionally independent given variable k, then

$$p(x_i, x_j|x_k) = p(x_i|x_k) \cdot p(x_j|x_k).$$

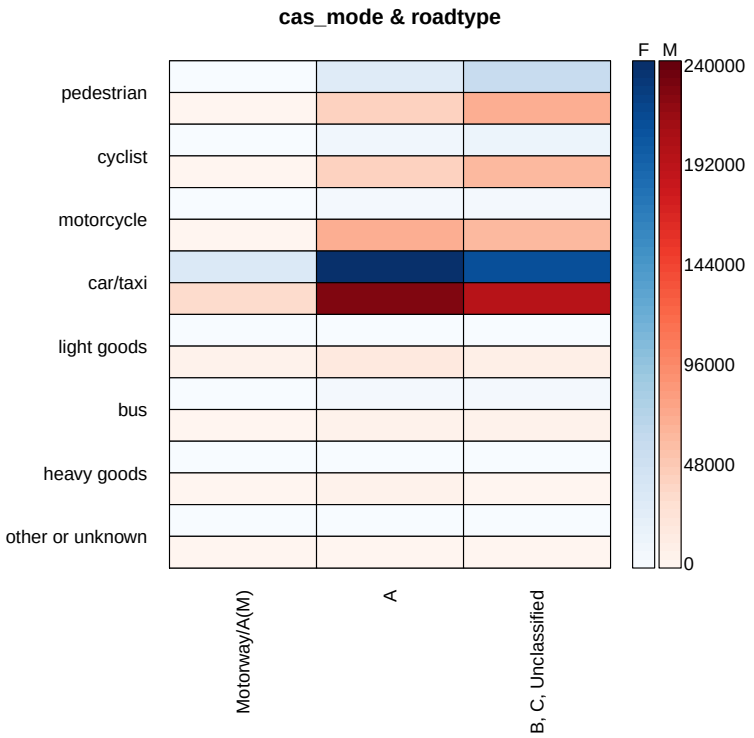

Figure 12: Number of injuries per casualty mode and road type.

Treating all variables as categorical, there is no support for conditional (marginal) independence. Using the linear-by-linear association test `lbl_test` in the `coin R` package for dual ordinal and ordinal–binary covariate combinations, we do not reject the following hypotheses for conditional (marginal) independence:

- Road type and casualty gender are conditionally independent given casualty severity ($p = 0.799$);
- Road type and casualty gender are conditionally independent given casualty age band ($p = 0.690$);
- Road type and casualty gender are conditionally independent given strike mode ($p = 0.00120$);
- Road type and casualty gender are conditionally independent given striker age band ($p = 0.00384$);
- Striker age band and casualty gender are conditionally independent given road type ($p = 0.177$);
- Striker age band and casualty gender are conditionally independent given casualty severity ($p = 0.642$);
- Striker age band and casualty gender are conditionally independent given strike mode ($p = 0.0189$);
- Striker age band and casualty gender are conditionally independent given striker gender ($p = 0.553$).

(From Section A.1, casualty gender is likely marginally independent of both road type and striker age band.)

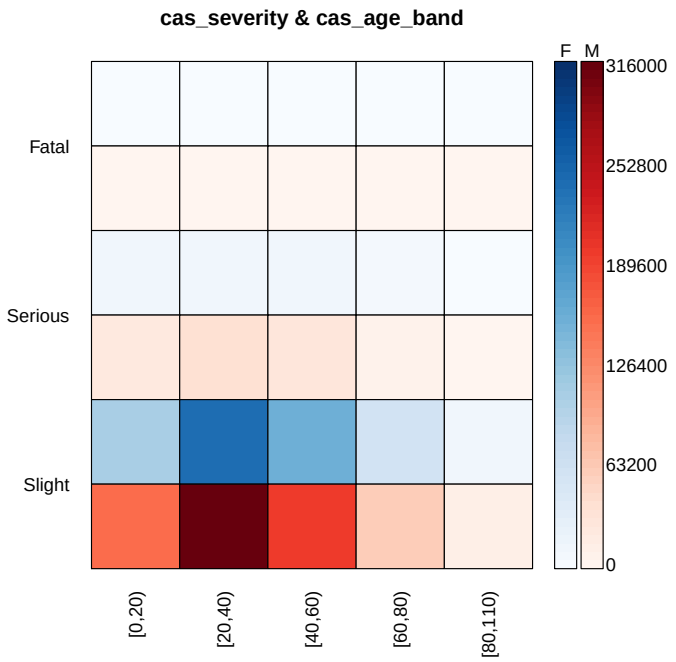

Figure 13: Number of injuries per casualty age and severity.

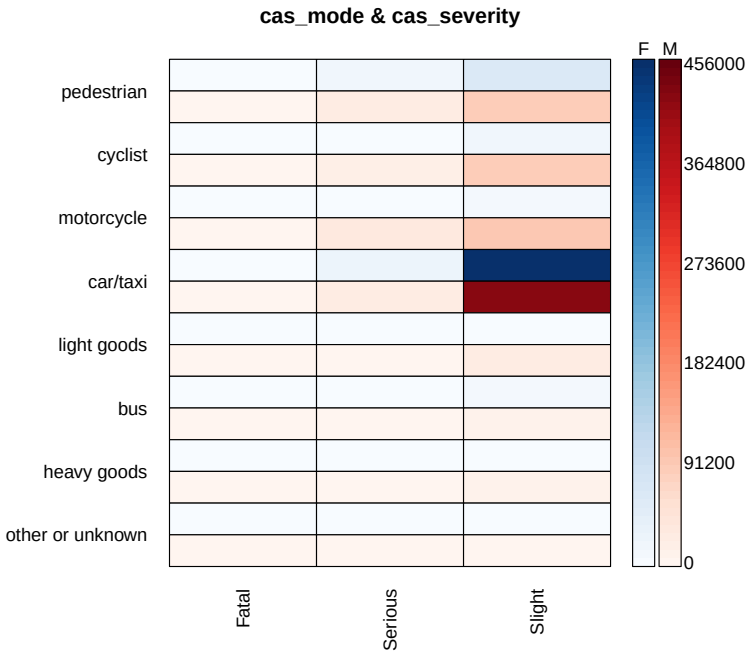

Figure 14: Number of injuries per casualty mode and severity.

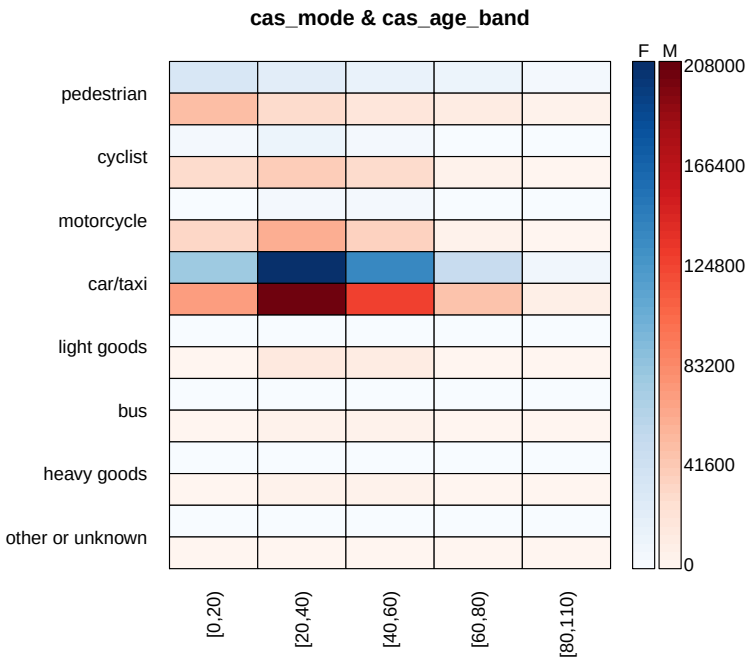

Figure 15: Number of injuries per casualty age and mode.

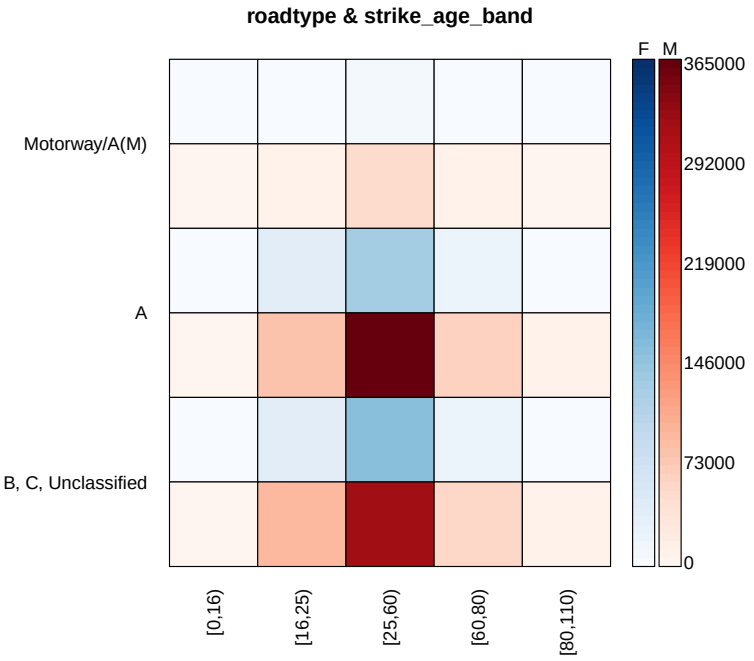

Figure 16: Number of injuries per striker age and road type.

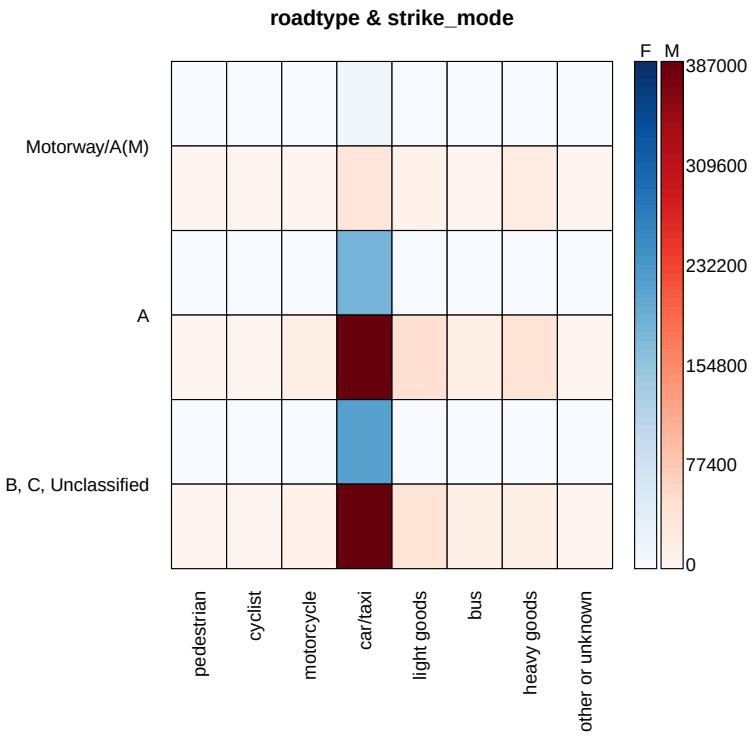

Figure 17: Number of injuries per striker mode and road type.

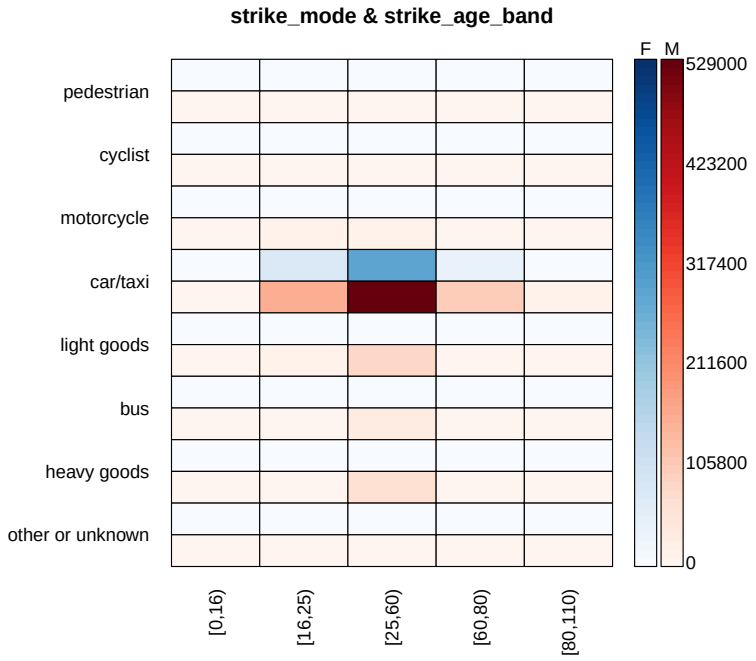

Figure 18: Number of injuries per striker age and mode.

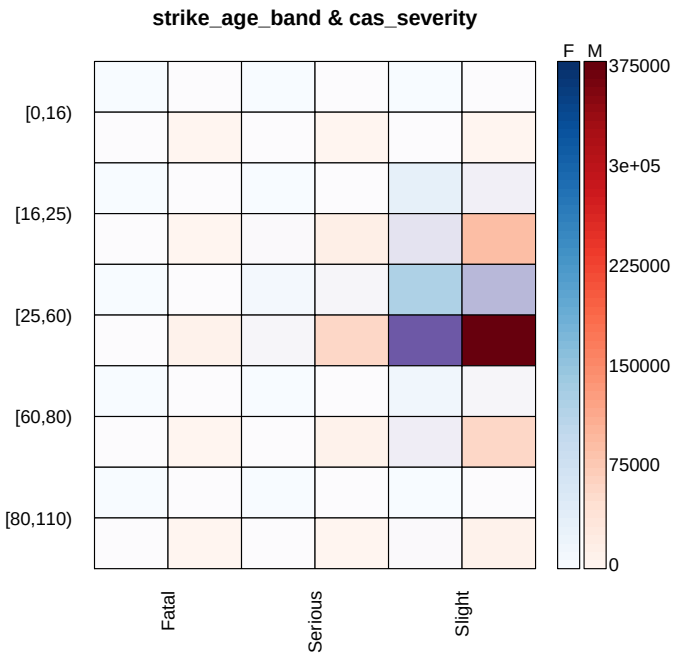

Figure 19: Number of injuries per casualty severity and striker age.

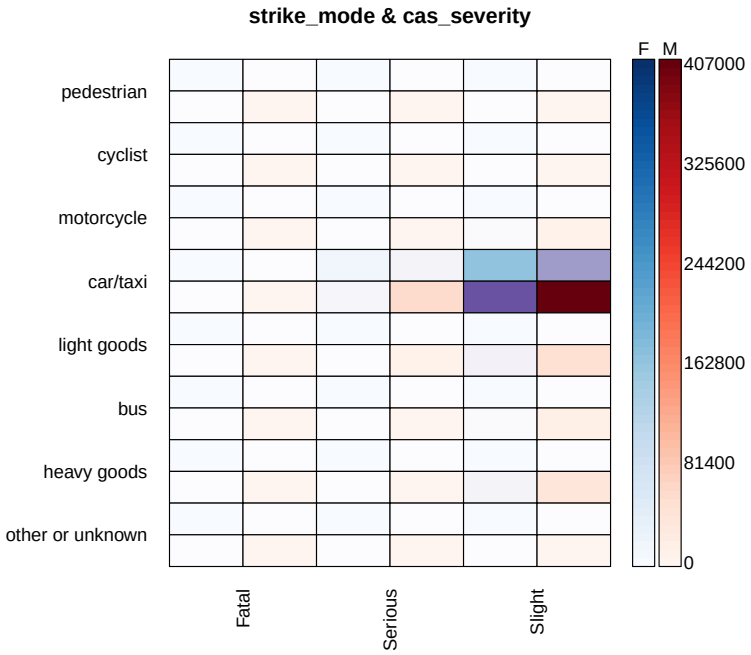

Figure 20: Number of injuries per casualty severity and striker mode.

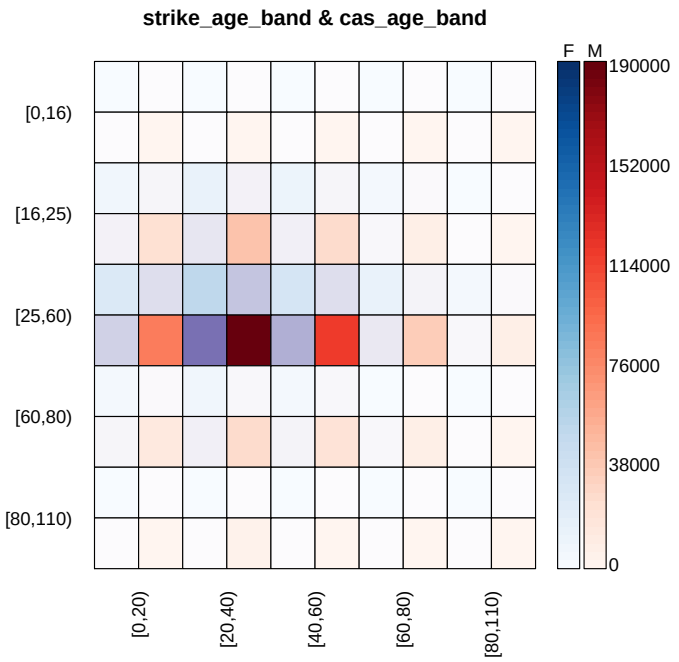

Figure 21: Number of injuries per casualty age and striker age.

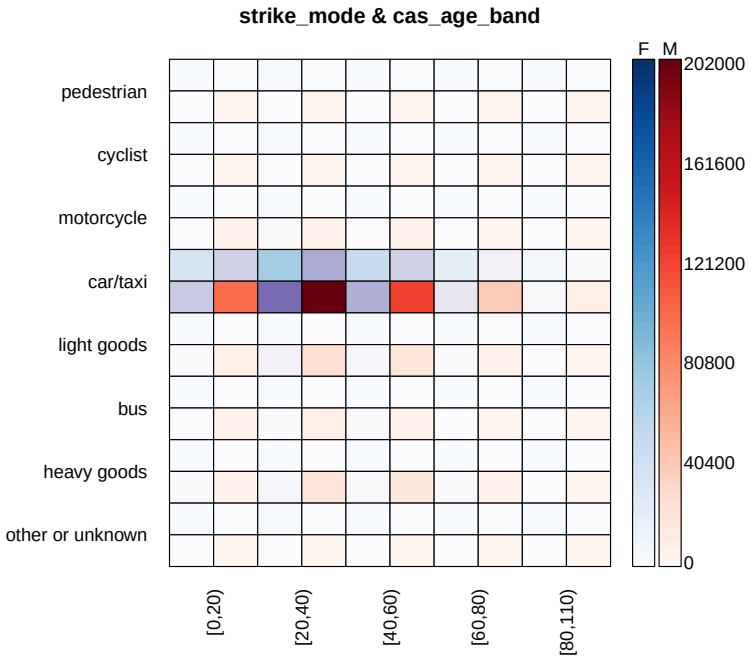

Figure 22: Number of injuries per casualty age and striker mode.

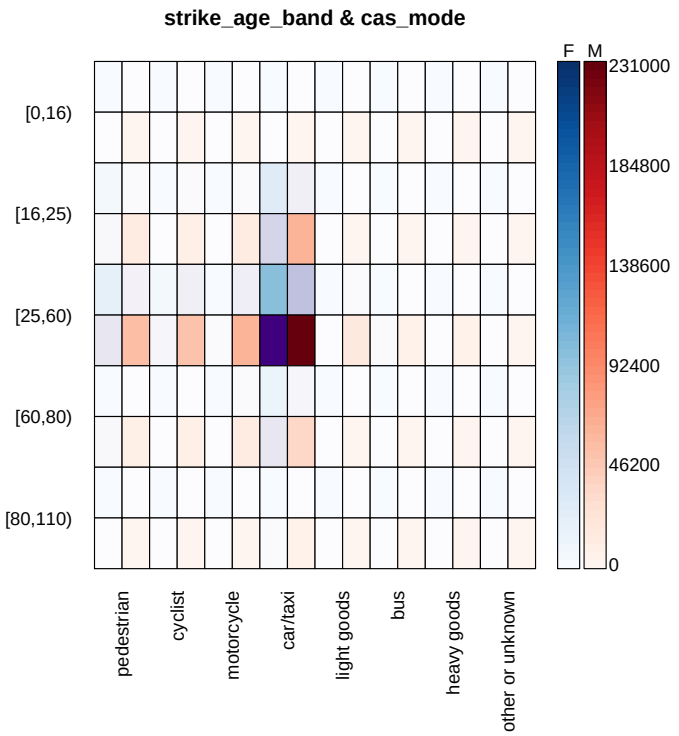

Figure 23: Number of injuries per casualty mode and striker age.

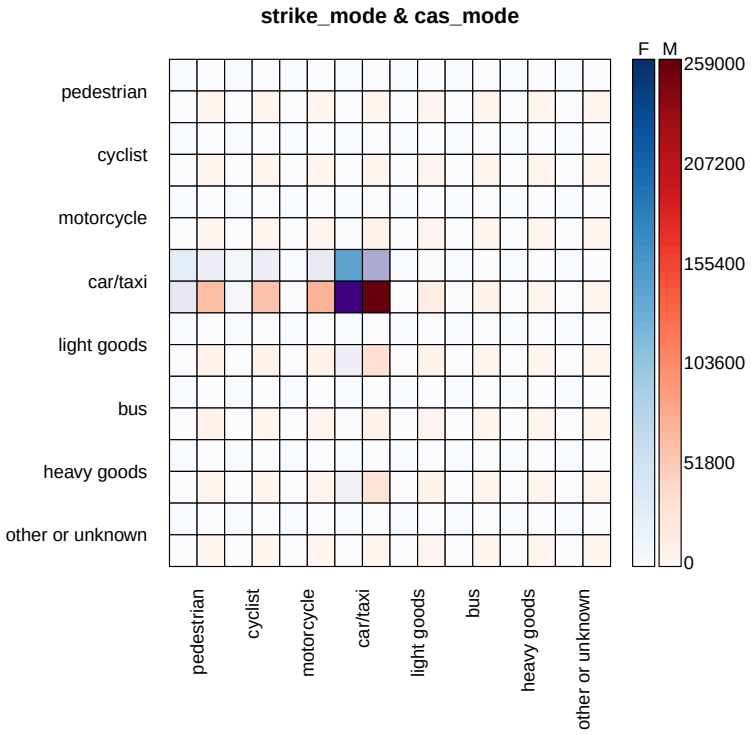

Figure 24: Number of injuries per casualty mode and striker mode.

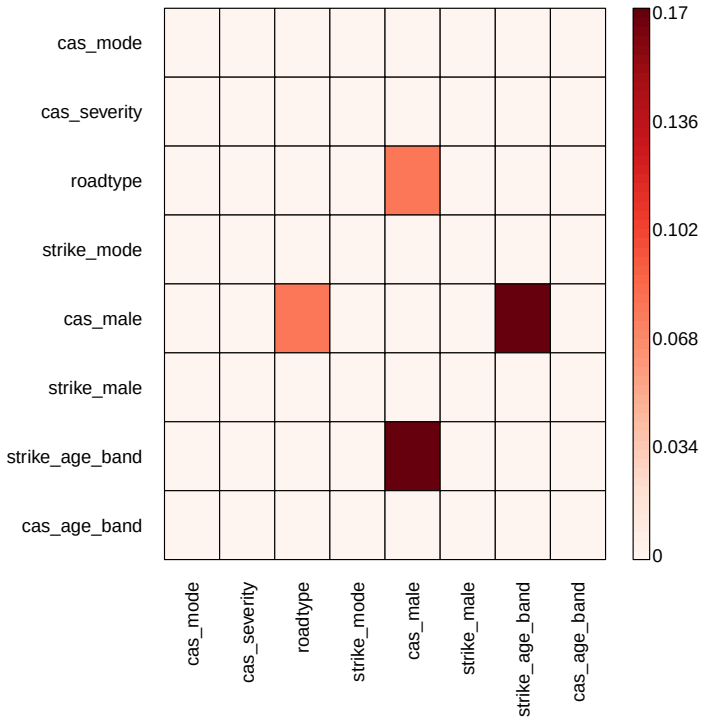

Figure 25: χ^2 tests for marginal independence. Red hue indicates p value.

B Preliminary regression studies

Which one pairwise interaction most improves a model compared to a main-effects-only model in terms of AIC? Figure 26 shows AICs for regressions with eight main effects and one first order interaction term. The term that offers most improvement is casualty gender and casualty mode. The two variables that seem to offer the most gains are casualty gender and strike mode.

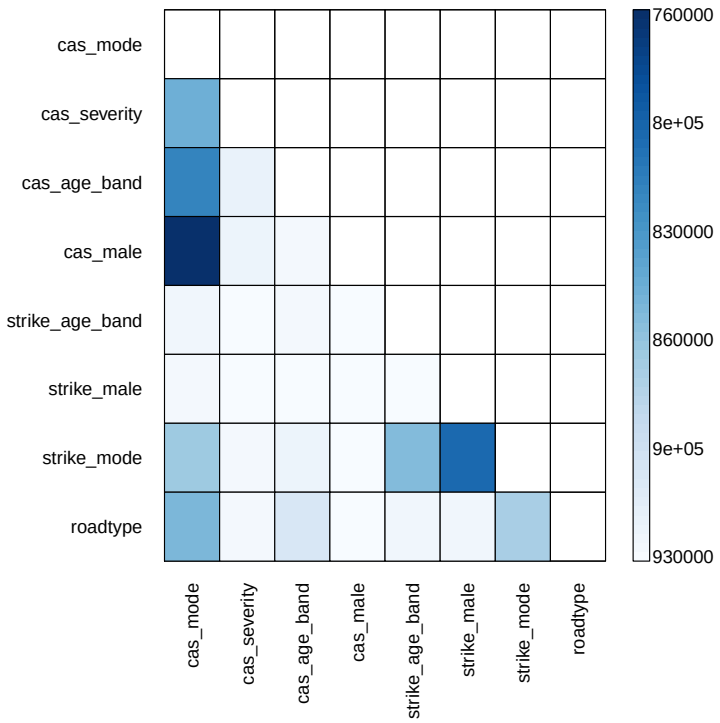

Figure 26: AIC values for single interactions in a model with all eight main effects.

Explicitly including the term for casualty gender and casualty mode and recreating the heatmap leads to the same pattern, and likewise with the next largest, striker gender and strike mode, followed by casualty age band and casualty mode.

The analogous study for three-way interactions is shown in Figure 27. I compared the difference between the 56 three-way hierarchical models, and the corresponding 56 models with only three two-way interactions and no three-way interaction. The only model for which the latter was better than the former was casualty severity, striker age band and road type. The greatest gain, of more than 5000, was for road type, casualty age band and casualty mode.

These studies motivate the method outlined in Section 2.1.

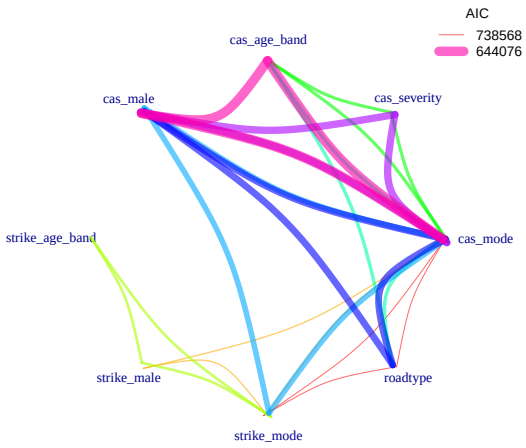

Figure 27: AIC values for three-way interactions in a model with all eight main effects. They are hierarchical models, so three two-way interactions are implicitly included. The top 11 two-way interactions are shown, with line widths indicative of the AIC.

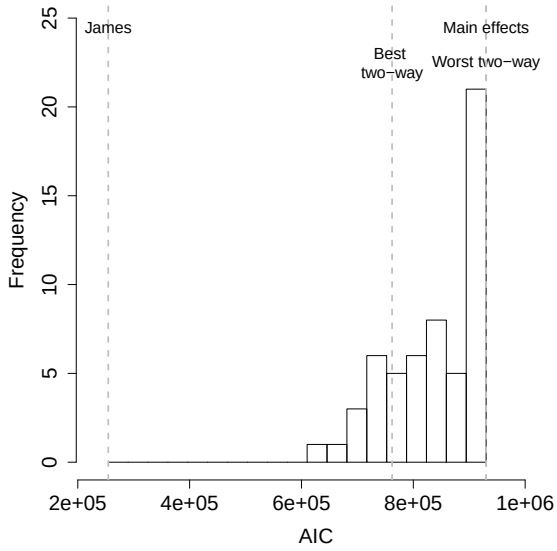

Figure 28: AIC values for three-way interactions in a model with all eight main effects. They are hierarchical models, so three two-way interactions are implicitly included. The top 11 two-way interactions are shown in Figure 27. Only 13 of the 56 three-way models have a better AIC than the best two-way model. The two-way model includes an interaction between casualty gender and casualty mode; this interaction is implicitly included in six of the 56 three-way models.

C Model-building results tables

Table 6: Building the regression model interaction by interaction for eight covariates (see Section 2.2). Upon each iteration, the interaction that most minimises the AIC of the present model is added to the model. After the eighth iteration we have the possibility for a three-way interation.

Iteration	Covariates	AIC
1	cas_male, cas_mode	762040
2	strike_mode, strike_male	632000
3	cas_age_band, cas_mode	517859
4	cas_severity, cas_mode	434537
5	roadtype, cas_mode	356169
6	strike_mode, strike_age_band	282395
7	strike_mode, cas_mode	217995
8	roadtype, strike_mode	172991
9	roadtype, cas_age_band	159051
10	cas_age_band, cas_severity	147154
11	strike_mode, cas_severity	139292
12	strike_age_band, cas_mode	131688
13	roadtype, cas_age_band, cas_mode	124666
14	strike_age_band, cas_age_band	119269
15	strike_mode, cas_age_band	115187
16	roadtype, strike_mode, cas_mode	111426
17	roadtype, strike_age_band	108551
18	strike_mode, cas_age_band, cas_mode	105731
19	strike_male, strike_age_band	103262
20	roadtype, strike_male	101360
21	cas_male, cas_severity	99552
22	cas_age_band, cas_severity, cas_mode	97880
23	strike_male, cas_age_band	96594
24	roadtype, cas_severity	95470
25	roadtype, cas_severity, cas_mode	93887
26	strike_mode, strike_age_band, cas_age_band	92790
27	strike_male, cas_male	91754
28	roadtype, strike_mode, strike_age_band	90818
29	strike_mode, strike_age_band, cas_mode	89928
30	strike_male, cas_mode	89074
31	strike_male, cas_age_band, cas_mode	87874
32	roadtype, strike_male, strike_age_band	87039
33	cas_male, cas_age_band	86266
34	cas_male, cas_age_band, cas_mode	82839
35	cas_male, cas_age_band, cas_severity	81973
36	strike_mode, cas_severity, cas_mode	81216
37	strike_male, cas_severity	80614
38	strike_age_band, cas_age_band, cas_mode	80031
39	strike_mode, strike_male, strike_age_band	79563
40	strike_age_band, cas_male	79149
41	roadtype, cas_male	78793
42	roadtype, cas_male, cas_age_band	78045
43	roadtype, strike_age_band, cas_mode	77701
44	strike_mode, cas_male	77413

45	strike_mode, cas_male, cas_mode	76747
46	strike_male, strike_age_band, cas_age_band	76503
47	strike_mode, strike_male, cas_mode	76284
48	roadtype, cas_male, cas_mode	76079
49	strike_age_band, cas_severity	75898
50	strike_age_band, cas_severity, cas_mode	75335
51	strike_male, strike_age_band, cas_mode	75149
52	strike_mode, cas_age_band, cas_severity	74974
53	cas_male, cas_severity, cas_mode	74839
54	strike_mode, strike_age_band, cas_age_band, cas_mode	74708
55	roadtype, strike_male, cas_mode	74582
56	roadtype, cas_male, cas_age_band, cas_mode	74458
57	strike_mode, cas_male, cas_age_band	74349
58	roadtype, strike_mode, cas_severity	74252
59	strike_male, cas_male, cas_age_band	74162
60	roadtype, strike_mode, cas_age_band	74083
61	roadtype, strike_age_band, cas_age_band	73997
62	strike_male, cas_severity, cas_mode	73919
63	strike_male, cas_male, cas_mode	73852
64	strike_male, cas_male, cas_age_band, cas_mode	73784
65	strike_mode, strike_age_band, cas_male	73721
66	cas_male, cas_age_band, cas_severity, cas_mode	73661
67	strike_male, strike_age_band, cas_male	73616
68	strike_mode, strike_male, cas_age_band	73571
69	strike_age_band, cas_male, cas_mode	73530
70	roadtype, strike_mode, cas_male	73490
71	strike_male, strike_age_band, cas_severity	73451
72	strike_mode, strike_male, cas_male	73415
73	strike_mode, strike_age_band, cas_severity	73379
74	roadtype, strike_mode, strike_male	73346
75	strike_mode, strike_male, cas_male, cas_mode	73314
76	strike_age_band, cas_male, cas_age_band	73287
77	strike_male, strike_age_band, cas_male, cas_age_band	73262
78	roadtype, cas_age_band, cas_severity	73242
79	roadtype, cas_age_band, cas_severity, cas_mode	73208
80	strike_mode, strike_male, cas_severity	73192
81	roadtype, strike_male, cas_age_band	73179
82	roadtype, strike_male, cas_age_band, cas_mode	73139
83	strike_male, cas_age_band, cas_severity	73130
84	roadtype, cas_male, cas_severity	73121
85	strike_mode, cas_male, cas_severity	73110
86	roadtype, cas_male, cas_age_band, cas_severity	73105
87	roadtype, strike_age_band, cas_male	73100
88	strike_male, cas_male, cas_severity	73101
89	strike_male, cas_male, cas_severity, cas_mode	73099
90	roadtype, cas_male, cas_severity, cas_mode	73099
91	strike_mode, strike_male, cas_male, cas_severity	73100
92	roadtype, strike_male, cas_male	73103
93	roadtype, strike_male, cas_severity	73107

94	roadtype, strike_male, cas_male, cas_severity	73110
95	roadtype, strike_mode, strike_male, cas_male	73114
96	roadtype, strike_male, strike_age_band, cas_male	73116

C.1 Results: seven covariates

James' model has the following code: `cas_mode*cas_male*cas_severity*(strike_mode + cas_age_band) + strike_age_band*strike_mode*(strike_male + cas_mode)`. This has the following four-, three-, and two-way interactions:

1. Four-way
   - Casualty mode, casualty gender, casualty severity, strike mode
   - Casualty mode, casualty gender, casualty severity, casualty age band
2. Three-way (in addition to those implied by four-way)
   - Striker age band, strike mode, striker gender
   - Striker age band, strike mode, casualty mode
3. No two-way interactions in addition to those implied by three- and four-way interactions.

We compare this to the greedy-algorithm model, listed in Table 7 and shown in Figure 29. The 48-factor model includes 21 two way interactions, 26 three-way interactions, and one four-way interaction.

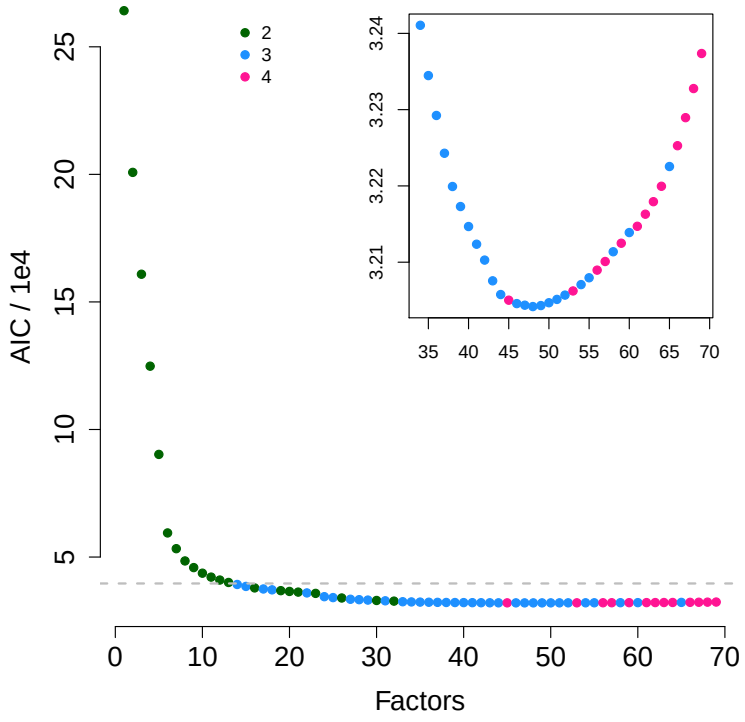

Figure 29: AIC values as the model grows. Each point is an additional interaction in the model. Points are added to the model from left to right. Green indicates a two-variable interaction; blue, a three-variable interaction. Grey line: James' model. Inset: factors 34 and onward.

Table 7: Building the regression model interaction by interaction for seven covariates (see Section C.1). Upon each iteration, the interaction that most minimises the AIC of the present model is added to the model. After the eighth iteration we have the possibility for a three-way interation.

Iteration	Covariates	AIC
1	cas_male, cas_mode	264157
2	strike_mode, strike_male	200785
3	cas_severity, cas_mode	160866
4	strike_mode, cas_mode	124823
5	cas_age_band, cas_mode	90238
6	strike_mode, strike_age_band	59459
7	cas_age_band, cas_severity	53288
8	strike_mode, cas_severity	48477
9	strike_age_band, cas_mode	45875
10	strike_age_band, cas_age_band	43682
11	strike_mode, cas_age_band	42142
12	cas_male, cas_severity	40994
13	strike_male, strike_age_band	40044
14	cas_age_band, cas_severity, cas_mode	39257
15	strike_mode, cas_age_band, cas_mode	38495
16	strike_male, cas_male	37905
17	strike_mode, cas_severity, cas_mode	37492
18	strike_mode, strike_age_band, cas_mode	37151
19	strike_male, cas_age_band	36830
20	strike_male, cas_severity	36524
21	strike_male, cas_mode	36267
22	strike_male, cas_age_band, cas_mode	35940
23	cas_male, cas_age_band	35761
24	cas_male, cas_age_band, cas_mode	34474
25	cas_male, cas_age_band, cas_severity	34115
26	strike_mode, cas_male	33956
27	strike_mode, cas_male, cas_mode	33460
28	strike_mode, strike_male, cas_mode	33298
29	strike_mode, strike_age_band, cas_age_band	33140
30	strike_age_band, cas_male	32984
31	strike_mode, strike_male, strike_age_band	32853
32	strike_age_band, cas_severity	32734
33	strike_age_band, cas_severity, cas_mode	32491
34	strike_age_band, cas_age_band, cas_mode	32411
35	strike_male, cas_severity, cas_mode	32345
36	cas_male, cas_severity, cas_mode	32292
37	strike_mode, cas_age_band, cas_severity	32243
38	strike_male, strike_age_band, cas_age_band	32199
39	strike_mode, cas_male, cas_age_band	32173
40	strike_male, strike_age_band, cas_mode	32147
41	strike_male, strike_age_band, cas_male	32124
42	strike_male, cas_male, cas_mode	32103
43	strike_male, cas_male, cas_age_band	32076
44	strike_male, strike_age_band, cas_severity	32058

45	strike_male, cas_male, cas_age_band, cas_mode	32050
46	strike_mode, strike_male, cas_severity	32046
47	strike_male, cas_age_band, cas_severity	32044
48	strike_mode, strike_male, cas_male	32042
49	strike_mode, strike_age_band, cas_male	32043
50	strike_male, cas_male, cas_severity	32047
51	strike_age_band, cas_male, cas_mode	32051
52	strike_mode, cas_male, cas_severity	32057
53	strike_male, cas_male, cas_severity, cas_mode	32062
54	strike_mode, strike_male, cas_age_band	32071
55	strike_mode, strike_age_band, cas_severity	32080
56	strike_mode, strike_male, cas_male, cas_severity	32090
57	strike_male, cas_male, cas_age_band, cas_severity	32101
58	strike_age_band, cas_male, cas_severity	32114
59	strike_male, strike_age_band, cas_male, cas_severity	32125
60	strike_age_band, cas_male, cas_age_band	32139
61	strike_male, strike_age_band, cas_male, cas_age_band	32147
62	cas_male, cas_age_band, cas_severity, cas_mode	32163
63	strike_mode, strike_male, cas_male, cas_mode	32179
64	strike_mode, strike_male, cas_male, cas_age_band	32200
65	strike_age_band, cas_age_band, cas_severity	32225
66	strike_male, strike_age_band, cas_male, cas_mode	32253
67	strike_age_band, cas_male, cas_age_band, cas_severity	32290
68	strike_mode, strike_male, strike_age_band, cas_male	32328
69	strike_age_band, cas_male, cas_age_band, cas_mode	32374

James' models' AICs are shown as grey dashed lines in Figures 3 and 29. Their values are 95,550 and 39,641, respectively. They intersect the greedy models at 24 and 14 factors, respectively. So which interactions are included up to these points that are omitted from James' model?

The first 24 eight-covariate interactions include the following that are excluded from James' model:

- striker age & casualty age
- strike mode & casualty age
- strike mode, casualty mode & casualty age
- striker gender & casualty age

The first 14 seven-covariate interactions include the following that are excluded from James' model:

- striker age & casualty age
- strike mode & casualty age

Table 8: Building the regression model interaction by interaction for eight covariates, of which one is modelled with a spline (see Section 3.1). Upon each iteration, the interaction that most minimises the AIC of the present model is added to the model.

Iteration	Covariates	AIC

1	cas_mode, cas_age	1303126
2	cas_male, cas_mode	1134981
3	strike_mode, strike_male	1004940
4	cas_severity, cas_mode	921618
5	roadtype, cas_mode	843250
6	strike_mode, strike_age_band	769476
7	strike_mode, cas_mode	705076
8	roadtype, strike_mode	660072
9	roadtype, cas_age	645045
10	cas_severity, cas_age	631786
11	roadtype, cas_mode, cas_age	619173
12	strike_mode, cas_severity	611303
13	strike_age_band, cas_mode	603699
14	strike_age_band, cas_age	597518
15	strike_mode, cas_age	592279
16	strike_mode, cas_mode, cas_age	588580
17	roadtype, strike_mode, cas_mode	585200
18	strike_male, strike_age_band	582730
19	roadtype, strike_age_band	579833
20	roadtype, strike_male	577932
21	cas_male, cas_severity	576123
22	strike_male, cas_age	574434
23	roadtype, cas_severity	573229
24	cas_severity, cas_mode, cas_age	570892
25	strike_mode, strike_age_band, cas_age	569675
26	cas_male, cas_age	568621

D Penalised regression

D.1 Method

We would like to use the group-ℓ_1-penalty method with ten-fold cross validation via the R package `logilasso` (Dahinden et al., 2007). However, this methodology is specific to two-level covariates. The method will need to be adapted and re-implemented to model the covariates listed in Table 1. However, we run `logilasso` naively in order to see what outputs we might expect.

D.2 Results

The penalty is captured by λ and the covariates by β. The relationship between the two is shown in Figure 30. We choose some threshhold for β, and plot the resulting residual connections in Figure 31.

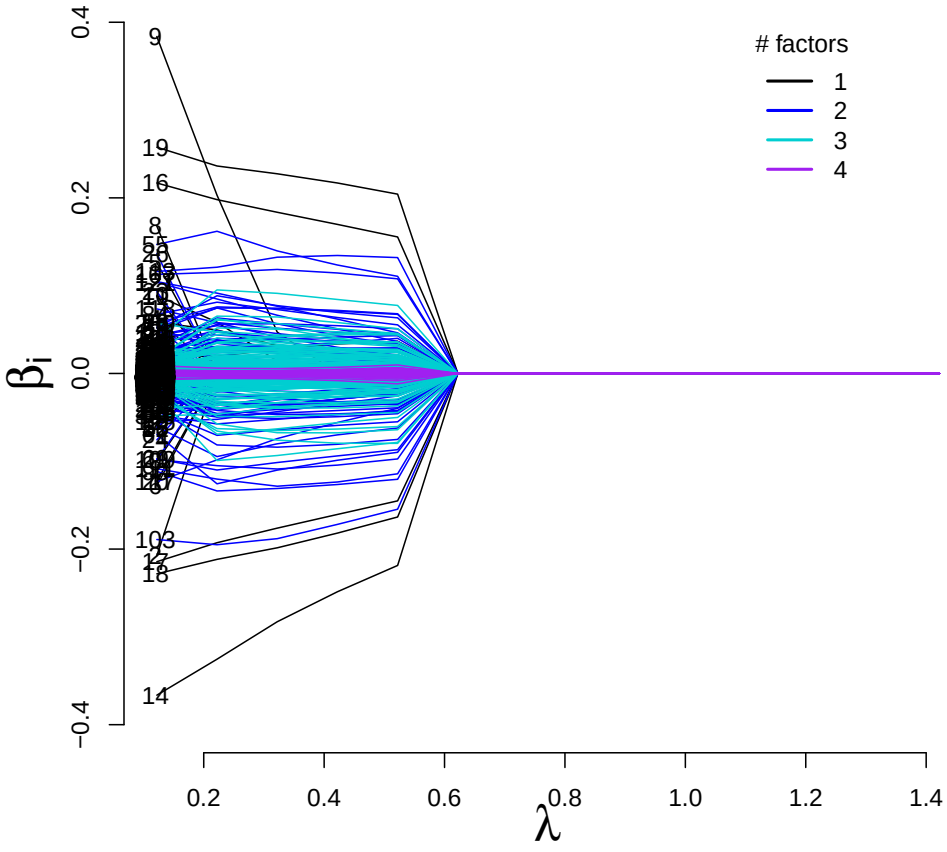

Figure 30: Each trajectory corresponds to one coeffient (β) in the model. The number on the left is its (numeric) index. The colour of the line corresponds to how many factors constitute the variable that multiplies the coefficient, e.g. black (indices 1–19) are main effects; blue (indices 20–128) are first-order interactions. On the x axis, the penalty term λ is varied.

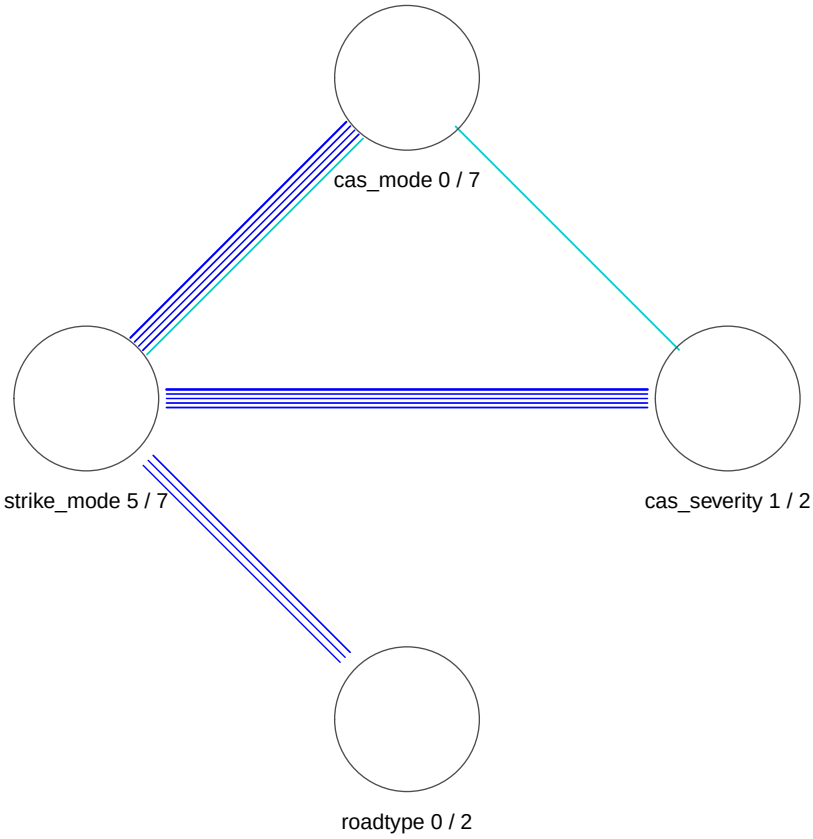

Figure 31: Network showing covariate combinations whose coefficient β exceeds some threshhold. Each node corresponds to a covariate. Next to the label names are two numbers (a / b), showing how many (a) of the total possible main effects (b) had a value of β greater than the threshold. Between the nodes are edges indicating an interaction effect whose coefficient exceeds the threshhold. Colours are the same as those in Figure 30. For example, cas_severity has 2 possible main effects, of which 1 has a coefficient greater than the threshold; cas_mode has 7 possible main effects, of which 4 have coefficients greater than the threshhold; between them, there are 14 possible interactions, of which 4 have coefficients greater than the threshhold.

E Bayesian

The Bayesian method of Ntzoufras et al. (2017) uses stochastic search variable selection. Not yet implemented.

F PReMiuM

Following Papathomas and Richardson (2016), we apply the variable selection function of PReMiuM (Liverani et al., 2015) to the dataset. We find the clustering process inconclusive (Figure 32), and the variables selected as shown in Figure 33. This is somewhat consisent with the results of Section D.

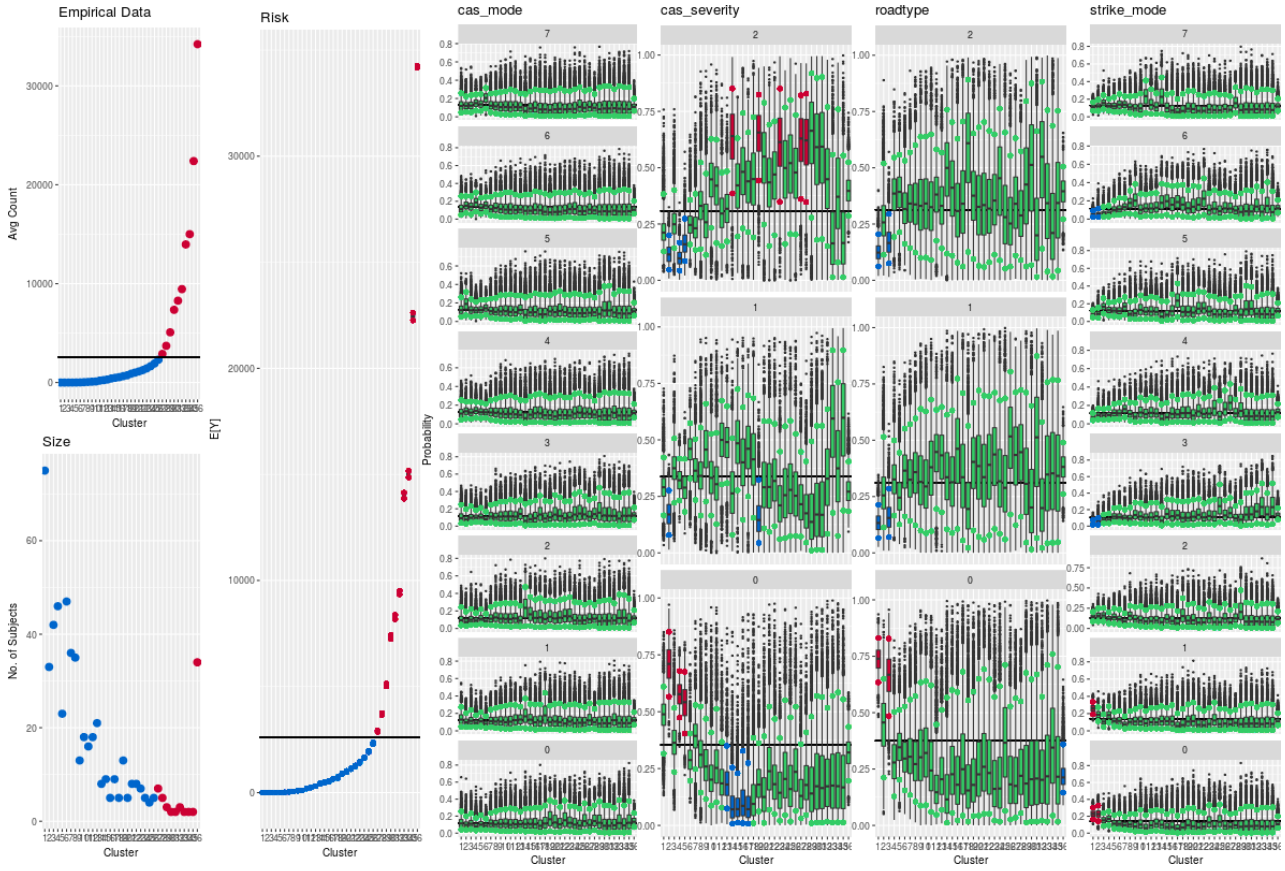

Figure 32: Summary of clustering from PReMiuM. There are many more clusters than can be usefully interpreted.

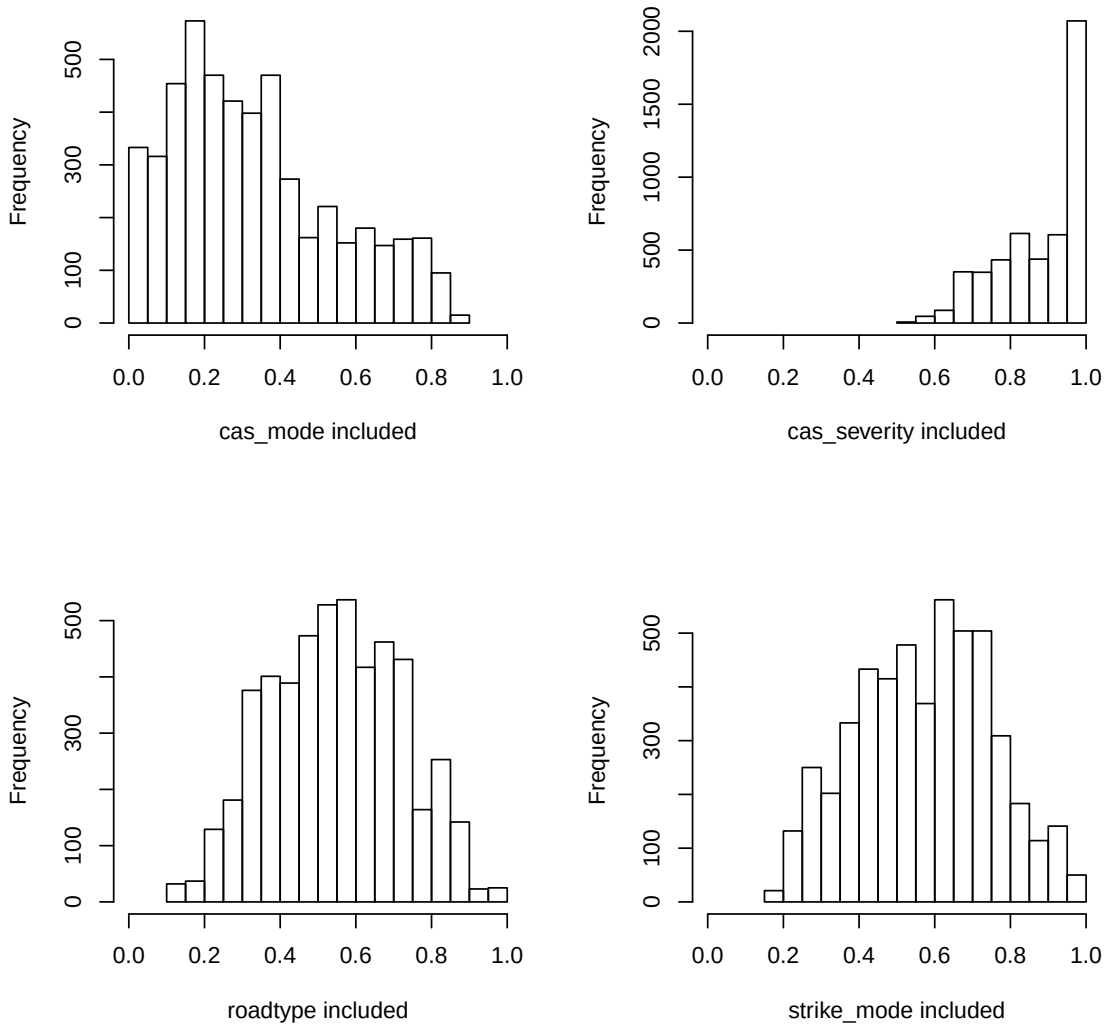

Figure 33: Results of variable selection in PReMiuM. It seems that casualty severity and road type are more selected than casualty mode and striker mode.

References

Dahinden, C., Parmigiani, G., Emerick, M. C. and Bühlmann, P. (2007), 'Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries', *BMC Bioinformatics* **8**(476).

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. and Richardson, S. (2015), 'PReMiuM: An R package for profile regression mixture models using Dirichlet processes.', *Journal of Statistical Software* **64**(7), 1–30.

Ntzoufras, I., Forster, J. J. and Dellaportas, P. (2017), 'Stochastic search variable selection for log-linear models for log-linear models', *Journal of Statistical Computation and Simulation* **61**(1), 23—-37.

Papathomas, M. and Richardson, S. (2016), 'Exploring dependence between categorical variables: benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms', *Journal of Statistical Planning and Inference* **173**, 47—-63.
URL: *http://dx.doi.org/10.1016/j.jspi.2016.01.002*