Crash risk relationships – 2000-2009 data

1 Introduction and main results **3**

1.1 Introduction *3*

1.2 Main results *4*

2 The data preparation **4**

2.1 Overview *4*

2.2 Initial processing *5*

2.3 Final assembly of the data *6*

3 The Data **7**

3.1 Year *7*

3.2 Region *7*

3.3 Urban-rural *9*

3.4 The crash data *10*

3.5 Estimated traffic average daily travel *12*

3.6 Curvature *14*

3.7 The Advisory Speed calculation *15*

3.8 Out-of-context-curve indicator (OOCC) *16*

3.9 SCRIM coefficient *18*

3.10 IRI (roughness) *19*

3.11 Adjusted IRI *20*

3.12 Gradient *22*

3.13 Skid-site and adjusted skid-site categories *22*

3.14 Lane width *23*

3.15 Rut mean and rut standard deviation *23*

3.16 Crossfall *24*

3.17 Texture depth *25*

4 The Poisson regression model **26**

4.1 The model *26*

4.2 Error estimates *27*

5 The model fit **28**

5.1 The regression analysis *28*

5.2 Analysis of variance tables *28*

5.2.1 All casualty crashes 29

5.2.2 Wet casualty crashes 29

5.2.3 Selected casualty crashes 30

5.2.4 Wet selected casualty crashes 30

5.3 Predicted crash rate graphs *31*

5.3.1 All casualty crashes - year 32

5.3.2 All casualty crashes - region 32

5.3.3 All casualty crashes – urban/rural 33

5.3.4 All casualty crashes – adjusted skid-site 33

5.3.5 All casualty crashes – OOCC 34

5.3.6 All casualty crashes – curvature 34

5.3.7 All casualty crashes – OOCC and curvature 35
5.3.8 All casualty crashes – ADT 35
5.3.9 All casualty crashes – scrim 36
5.3.10 All casualty crashes – gradient 36
5.3.11 All casualty crashes – adjusted iri 37
5.3.12 All casualty crashes – adjusted iri with tps interaction 38
5.3.13 Wet casualty crashes – region 39
5.3.14 Wet casualty crashes – OOCC 40
5.3.15 Wet casualty crashes – curvature 40
5.3.16 Wet casualty crashes – scrim 41
5.3.17 Wet casualty crashes – adjusted iri 41
5.3.18 Selected casualty crashes – urban/rural 42
5.3.19 Selected casualty crashes – adjusted skid-site 43
5.3.20 Selected casualty crashes – scrim 43
5.3.21 Selected casualty crashes – adjusted iri 44
5.3.22 Selected casualty crashes – adjusted iri with tps interaction 45
5.3.23 Wet selected casualty crashes - scrim 46
5.3.24 Wet selected casualty crashes – adjusted iri 46

5.4 The coefficients and calculation of predicted crash rates 47

5.5 Comparison of fitted and observed counts 48

5.6 Effect of the averaging 52

6 Application of the model 52

6.1 What-if study for skid resistance 52

6.2 What-if study for IRI 54

7 Variations on model 55

7.1 Include curvature × SCRIM interaction 55
7.1.1 All crashes 55
7.1.2 Wet crashes 56

7.2 Include texture and interaction 57

7.3 Include mean rut depth plus interaction 58

7.4 Include year × region interaction 58

7.5 Serious and fatal crashes 59
7.5.1 All serious/fatal crashes 60
7.5.2 Wet serious/fatal crashes 62
7.5.3 Fit test 63

8 IRI variance terms 63

9 Curve model 68

10 Discussion 71

1 Introduction and main results

1.1 Introduction

This is an update of my reports of 2004 and 2009[1] which relate vehicle crash rates to road surface characteristics using data from 1997 to 2002. We now have crash data for 1997 to 2009. The survey data is incomplete for 1997 and 1999 so I am limiting the analysis to the years 2000 to 2009.

As in the previous reports I am fitting a Poisson regression model to predict the crash rate from the road surface data, particularly *curvature*, *skid resistance* and *roughness*. But in this report, I am particularly concerned with the effect of *roughness*. However, since the approach is to fit the model of the earlier papers I also look at the effects of the other variables, particularly *skid-resistance*.

The approach is the same as that of the 2004 report except that I am omitting the preliminary analyses and include the out-of-context-curve effect described in the 2008 update[2].

Most of this report is concerned with vehicle crashes in which at least one person is killed or suffers serious or minor injuries. I refer to these as *casualty crashes* or simply *crashes*. In section 7.5 I consider crashes in which at least one person is killed or suffers serious injuries. I call these *serious/fatal* crashes. In all cases *crashes* means *reported crashes*.

Currently, I do not have the complete 2009 crash data and the results here will need to be updated when the complete data is available.

The main results from this report are summarised in section 1.2 below.

Section 2 gives an overview of the data preparation and section 3 describes the data used in the analyses particularly by presenting histograms of the data.

Section 4 describes the Poisson model and section 5 gives the results of the model fit. Spreadsheets accompanying this report give the coefficients from the fitted models.

Section 6 carries out two "what-if" analyses to see the effect of changes in either skid resistance or roughness on the selected parts of the network in 2008.

---

[1] see web pages http://www.robertnz.net/pdf/crashrisk8.pdf and http://www.robertnz.net/pdf/report_curves3.pdf

[2] http://www.robertnz.net/pdf/OOCC_report2.pdf

Section 7 looks at a number of variations on the model including looking at additional variables and rerunning the model with only serious and fatal crashes.

Section 8 runs the model with the IRI variable replaced by IRI variance for various wavelengths and section 9 reruns the curve model[3] with the updated data.

Finally the results are discussed in detail in section 10.

1.2 Main results

- The results are very similar to those from the previous studies with important predictors of crash rate including out-of-context-curve effect, curvature, ADT, skid resistance, roughness.
- There is still quite a high unexplained variability in the data.
- The effect of roughness depends on curvature, the effect being strongest on curves with radii of curvature in the range 500 to 5000 metres.
- Considering the IRI variances measurements, the IRI variance for wavelength 10 metre is a better predictor than the variances for wavelengths 3 or 30 but is only slightly better, if at all, than ordinary IRI.
- The skid resistance effect is stronger for crashes on wet roads rather than for all crashes and there is some evidence that it is stronger on curves than on straight or near straight roads.
- Where I have repeated the analyses using only serious/fatal crashes, I have found results very similar to those for all casualty crashes.
- Including a year × region interaction has little influence on the results.
- Rerunning the curve model of 2009 with the new data yields very similar results to the original 2009 analysis. These is some evidence for an effect of roughness for the curves with greater radii of curvature.

2 The data preparation

2.1 Overview

We have the following sets of data:

- Data collected by the SCRIM+ machine at 10 metre intervals on each side of the road
    - Geometry: gradient; curvature; crossfall; GPS coordinates (2010 only)
    - Scrim coefficient for each wheel path; skid_event
    - Mean texture depth for each wheel path

[3] http://www.robertnz.net/pdf/report_curves3.pdf

- Data collected by the SCRIM+ machine at 20 metre intervals on each side of the road
    - IRI roughness for left and right wheel paths; IRI variances (2006-9 only)
    - Mean rut depth and standard deviation of rut depth and related measurements
- Carriageway data: urban/rural; number of lanes; lane width; estimated traffic ADT
- Road names: state highway number; road region
- Crash data:
    - Crash location (two versions); crash details; movement code; etc
    - Crash vehicle details
    - Crash causes
- Survey number to year correspondence (also survey vehicle type)

We can think of the state-highway network being divided up into state-highways. These, in turn, are divided into roads. The roads have a unique name and unique identification number. The correspondence between the names and identification numbers is given in the *road names* table. Road names are important because they give us the order of the roads on a state-highway. The roads are divided into 10 metre segments for which we have the *SCRIM+ data*. The roads are also divided into longer segments for which we have the *carriageway data*.

The data was initially loaded into a MySQL database for initial processing. This processing is described in section 2.2 below.

2.2 Initial processing

Initial processing was carried out using the MySQL database program.

Each line in each SCRIM+ table has a unique identity value and the following variables

survey_number; *road_id*; *start_m*; *end_m*; *lane*

which identify the segment of road being surveyed.

For the geometry, scrim and texture tables we usually have *end_m* – *start_m* = 10 and *start_m* is a multiple of 10. Initial processing consisted of using the survey number to year correspondence table to establish the *survey_year* of each survey, rounding *start_m* and *end_m* to be multiples of 10 and rejecting lines where the rounded values didn't differ by exactly 10. Where there were duplicate measurements choose the one with the highest identity value.

For the roughness and rutting tables we usually have *end_m* – *start_m* = 20 and *start_m* is a multiple of 20. Initial processing consisted of using the survey number to year correspondence table to establish the *survey_year* of each survey, rounding *start_m* and *end_m* to be multiples of 10 and rejecting lines where the rounded values didn't differ by exactly either 10 or 20 or where *end_m* – *start_m* < 6. Where there were duplicate measurements choose the one with the highest identity value.

Create a base set of 10 metre road segments. This consists of all values of *survey_year*, *road_id* and the rounded version of *start_m* that appear in any of the 10 metre SCRIM+ variables. Join with the road names table. We can sort by the *road_name* variable to ensure that the sections of road identified by each *road_id* are in consecutive order along each state highway.

Now join in the 10 metre data to this base set with separate columns for the left and right lanes – choose only the data with lane = *L1* or *R1*.

The 2009-2010 survey data includes GPS locations. Find the locations of the beginning and end of each *road* as identified by the *road_id*.

Now considering the roughness and rutting data. Make table columns showing where, for each 10 metre segment in the base set, the corresponding roughness and rutting data values occur in the roughness and rutting tables. Where we have 20 metre roughness or rutting segments each value will typically be referenced twice in these columns. Where there are both 10 metre and 20 metre roughness or rutting segments that could correspond to a 10 metre segment in the base set, the 10 metre one is chosen preferentially. I don't know whether this situation actually arises in this dataset. The actual combining of the roughness and rutting data into the rest of the data is done in the subsequent analysis carried out by the C++ programs.

Now considering the crash data. The crash vehicle type data for each crash and crash causes for each crash are each amalgamated into single fields and joined in to the crash table. The linking of the crashes to the 10 metre data is carried out by the C++ programs.

2.3 Final assembly of the data

The C++ programs set up the data structures needed for carrying out the analyses described in the rest of this report, read in the data generated by MySQL, carry out some checking and generate the transformed data where required.

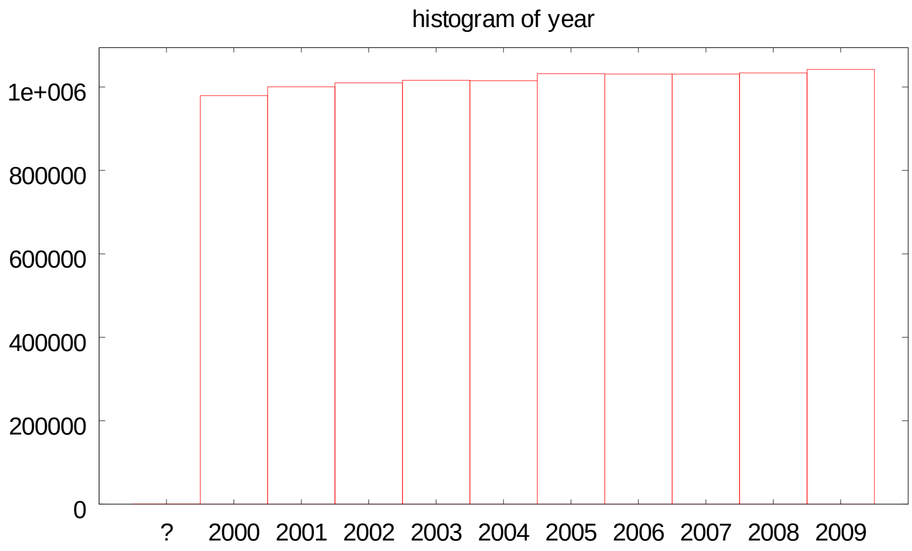
In particular, they link in the roughness, rutting, crash and road data and calculate the adjusted skid site, adjusted IRI, and the OOCC variables described in the section 3. They check for isolated missing values in the predictor variables and attempt to estimate these from neighbouring variables.

As noted in the previous section the state highways are divided into lengths of roads identified by their *road_id*s. It is important that any gaps between these lengths of roads be identified, since gaps will invalidate the averaging described in section 4.1 or the identification of the out of context curve effect described in section 3.8. The GPS data collected in the 2009-2010 as part of the geometry data is used to identify these gaps.

3 The Data

The following sections show histograms of the 10 metre data.

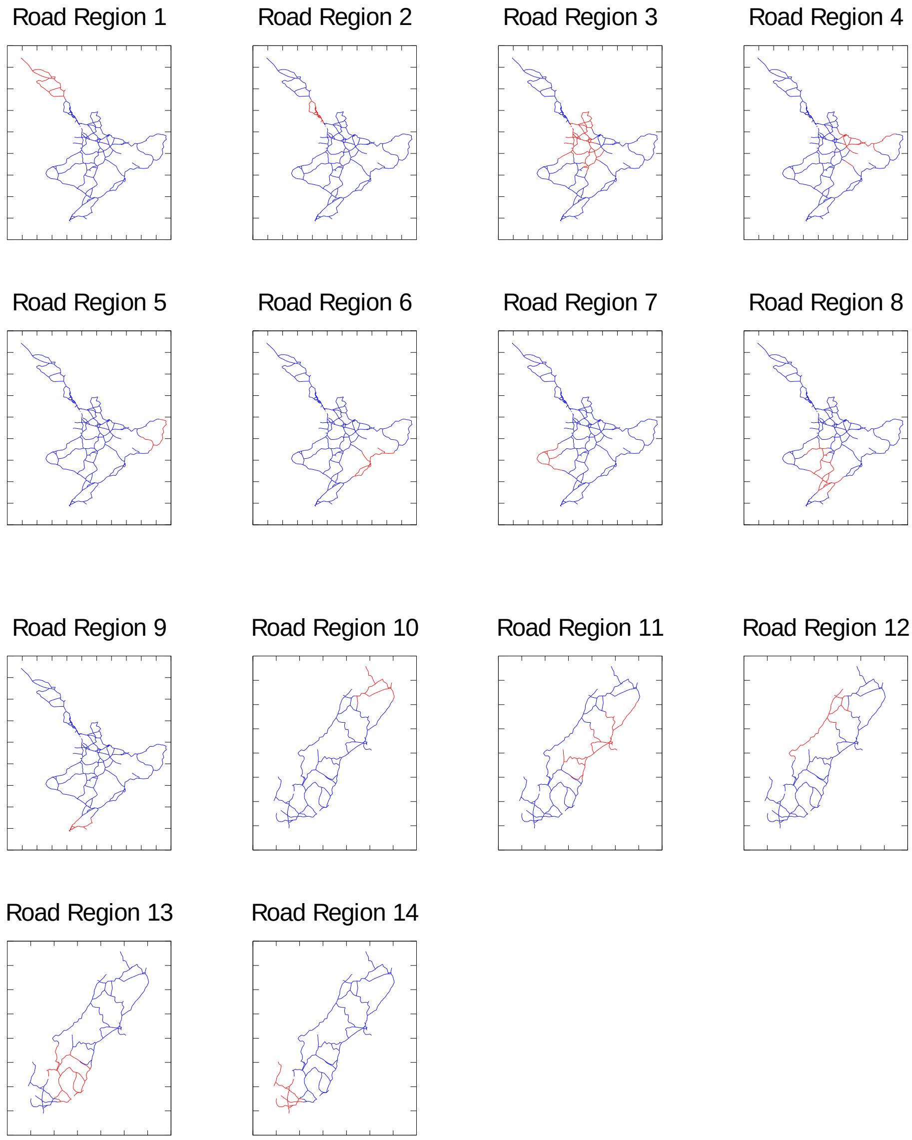
3.1 Year



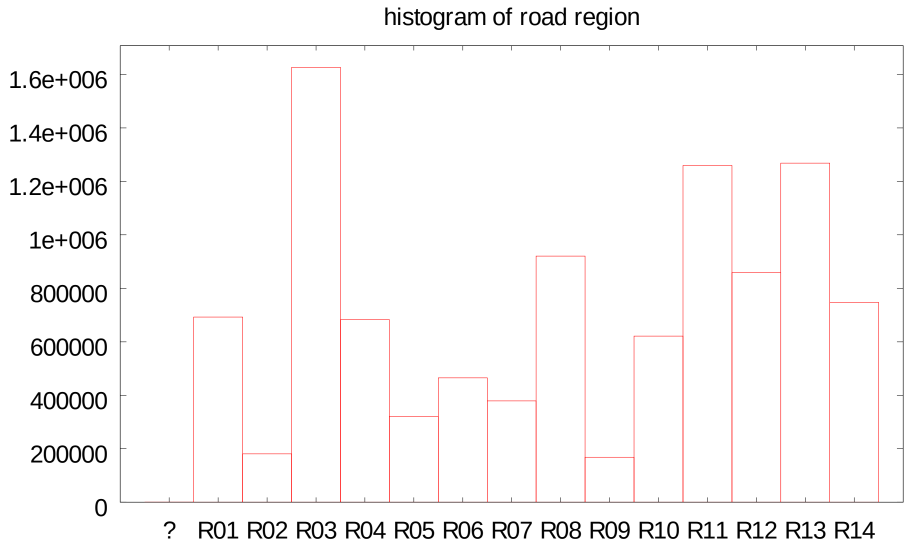
The histogram shows the number of ten metre segments surveyed each year. It is around 1 million and is increasing slightly as we go from 2000 to 2009. This could indicate an increase in the length of road surveyed or a reduction in the number of missing values. As noted in the introduction, I am limiting the analysis to the data from 2000 to 2009.

3.2 Region

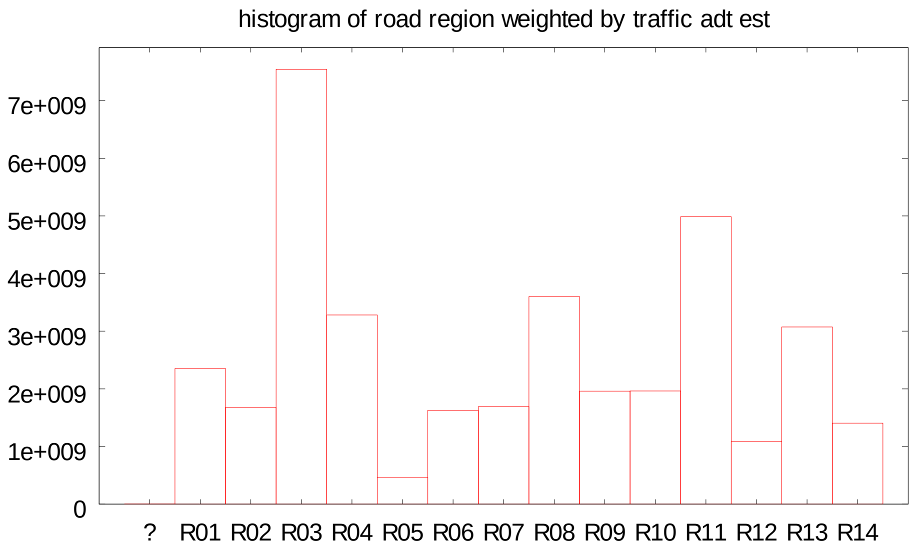
The state highway network is divided into 14 road regions. In the 2010 survey GPS locations of the roads were reported. This enables us to plot the locations of the road regions. North and South Islands are shown separately. Each graph shows the network in either the North or South Island with the roads corresponding to the region being graphed shown in red.



The following histogram shows the number of 10 metre segments in each region. Since we are looking at 10 years most segments will be counted ten times.



In the following graph the histogram has been weighted by the estimated traffic ADT.



The Auckland and Wellington regions (2 and 9) are relatively larger in the weighted histogram compared with the unweighted and East Cape (5) smaller.

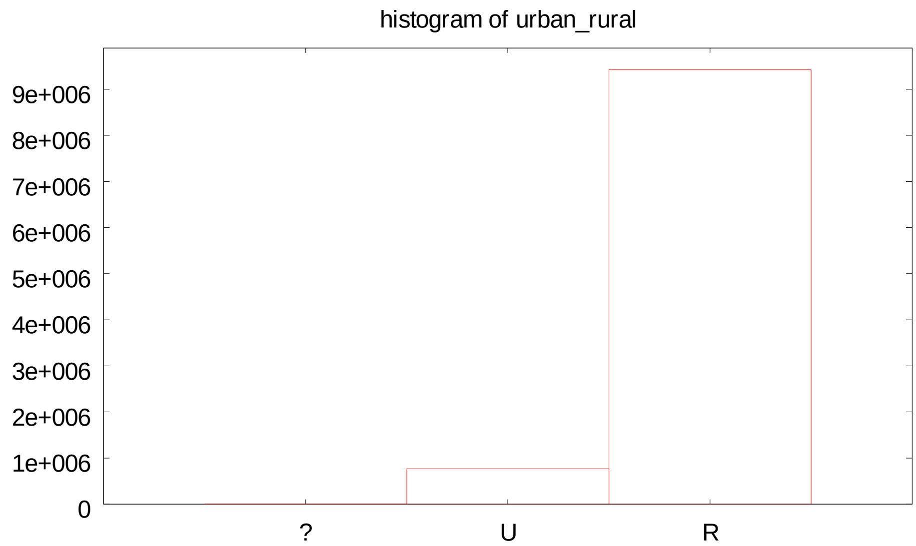
There is quite a lot of variability in the size of the road regions, whether or not they are weighted by the ADT. Nevertheless, they provide a convenient way of breaking the total network up into smaller regions. The differing sizes should not present a problem in the analyses.

3.3 Urban-rural

The roads are classified as *urban* or *rural*. *Rural* means a speed limit of more than 70 km an hour. So this is a very rough classification. I have included both rural and urban roads in the analysis but it possible that the analyses should have been limited to rural roads.

So, in this report *urban* means those parts of the state highway network where the speed limit is less than or equal to 70 km/hr. Hence the results don't have a lot of relevance to urban roads, in general.

The following histogram shows the most of the road length is *rural*.



3.4 The crash data

The analyses were applied to each of 4 subsets of the crash dataset described in the following table.

Group	**Criteria**
All	All casualty crashes
Wet	All casualty crashes with the road wet field being W or the cause code was 801, 823 or 901.
Selected	All casualty crashes with MVMT_IDA being one of A, B, C, D, F
Wet & selected	Satisfying both the wet and selected criteria

where the MVMT_IDA codes are described in the following table.

A	Overtaking and Lane Change
B	Head On
C	Lost Control or Off Road (Straight Roads)
D	Cornering
E	Collision with obstruction
F	Rear End
G	Turning Versus Same Direction
H	Crossing (No Turns)
J	Crossing (Vehicle Turning)
K	Merging
L	Right Turn Against

M	Manoeuvring
N	Pedestrians Crossing Road
P	Pedestrians Other
Q	Miscellaneous

The definitions of the relevant cause codes for the wet crashes are

Cause code	Contributing factor
801	Rain
823	Flood waters, large puddles, ford
901	Heavy rain

Cause 823 is additional to the ones used in the 2004 study.

Here are the numbers of crashes by year for the sections of road used in the analyses.

Year	All crashes	Wet crashes	Selected crashes	Wet selected crashes
2000	1800	525	1307	421
2001	2023	642	1425	506
2002	2274	654	1607	505
2003	2391	641	1726	518
2004	2379	740	1699	596
2005	2463	690	1809	569
2006	2619	728	1922	587
2007	2801	771	2079	631
2008	2554	658	1816	525
2009	1566	427	1126	349
Total	**22870**	**6476**	**16516**	**5207**

The numbers are low for 2009 because we don't have complete data for this year.

Unlike the data for the 2004 study most of the crashes have been assigned locations and the number of crashes which could not be located is not an issue. In the 2004 study about ¾ of the crashes were located.

If we limit the crashes to those which resulted in fatalities or serious injuries the numbers for all crashes and wet crashes become

Year	All crashes	Wet crashes
2000	729	201
2001	731	209
2002	742	202
2003	755	177
2004	735	219
2005	751	201

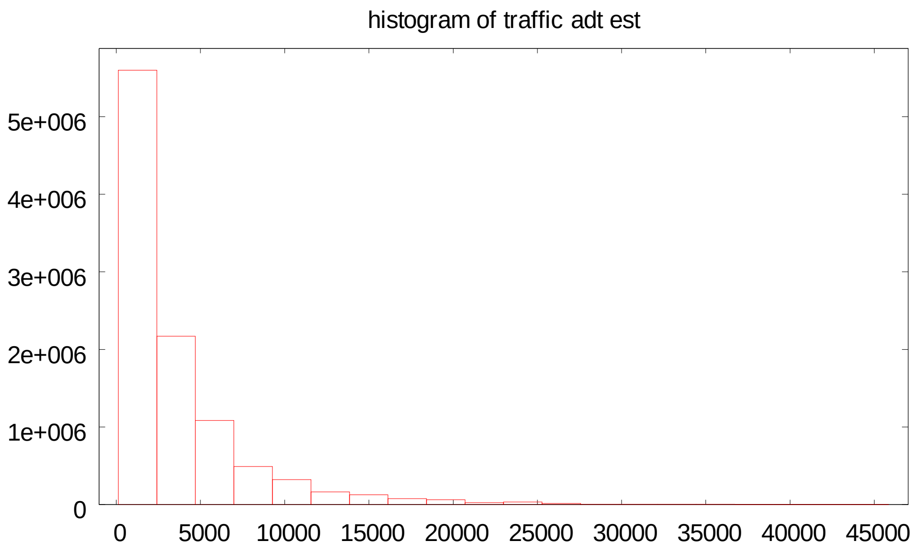
Year	All crashes	Wet crashes
2006	760	197
2007	759	182
2008	691	154
2009	435	111
Total	**7088**	**1853**

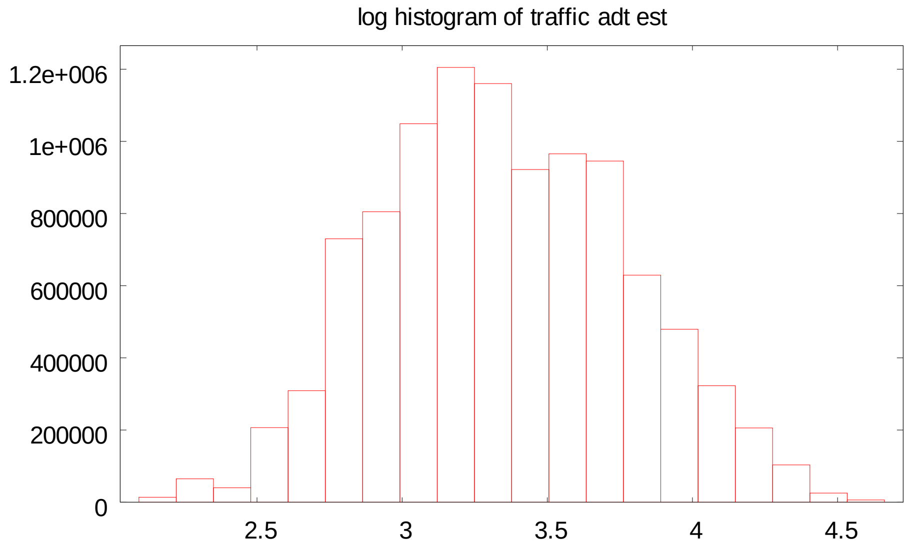
3.5 Estimated traffic average daily travel

I am using the estimated average daily travel for 2009. This will not be a problem for the analyses provided that the travel has changed by the same relative amount for each road over the 10 years of the study. This is unlikely to be exactly true and ideally we should be using the estimate for each year.

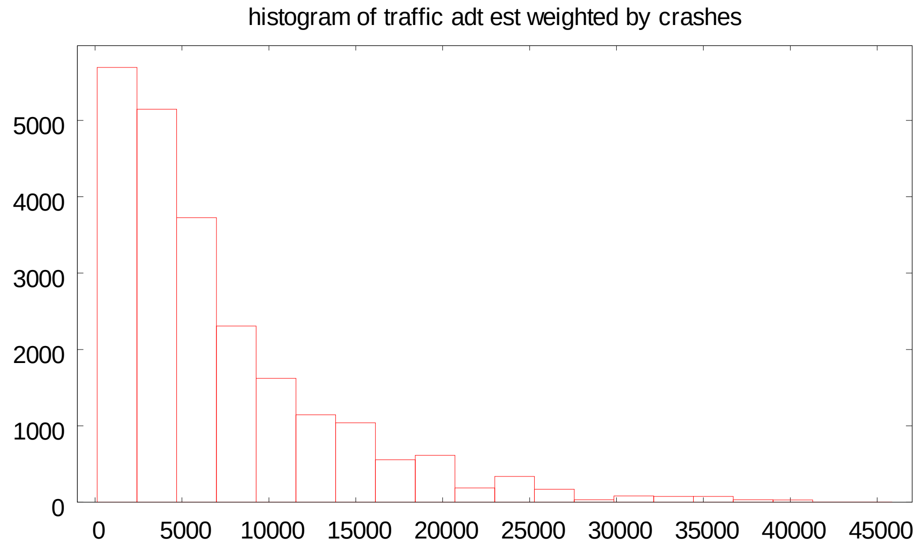
I am omitting the roads for the analysis where the ADT is less than 100.

The following histograms are for the estimated average daily travel (ADT). The second one is for $\log_{10}$(ADT).



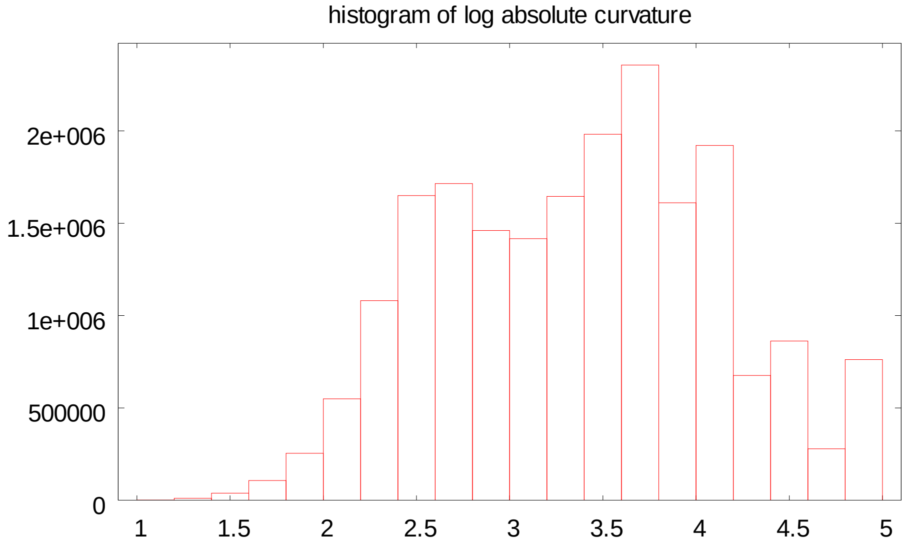


The following two graphs show the histogram of *traffic adt est* weighted by *traffic adt est* and by the number of crashes.



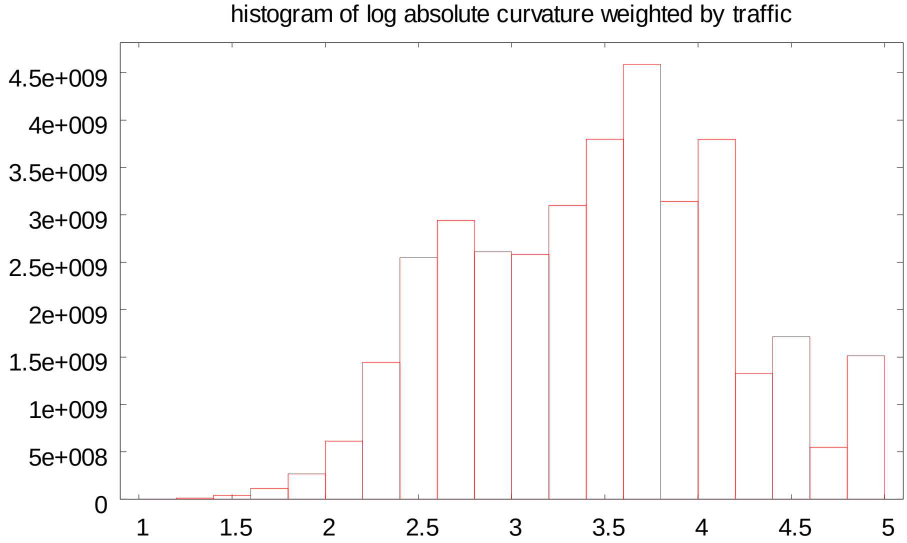
These show that the low ADT roads are still dominant even when you weight by ADT and even more so when you weight by crashes.

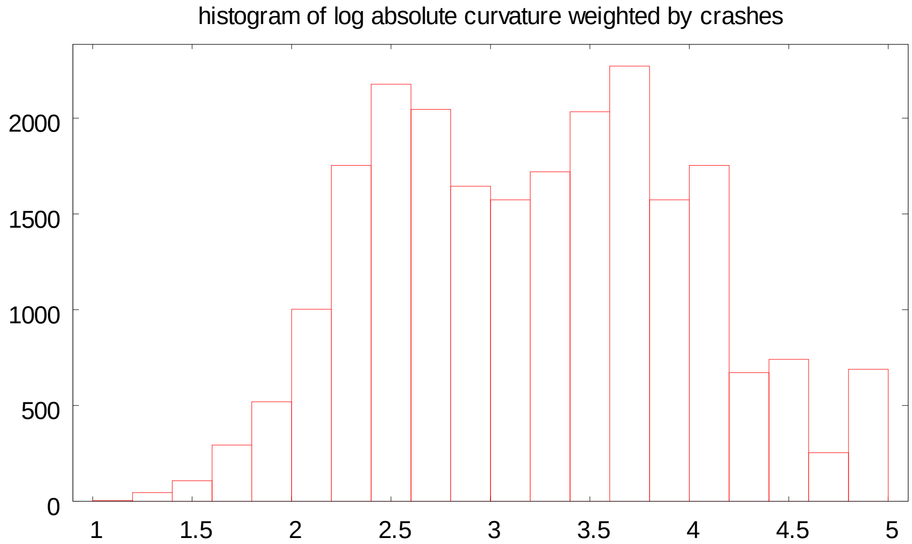
3.6 Curvature



The histogram is of log_{10} *absolute radius of curvature*. Straighter roads are to the right hand side of the graph, sharp bends to the left hand side. By convention, straight roads are assigned a radius of curvature 100,000.

The following two graphs show the histogram of log_{10} *absolute curvature* weighted by *traffic adt est* and by the number of crashes.





Weighting by *traffic adt est* doesn't make a lot of difference to the general shape of the graph. However weighting by the number of crashes shows increased heights in the histogram corresponding to low radii of curvature.

3.7 The Advisory Speed calculation

Advisory speed is not used directly in the main models but is used in calculating the *out of context curve* (OOCC) effect.

This is the formula I used for calculating the *Advisory Speed*.

$$AS = -\left(\frac{107.95}{H}\right) + \sqrt{\left(\frac{107.95}{H}\right)^2 + \left[\frac{127{,}000}{H}\right]\left[0.3 + \frac{X}{100}\right]}$$

where AS = RGDAS Advisory Speed (km/h)

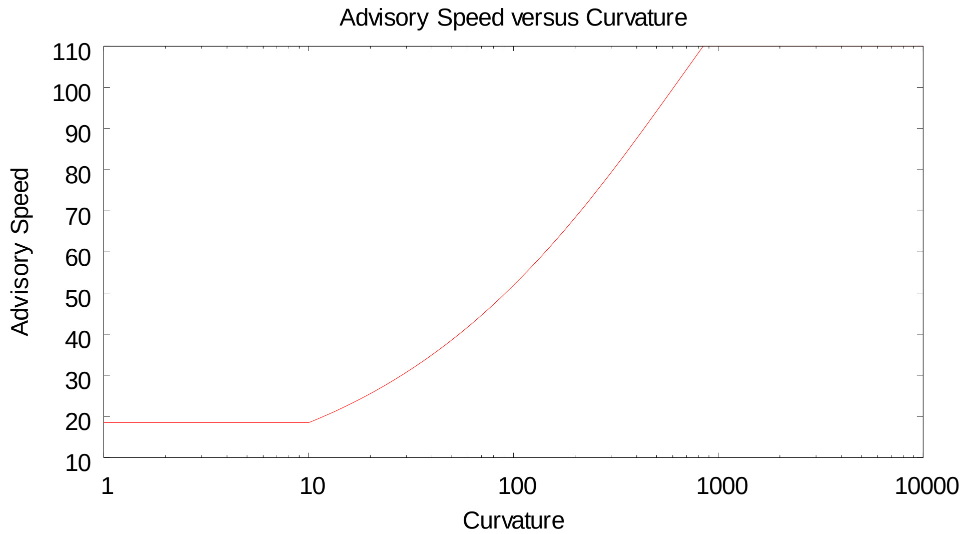
$$X \quad = \quad \text{\% Crossfall (sign relative to curvature)}$$

$$H \quad = \quad \text{Absolute Curvature (radians/km)} = (1000\text{m} / R)$$

X and *R* were taken from the road geometry data collected by the SCRIM machine. If $R < 0$ then the sign of *X* was switched. Then the range of *X* was limited to 0 to 30.

The resulting value of *AS* was capped at 110 km/hr. In urban locations the cap for when calculating *AS2* (but not *AS1*) defined in the next section, 3.8, was set at 70 km/hr.

Here is a graph of Advisory Speed versus curvature with crossfall set to zero.



The curve is horizontal for radii of curvature less then 10 (in absolute value) because I have replaced all radii of curvature less than 10 by 10 in all the analyses.

3.8 Out-of-context-curve indicator (OOCC)

The *out-of-context-curve* indicator is intended to show to what an extent a curve is unexpected. So if a curve is preceded by a section of straight road it will be assigned a high value of the *out-of-context-curve* indicator.

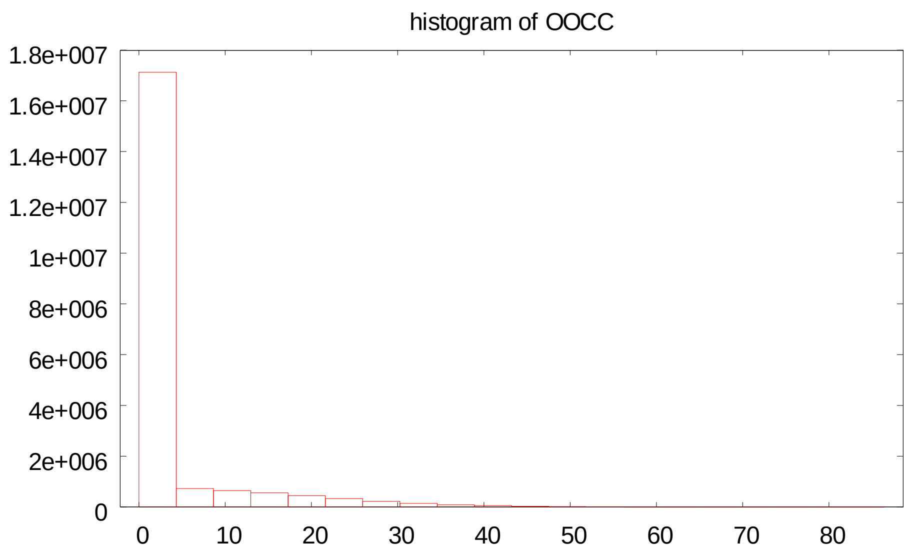
The out-of-context-curve indicator for a particular ten-metre road section is based in the advisory speed defined in section 3.7 above and is defined as follows: Take the average advisory speed, *AS1*, for the current and preceding two ten-metre sections and the average advisory speed, *AS2*, for the 50 ten-metre sections preceding the 3 used to calculate *AS1*. Then the *OOCC* indicator is zero if $AS2 \leq AS1$ and

$AS2 - AS1$

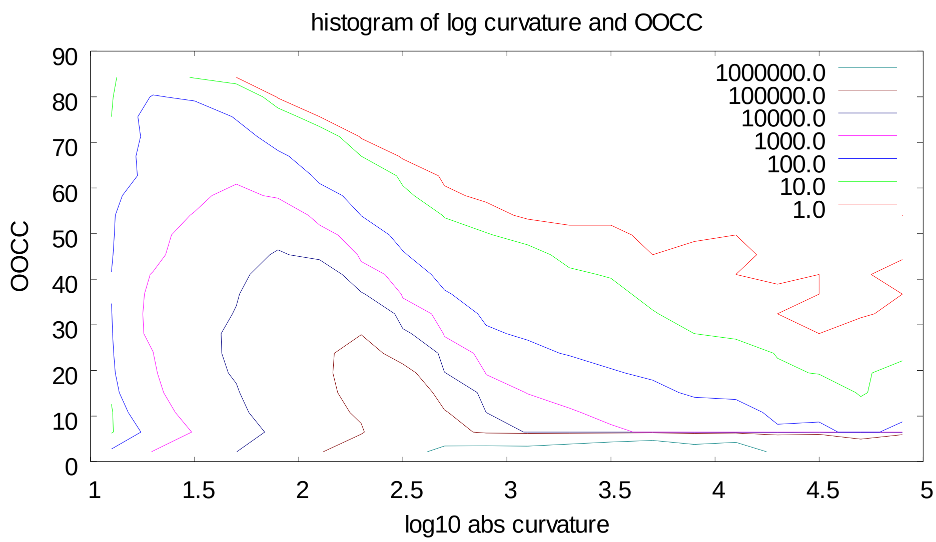
if *AS2* > *AS1*.

For the *decreasing* side of the road I do the same thing in the opposite direction.

This definition is not entirely satisfactory since the value does not die away quickly enough when one drives into a sequence of curves. So when we are fitting both the *curvature* effect and the *OOCC* effect, possibly some of the *curvature* effect is being fitted by the *OOCC* term.

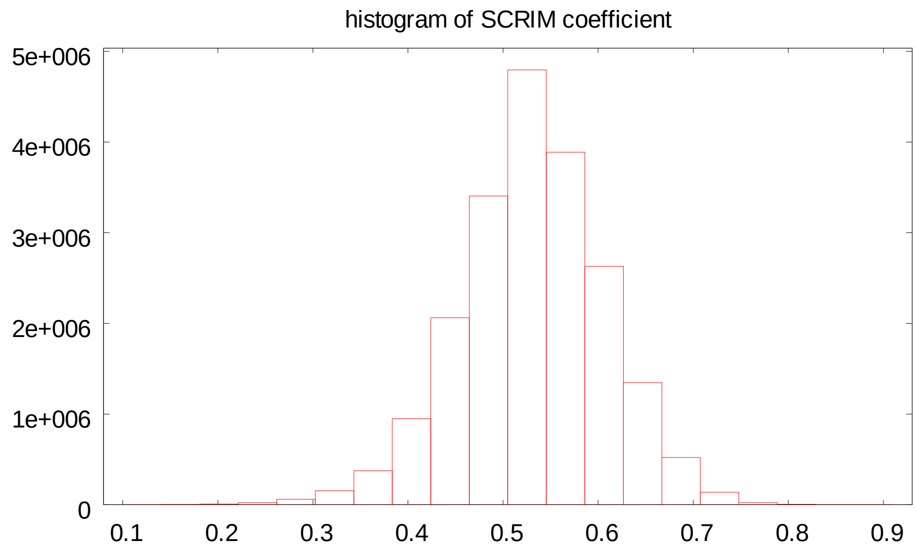


The following contour plot is a two-dimensional histogram of *OOCC* and log_{10}*(absolute curvature)*. The ranges of each variable were divided into 25 bins and a two dimensional histogram calculated. The results are fed to a contour plot program. The contours indicate the number of points per two-dimensional bin. So above the red line there are no points. Between the red and green lines there are 1 to 9 points per bin. Between the green and blue lines there are 10 to 99 points per bin and so on.

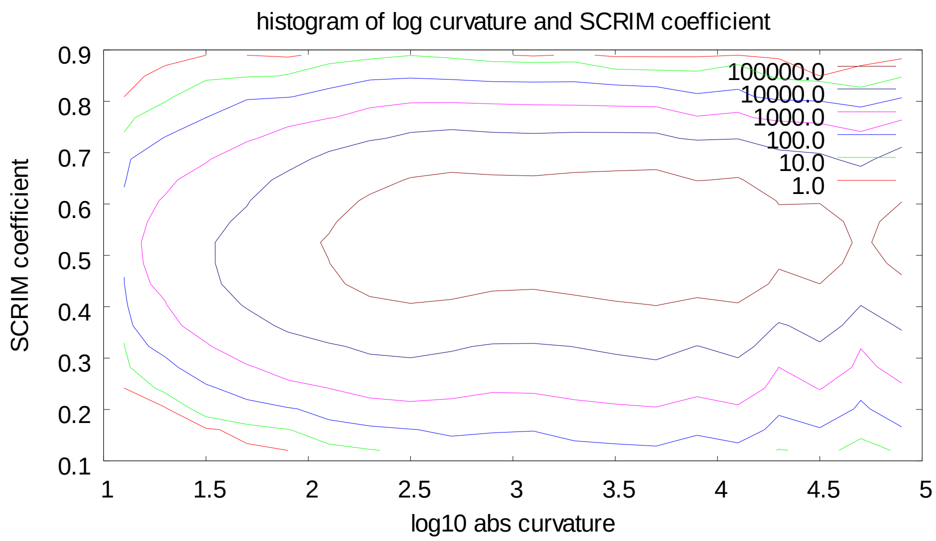


3.9 SCRIM coefficient

The SCRIM coefficient is the measure of skid resistance.



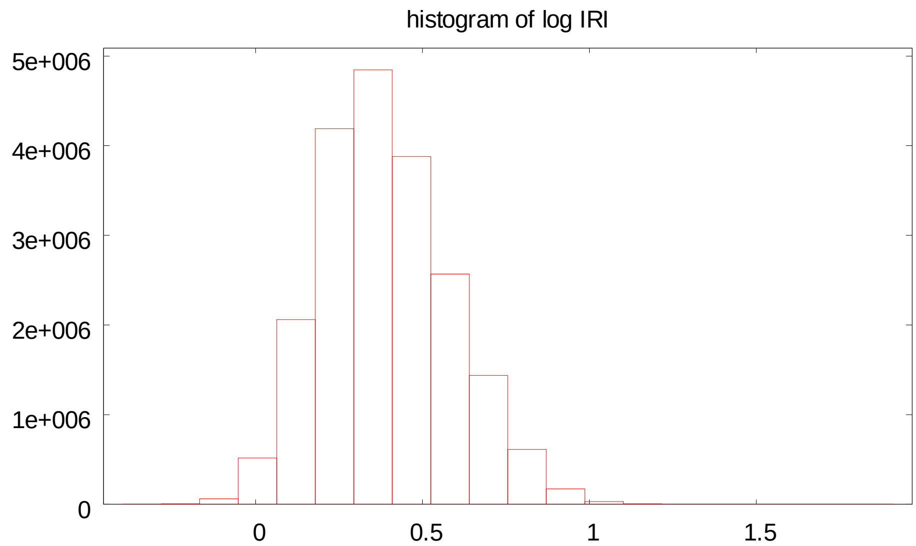
The following is a two-dimensional histogram of *SCRIM* and *log*$_{10}$*(absolute curvature)*. See section 3.8 for a description of two dimensional histograms.



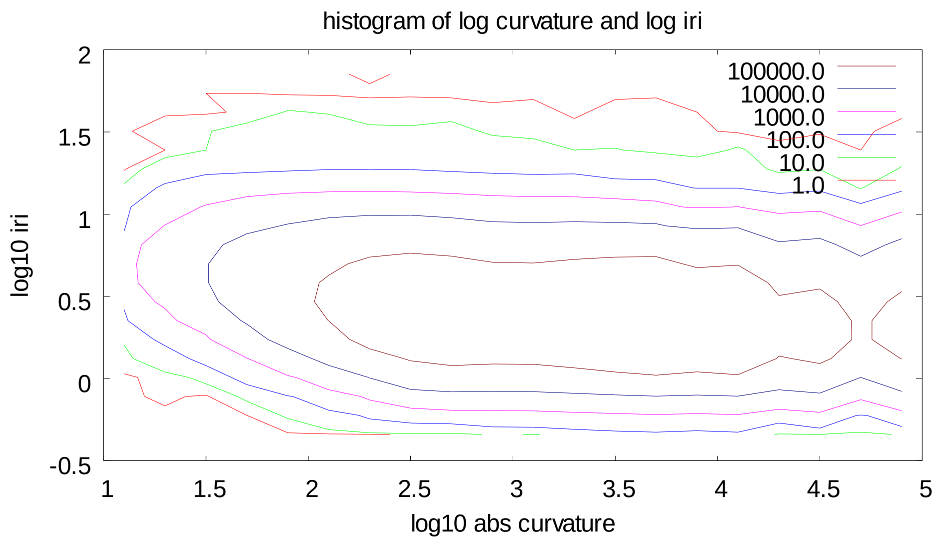
Unlike the corresponding graph for *IRI* in section 3.10 below there is only a slight, if any, relationship between *curvature* and *SCRIM*.

3.10 IRI (roughness)

IRI is the measure of roughness. The histogram is of values of $log_{10}(IRI)$.



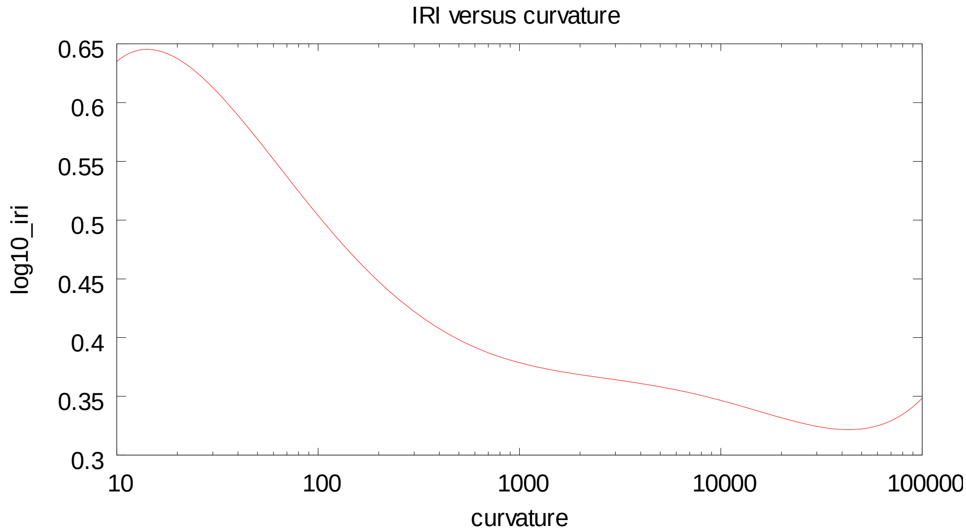
The following is a two-dimensional histogram of $log_{10}(IRI)$ and $log_{10}(absolute\ curvature)$. See section 3.8 for a description of two dimensional histograms.

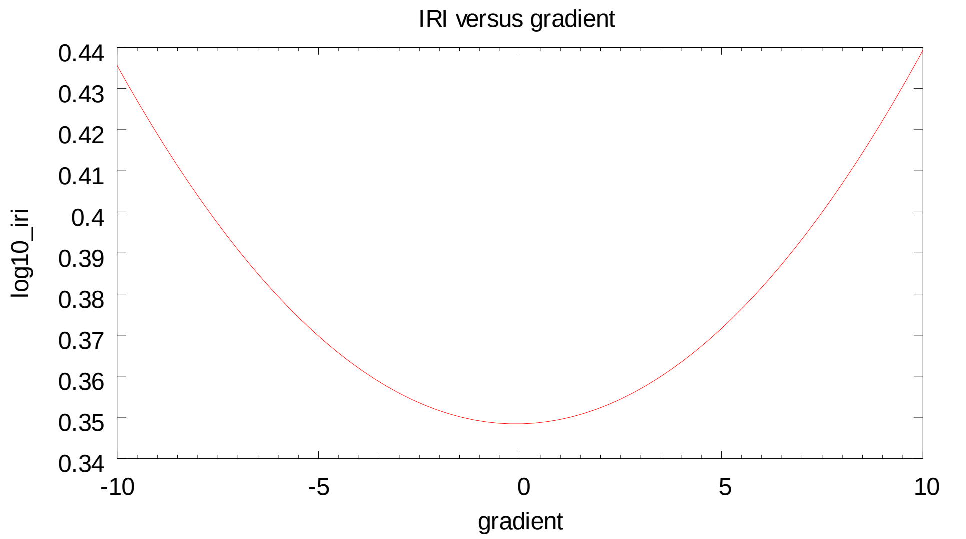


This histogram shows a slight curling up on the left hand side. Highly curved roads tend to have higher roughness. This is, at least partly, a measurement effect. I considered it advisable to try to reduce this effect and this is described in section 3.11 below.

3.11 Adjusted IRI

I carried out a regression analysis, predicting *log_{10}(IRI)* with a fifth degree polynomial of *log_{10}(absolute curvature)* and second degree polynomial of *gradient*. The predicted values are shown in the graphs below.

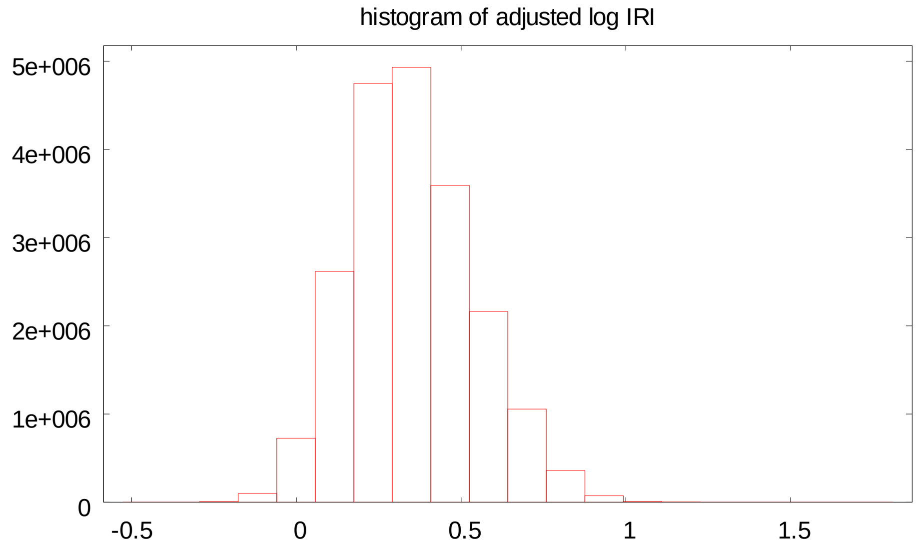


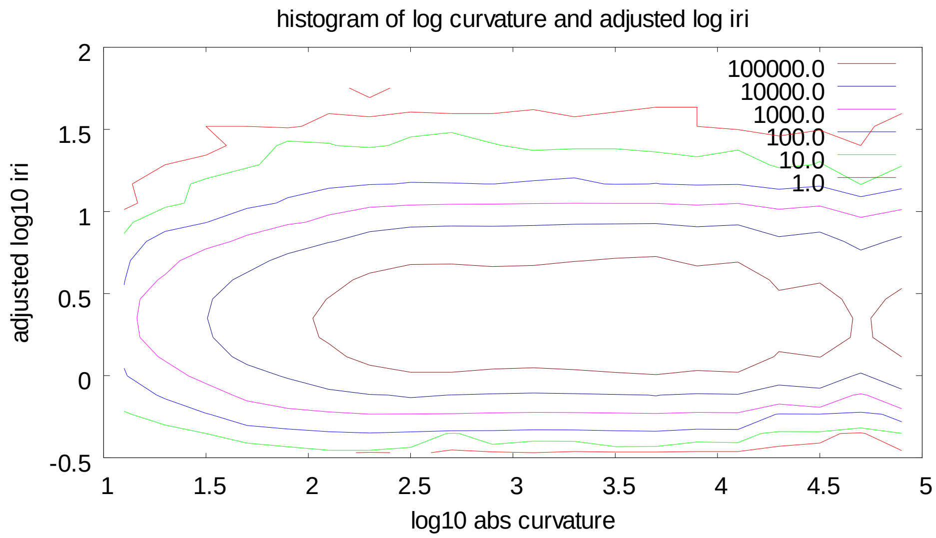


We can use the regression to adjust the *IRI* value to remove the effect of *curvature* and *gradient*. The adjustment has been set so that there is no adjustment for *curvature* = 10,000 and *gradient* = 0. The adjustment reduces the *IRI* by a factor of about 2 when the *curvature* is 10 and the *gradient* 0 or by a factor of about 1.2 when the *curvature* is 10,000 and the *gradient* is 10.

For the regression coefficients and details of the adjustment calculation see the worksheet “Adj IRI” in the spreadsheet “fitted_all.xls”.

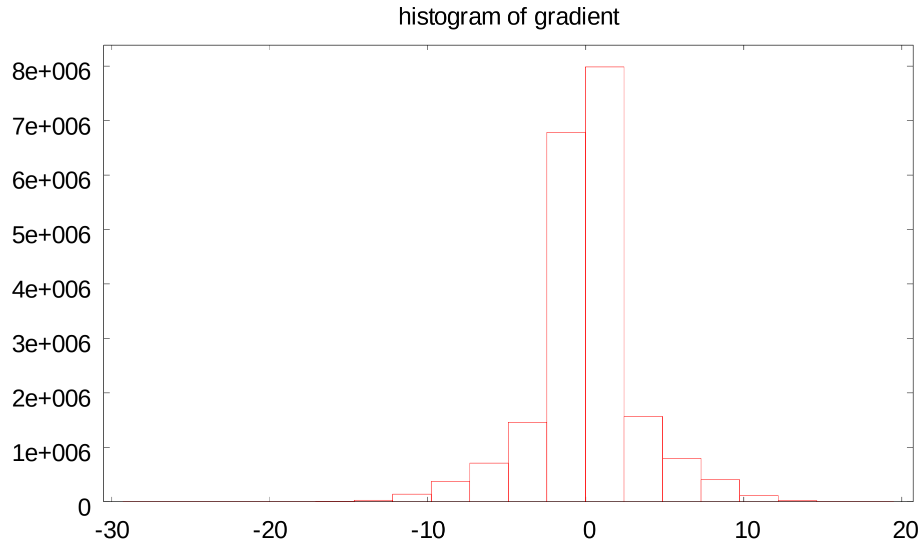
Here is the histogram of the *adjusted log_{10}(IRI)* and the two dimensional histogram of the *adjusted log_{10}(IRI)* and *log_{10}(absolute curvature)*.





The curling up on the left hand side in the corresponding histogram for $log_{10}(IRI)$ has now disappeared.

3.12 Gradient



A positive gradient means uphill, negative means downhill. Most values are between -10 and 10 but there are more extreme values in both directions.

3.13 Skid-site and adjusted skid-site categories

Skid-site values are given for each 10 metre segment. Here are the definitions

skid-site	Description
4	Normal roads (event free)
3	Approaches to road junctions, down gradient 5-10%, motorway junction area
2	Urban curve < 250m radius, rural curve < 400m radius, down gradient > 10%, on ramps with ramp metering
1	Railway level crossings, traffic signals, pedestrian crossings, stop and give way signs, roundabout approaches
5	Divided carriageway (event free)

We are not considering divided roads in this study so skid-site 5 is not relevant. We want curvature and gradient to be handled by the curvature and gradient predictors rather than skid-site and hence we need to define an adjusted skid-site variable.

adjusted skid-site	Description
4	Normal roads (event free)
3	Approaches to road junctions
2	Not used
1	Railway level crossings, traffic signals, pedestrian crossings, stop and give way signs, roundabout approaches
Missing	Divided carriageway

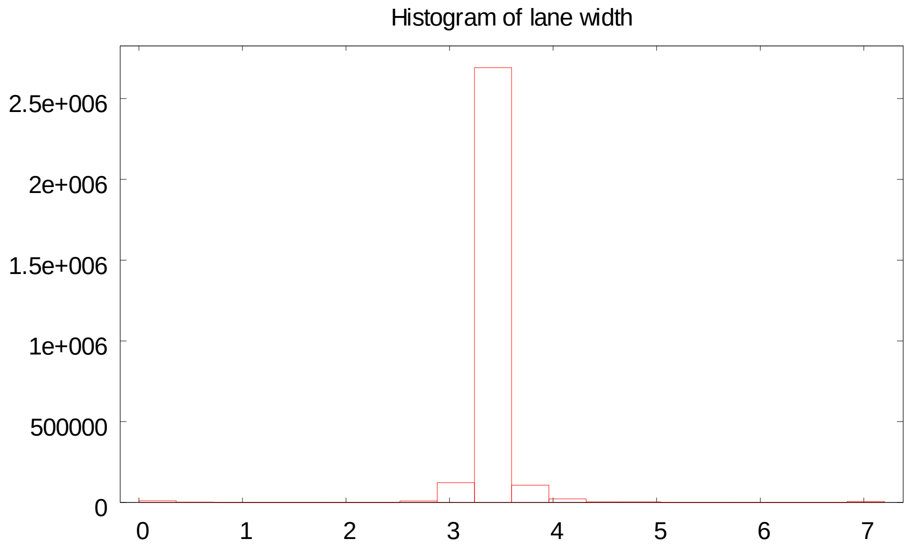
Event-codes are reported for each 10 metre segment and these enable us to calculate the adjusted skid-site value according to the following rules:

- 4 is the default
- *missing* if event code L (divided road) appears anywhere
- 1 for event codes A-F or 1
- 3 for event codes J, M or O

The skid-site effect calculated by the model is going to be very rough since there is likely to be considerable variation in the hazard at different skid-site locations.

3.14 Lane width

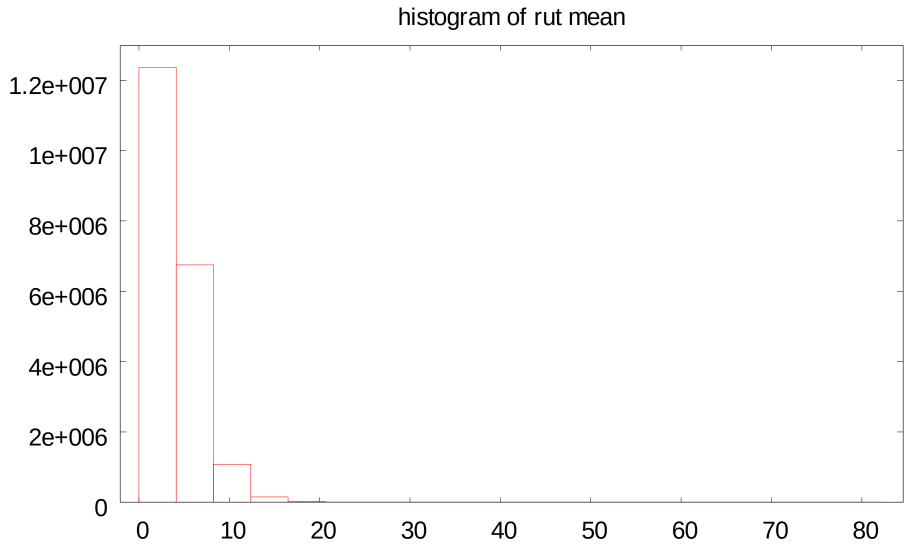
The carriage way table includes a lane width variable. This appears to be incomplete and I have not used it in the analyses.

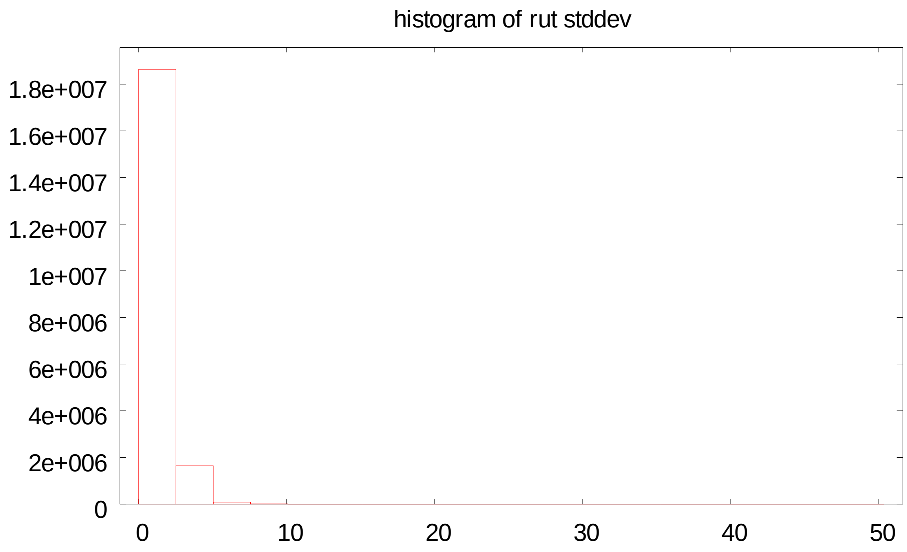


It may be possible to pick up the surface width from the surface data table, but this is not something I have investigated yet.

3.15 Rut mean and rut standard deviation

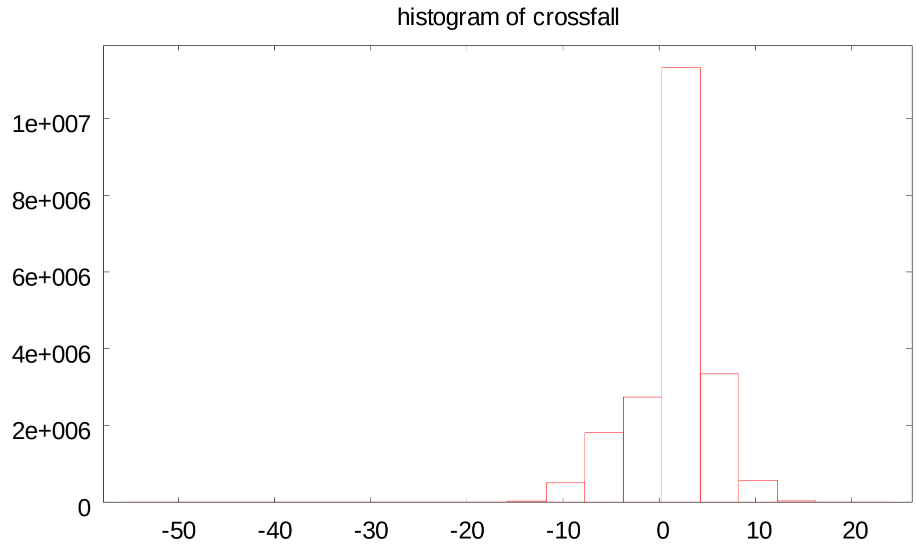
These weren't significant in the earlier analysis and I haven't included them in the main analysis. Rut mean depth is included in a separate analysis.





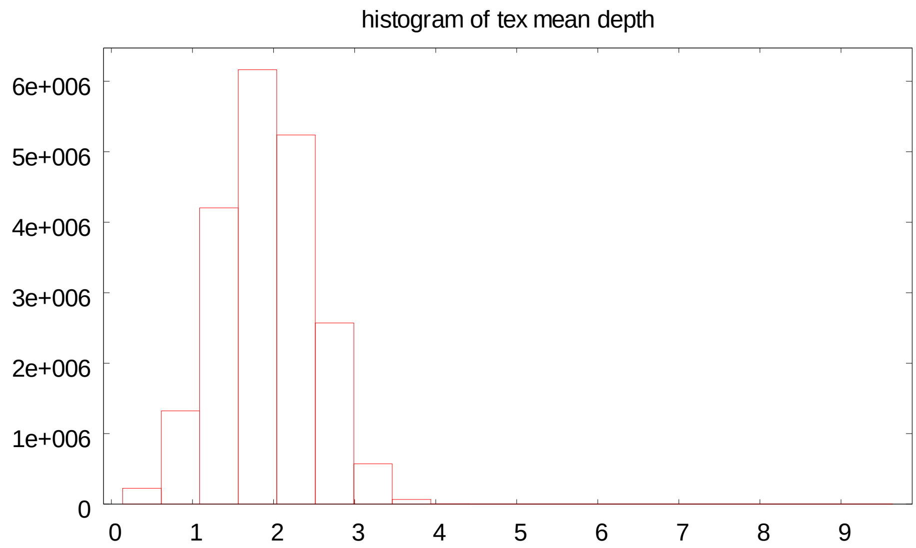
3.16 Crossfall

It was fairly marginal whether crossfall had a significant effect and it is not included in the analyses in this report. However the effect is likely to be very complicated in its interaction with curvature so it probably should not be completely ruled out at this stage.



3.17 Texture depth

This wasn't significant in the earlier analysis and I haven't included it in the main analysis. It is investigated briefly in a supplementary analysis. See section 7.2.



There is a problem with including texture depth in the analysis. Asphaltic concrete surfaces, as opposed to chip-seal, tend to be used in more populated areas. Roads with asphaltic concrete surfaces tend to have a lower texture, but being in more populated areas, may have a higher crash risk. Hence, there is the possibility of a spurious correlation. By fitting both the urban-rural effect and the skid site categories in the model one can reduce this spurious correlation, but one really needs to include the surface type as a predictor variable. The information is available but it is outside the scope of the present study to include it.

4 The Poisson regression model

4.1 The model

I suppose each side of each 10-metre length of road can generate crashes at the rate (per year)

$$a \exp(L) \tag{1}$$

where a is the ADT (per side) and L is a linear combination of the road characteristics being transformations of terms including

- a constant,
- gradient,
- curvature,
- out of context curve effect,
- skid-site classification,
- skid resistance,
- log(ADT),
- year,
- region,
- urban/rural classification.

The coefficients in the linear combination are the unknown parameters to be estimated.

The spreadsheets distributed with this report show how to calculate of the crash risk from the fitted model. See section 5.4 for more details.

Since we are taking the exponential of L, a linear combination of the road characteristics, the actual model is multiplicative.

Note that the average daily traffic (ADT) appears in the model in two places, a in equation (1) and as a component of L. These could have been combined into a single term in L. However, by using the formulation in (1) the component in L is present only if the crash risk (expected number of crashes per 100 million vehicle kilometres) depends on ADT. When there is dependence, this dependence is modelled by the size of the coefficient of log(ADT) in L. The crash risk is given by

$$\frac{10^{10}}{365} \exp(L)\,. \tag{2}$$

The *actual* rate that crashes are reported in a 10 metre length of road is the average of *generating* rates over the 10 metre lengths in its immediate neighbourhood (on the same road) and summed over the two sides of the road. In the results reported here the average is over 10 metre lengths within 100 metres of the length being considered. Typically, this gives an average over 210 metres on each side of the road. There is no weighting down of the more distant 10 metre lengths. This averaging allows for error in reporting the location and the possibility that a crash ends at a location some distance from the piece of road involved in generating the crash.

Because we are combining the sides of the road we don't have to know the directions of vehicles involved in the crash.

The model assumes that the crashes are statistically independent and the number in each 10-metre segment follows a Poisson distribution. (Of course, for most segments the number will be zero).

Fitting is by maximum likelihood and uses my C++ libraries for matrix manipulation and automatic differentiation from http://www.robertnz.net and my prototype array and statistical modelling package currently under development. I am writing a description of the modelling package at http://www.robertnz.net/cpp_stats.htm.

4.2 Error estimates

The error estimates and significance tests produced by the model are based on the assumption that the values of the *dependent variable*, the number of crashes in each ten metre segment each year, are distributed as independent Poisson variables. It was clear in the 2004 study, and the same is the case here, that this assumption is not correct.

The usual trick in this situation is to use the *residual deviance* calculated as part of the maximum likelihood estimation to provide a correction to the error estimates. This does not work in the present analysis because the average number of crashes per segment is very small – much less than 1.

The analysis in section 5.5 suggests error estimates should be increased by around a factor of around 2.3. This corresponds to increasing the critical points for the tests in the analysis of variance tables by a factor of around 5.4.

The increased error is probably due, an least partially, to an unknown factor that affects a length of road in the same way. One possible suspect is that the ADT estimate is subject to sufficient error to cause a problem.

If this is the case, the errors of corresponding to variables that vary slowly along a road such as SCRIM or IRI are likely to be subject to the increased error, but variables that change rapidly such as curvature, and their interactions with other variables will be less affected. A more advanced analysis could attempt to allow for this extra variability. For example, it could include a random effect or apply a statistical technique known as *The Jack-Knife*.

5 The model fit

5.1 The regression analysis

I have carried out the regression analysis using the four categories of crash data described in section 3.4. The analysis is essentially the same as the simplified one given in the 2004 report with the polynomial transforms of the variables. I am using the adjusted skid site categories as described in section 3.13 and the *out-of-context-curve* variable described in section 3.8. I am also including and interaction term between curvature and the adjusted IRI value. Road region with 14 levels replaces the region variable with 7 levels used in the 2004 report.

The following table shows details of the predictor variables

Predictor variable	**Bounds**	**Notes**
year		discrete variable, 10 levels
region		discrete variable, 14 levels
urban_rural		discrete variable, 2 levels
adj_skid_site		discrete variable, 3 levels
poly3_bound_OOCC	0, 35	3rd degree polynomial of bounded version of OOCC
poly2_bound_log10_abs_curvature	2,4	2nd degree polynomial of bounded version of log of absolute curvature
poly2_log10_ADT		2nd degree polynomial of ADT
poly2_scrim-0.5000		2nd degree polynomial of (scrim – 0.5)
poly3_bound_abs_gradient	4,10	3rd degree polynomial of bounded version of absolute curvature
poly3_bound_adj_log10_iri	-0.3, 1.2	3rd degree polynomial of bounded version of adjusted log IRI
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	as above	interaction between 2nd degree polynomial of bounded version of absolute curvature and 2nd degree polynomial of bounded version of adjusted log IRI

By and large I am using the same transformations as in simplified model in the 2004 report. The OOCC variable is new and the bounds on the adjusted log IRI variable have been changed. The inclusion of the interaction term between log absolute curvature and adjusted log IRI is new. This allows for a different sized effect for different curvatures.

I present the analysis of variance tables and predicted value graphs in this report. The estimates of effects are in the spreadsheets distributed with this report.

5.2 Analysis of variance tables

The analysis of variance tables show two versions of the chi-squared values.

The type III value is the version that seems often favoured by SAS users – each variable is tested in the presence of all other variables. This can be misleading if two variables are highly correlated since both can appear non-significant when tested in the presence of the other. This version does not make sense when you test a main effect when that effect is also part of an interaction term (curvature and IRI in our analyses). The type I version is when each variable is tested only in the presence of the variables above it in the table. This version tends to be favoured by S-plus and R users. One tries to arrange the order of variables so that the most important variables come first with interactions coming after main effects, but even then, apparent significance can be misleading when variables are highly correlated (as is the case with OOCC and curvature in our analyses). So I like to give both tables.

In the 2004 study the standard errors given by the Poisson model seemed to be underestimated by a factor of around 2. The same seems to be true here. This means that the 1% points given in the tables should be increased by a factor of 4, especially those that vary only slowly as we proceed along a road. See sections 4.2 and 5.5 for more details. In fact, 5.4 seems to be a better figure. However for curvature and interactions with curvature the correction is probably rather less.

5.2.1 All casualty crashes

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	519	526.35
region	13	27.7	298.88	665.07
urban_rural	1	6.63	27.811	485.58
adj_skid_site	2	9.21	4693.3	6288.5
poly3_bound_OOCC	3	11.3	454.48	5399.5
poly2_bound_log10_abs_curvature	2	9.21	110.88	459.57
poly2_log10_ADT	2	9.21	576.57	518.17
poly2_scrim-0.5000	2	9.21	221.65	265.24
poly3_bound_abs_gradient	3	11.3	46.334	65.975
poly3_bound_adj_log10_iri	3	11.3	83.129	108.94
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	107.14	107.14

With the exception of *urban_rural* (in the type III table) all variables are statistically significant although *gradient* is fairly marginal.

5.2.2 Wet casualty crashes

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	194.1	134.7
region	13	27.7	221.21	485.06

Predictor variable	df	1% pt.	Chi-squared	
urban_rural	1	6.63	38.927	8.1259
adj_skid_site	2	9.21	684.88	1005.1
poly3_bound_OOCC	3	11.3	246.9	4211.5
poly2_bound_log10_abs_curvature	2	9.21	77.204	428
poly2_log10_ADT	2	9.21	141.1	104.84
poly2_scrim-0.5000	2	9.21	389.38	436.88
poly3_bound_abs_gradient	3	11.3	68.251	83.437
poly3_bound_adj_log10_iri	3	11.3	32.167	39.785
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	42.743	42.743

The chi-squared value of the *scrim* variable has increased showing the importance of skid resistance for wet roads, otherwise the chi-squared values have mostly decreased as you would expect with the smaller number of crashes. The *iri* and *curvature* × *iri* terms are not significant when we use the 1% point multiplied by 5.4.

5.2.3 Selected casualty crashes

Predictor variable	df	1% pt.	Chi-squared	
			Type III	Type I
year	9	21.7	430.4	417.6
region	13	27.7	207.98	607.61
urban_rural	1	6.63	98.652	265.56
adj_skid_site	2	9.21	465.85	971.36
poly3_bound_OOCC	3	11.3	382.23	7057.3
poly2_bound_log10_abs_curvature	2	9.21	146.39	711.96
poly2_log10_ADT	2	9.21	710.14	600.84
poly2_scrim-0.5000	2	9.21	285.08	304.36
poly3_bound_abs_gradient	3	11.3	41.345	53.905
poly3_bound_adj_log10_iri	3	11.3	34.04	57.841
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	59.927	59.927

The chi-squared values are similar to those for all casualty crashes. The *iri* and *curvature* × *iri* terms are marginal when we use the increased value of the 1% point. The *urban_rural* variable is now significant in the type III version of the chi-squared value. I interpret this to mean that when we remove the types of crashes that tend to be associated with urban areas the increased safety due to the lower speed limit becomes apparent.

5.2.4 Wet selected casualty crashes

Predictor variable	df	1% pt.	Chi-squared	
			Type III	Type I
year	9	21.7	166.94	114.38

Predictor variable	df	1% pt.	Chi-squared	
region	13	27.7	170.96	464.49
urban_rural	1	6.63	75.633	251.1
adj_skid_site	2	9.21	80.89	196.44
poly3_bound_OOCC	3	11.3	194.9	4606.2
poly2_bound_log10_abs_curvature	2	9.21	87.782	570.04
poly2_log10_ADT	2	9.21	181.86	138.04
poly2_scrim-0.5000	2	9.21	429.95	462.8
poly3_bound_abs_gradient	3	11.3	58.002	69.765
poly3_bound_adj_log10_iri	3	11.3	17.233	23.964
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	32.325	32.325

The chi-squared values are similar to those for the wet casualty crashes. The *scrim* value has increased slightly, the others have generally decreased slightly, the *adjusted skid-site* affect seems a lot smaller.

5.3 Predicted crash rate graphs

In order to see how each variable in the model affects crash rate I have graphed the crash rate predicted by the model as each variable, in turn is varied. For the terms not being varied I have used the following values:

year	2008
region	R03
urban_rural	R
adj_skid_site	4
OOCC	0
curvature	5000
ADT	1000
gradient	0
scrim	0.5
adj_log10_iri	0.3

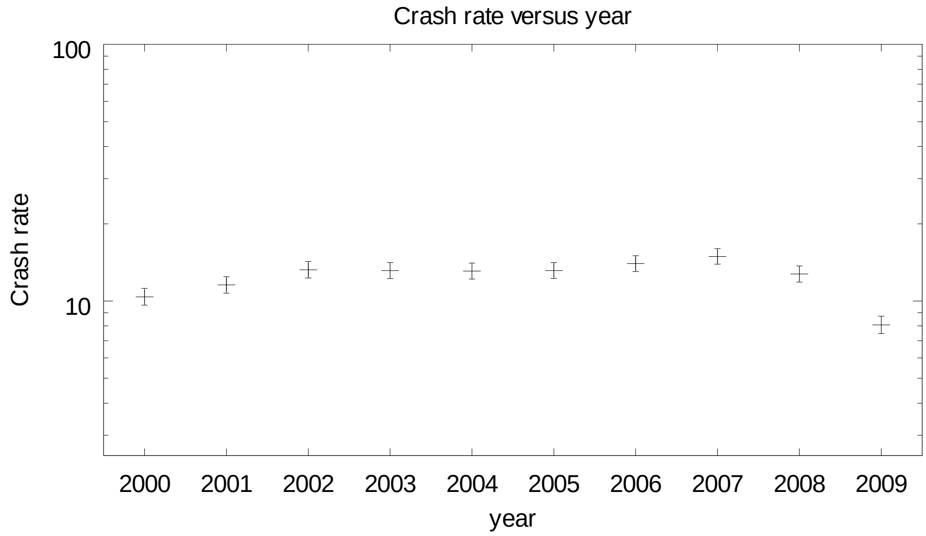
Crash rates are in crashes per 100 million vehicle kilometres. The error bounds show 2 standard deviations (roughly 95% confidence) and are based on the Poisson model – so lengths should be doubled. However they are for the overall crash rate and there is some error that is common to all the points on a graph. So when we look at differences the error may be less that is suggested by the graph (after we have doubled the length).

I show all the graphs for the *all casualty crashes* but only a selection for the others. (*All casualty* means all reported crashes which involve at least one fatal, serious or minor injury).

For the *wet casualty crashes* and *wet selected casualty crashes* we don't know the fraction of time the road is wet and are normalising by the total traffic, not the traffic when the road is wet. Hence the calculated crash rates are much less than for *all casualty crashes* and *selected casualty crashes*.

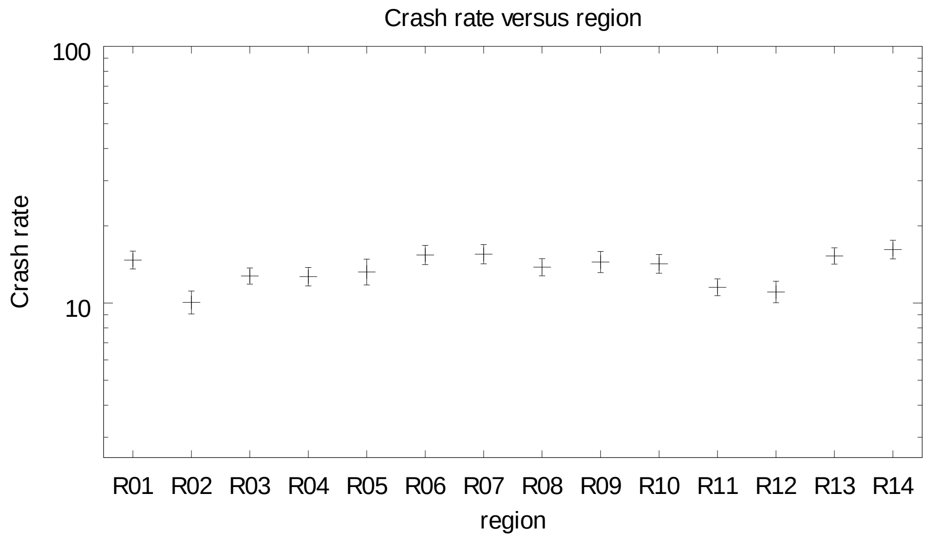
In all the crash rate graphs the vertical scale is logarithmic and there is a ratio of 40 between the crash rate at the lower end and the upper end.

5.3.1 All casualty crashes - year



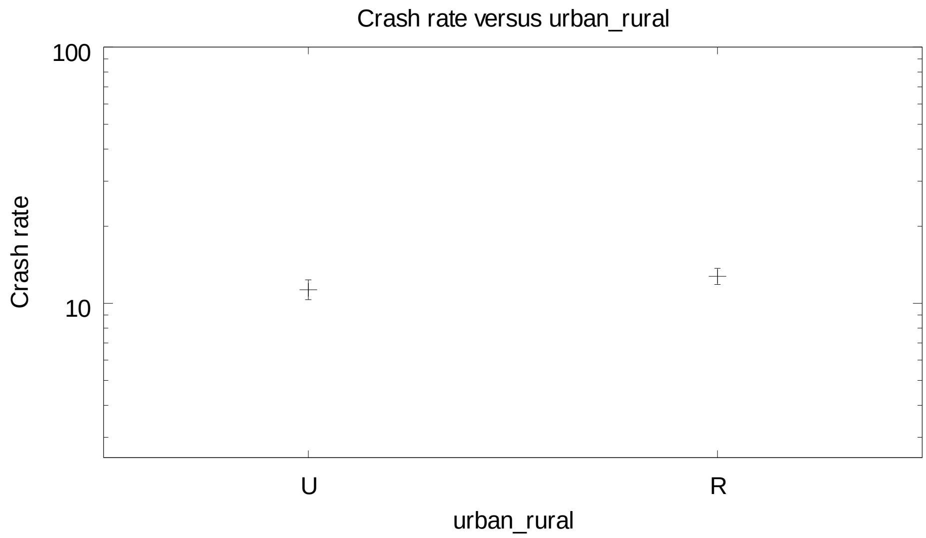
The downturn in 2008 is possibly due to the recession. The drop in 2009 is because not all the data is in yet and does not represent a real effect.

5.3.2 All casualty crashes - region



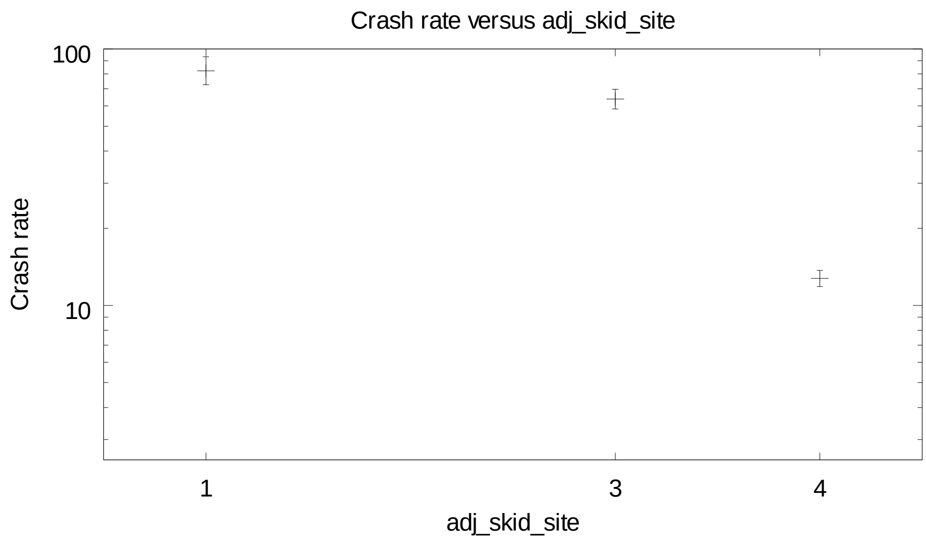
This shows some regional variation. This could be a real effect or it might be due to different reporting rates.

5.3.3 *All casualty crashes – urban/rural*



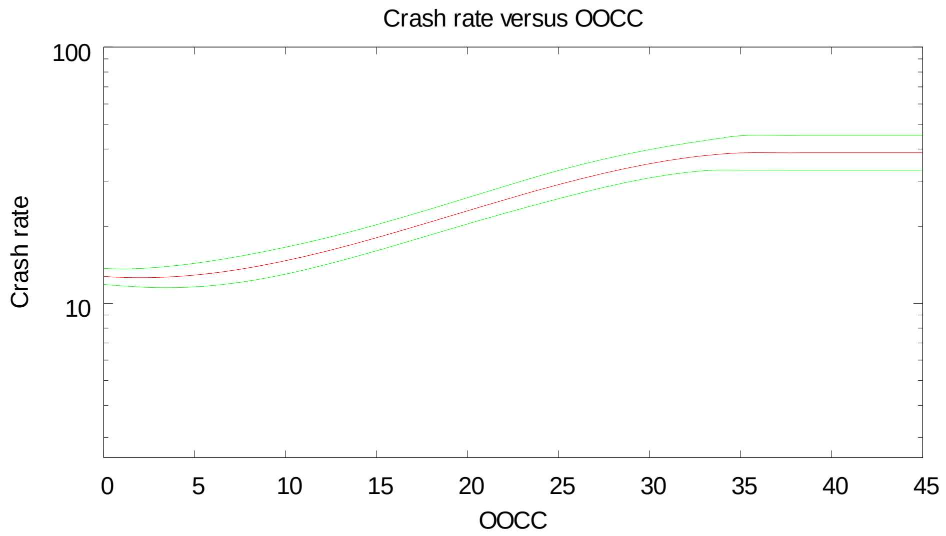
This shows very little difference between urban and rural. Possibly the decreased risk due to a lower speed limit in urban areas is balanced by additional causes of crashes.

5.3.4 *All casualty crashes – adjusted skid-site*



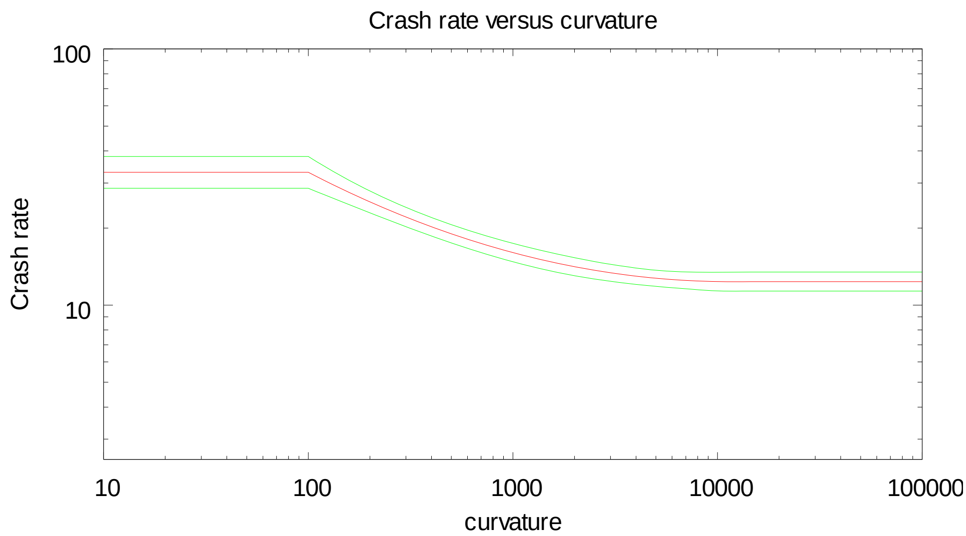
The crash rate for adjusted skid-site 4 is substantially lower than for adjusted skid-site 1 or 3. There needs to be some care in interpreting the actual rates for skid-sites 1 and 3 since these are essentially point events and we are looking at crash rates per kilometre.

5.3.5 All casualty crashes – OOCC



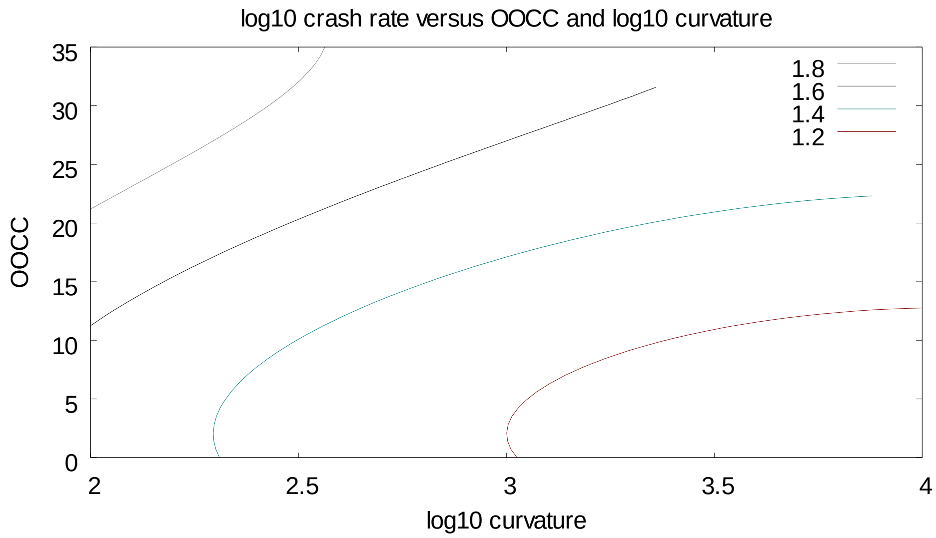
This shows the increasing crash rate as the OOCC (out of context curve) effect increases. Since we have lumped together values where OOCC is greater than or equal to 35 we have a horizontal line for these values.

5.3.6 All casualty crashes – curvature



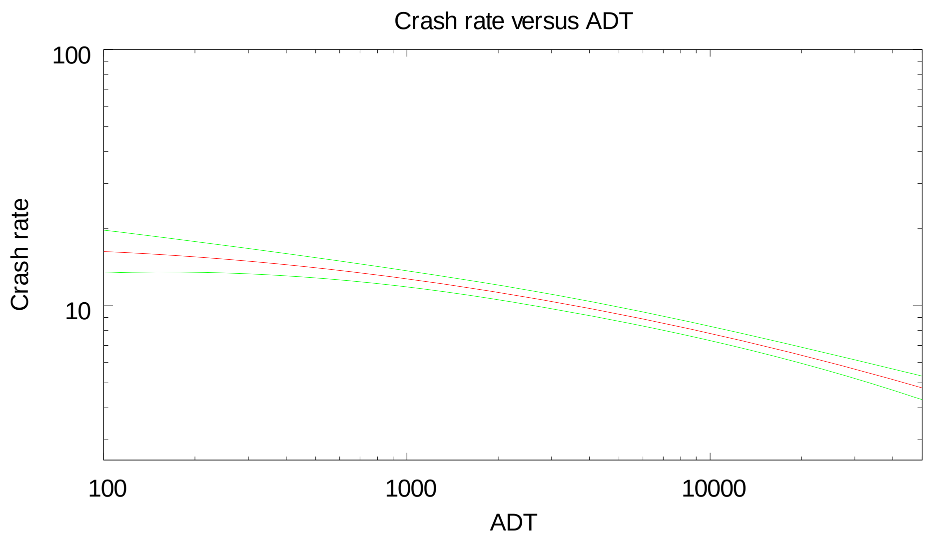
This shows the decreasing crash rate as the absolute radius of curvature increases (road becomes straighter). Since we have lumped together values where absolute curvature is less than or equal to 100 or greater than or equal to 10000 we have horizontal lines for these values.

5.3.7 All casualty crashes – OOCC and curvature



Because OOCC and curvature are closely related, I have drawn a contour plot to show what happens when both are varied. I have omitted the values where a combination doesn't make sense – you can't have high radius of curvature and high OOCC. The contours show $\log_{10}$ crash rate so 1.2 corresponds to a crash rate of 15.8 and 1.8 to a crash rate of 63.

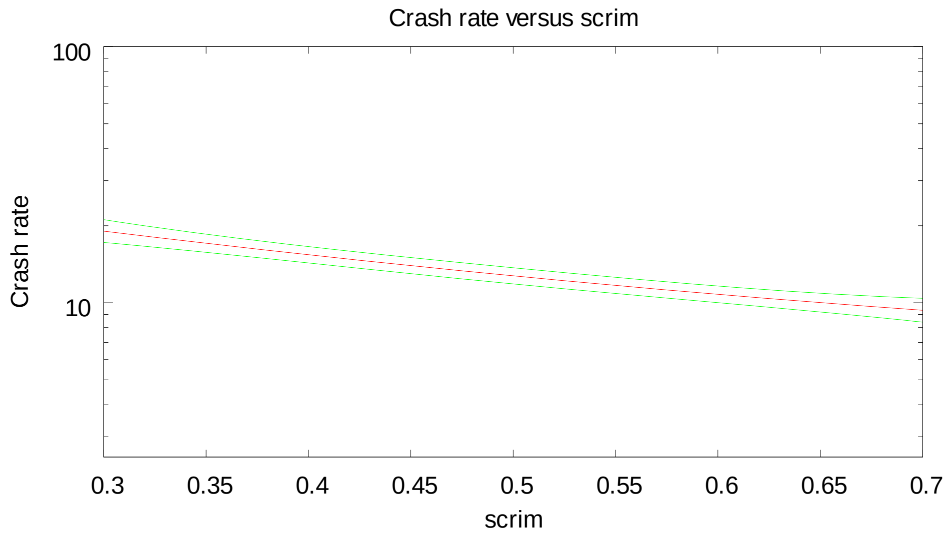
5.3.8 All casualty crashes – ADT



This shows the crash rate dropping as ADT increases. Even after allowing for sharper bends, lower SCRIM, increased roughness that you might expect on lesser used roads, these roads are still more dangerous than the high ADT roads.

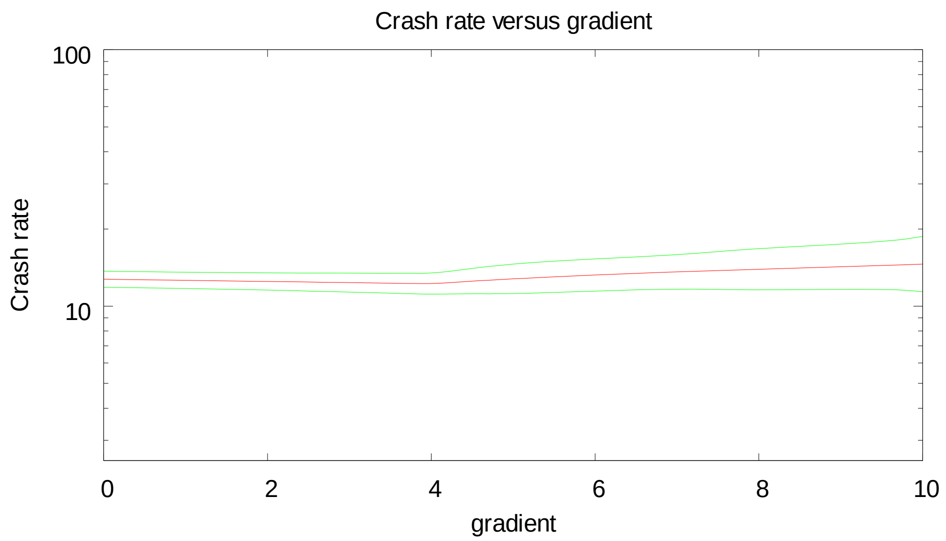
One could hypothesise that the effect is due to more risky driving, less policing, less signage and numerous other factors that are not included in the model and that are associated with lower ADT roads.

5.3.9 All casualty crashes – scrim



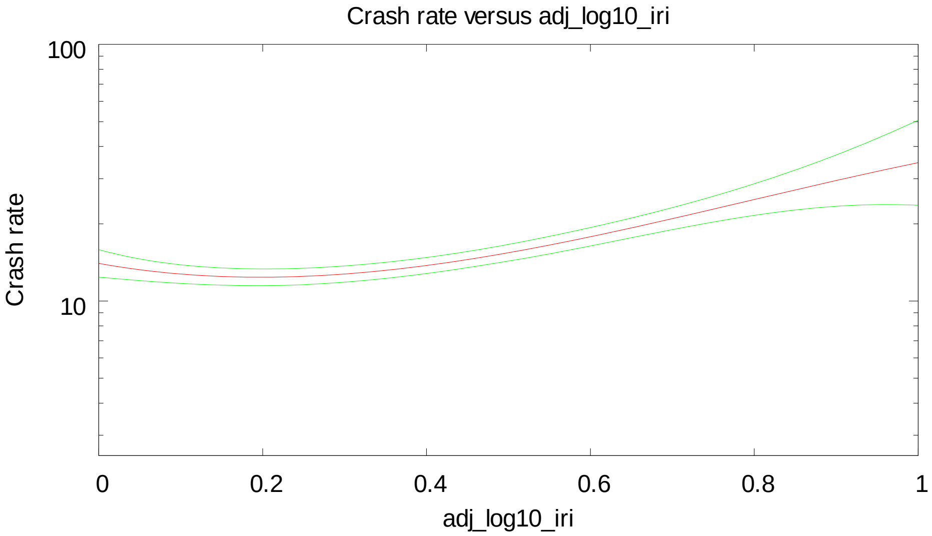
This shows the crash rate decreasing as scrim increases. I am fitting a quadratic curve although not much curvature is apparent. There isn't much data towards the extreme values of scrim values so this graph shouldn't be taken too seriously at the ends.

5.3.10 All casualty crashes – gradient



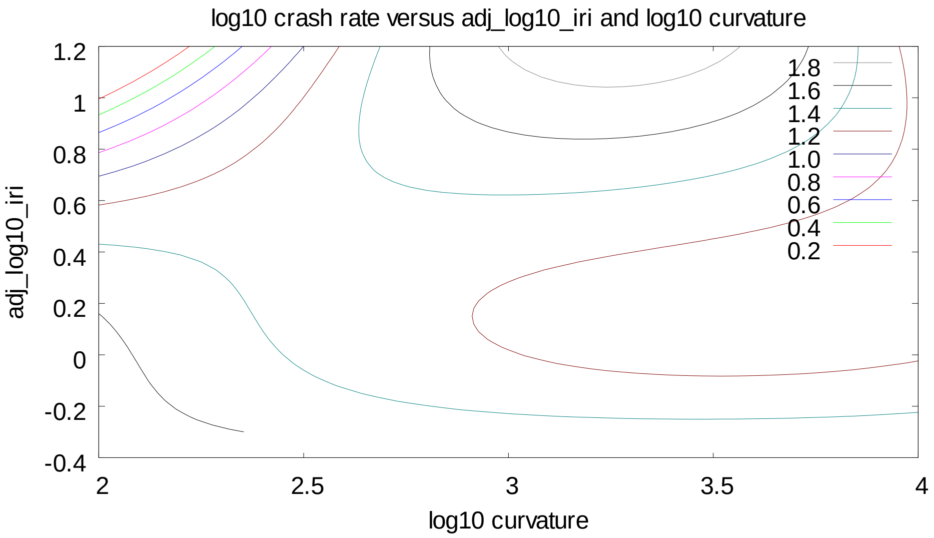
This shows the crash rate increasing slightly as absolute gradient increases. Because we don't know the direction of travel of vehicles involved in crashes we can't distinguish between uphill and downhill gradient. So, assuming uphill decreases risk and downhill increases it, the effects may cancel and we might not expect a big effect. Values less than or equal to 4 are lumped together so we get a horizontal line for these points.

5.3.11 All casualty crashes – adjusted iri

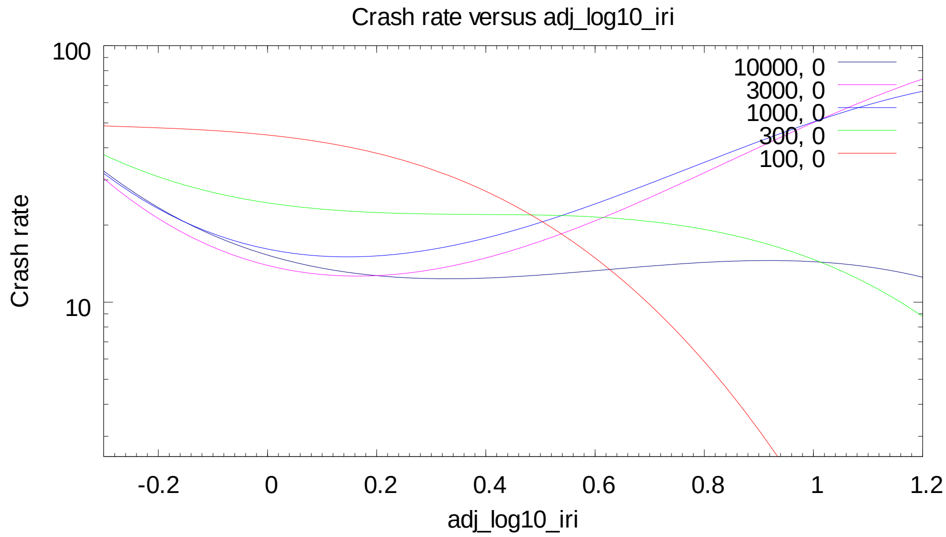


This shows crash risk increasing as the *adjusted iri* increases.

Because we are fitting and interaction between curvature and *adjusted iri* one needs to look at both effects together. The following graph shows a contour plot as both of these are varied.



It is easier to understand the relationship if instead we graph the crash rate against the *adjusted iri* for a selection of values of curvature.



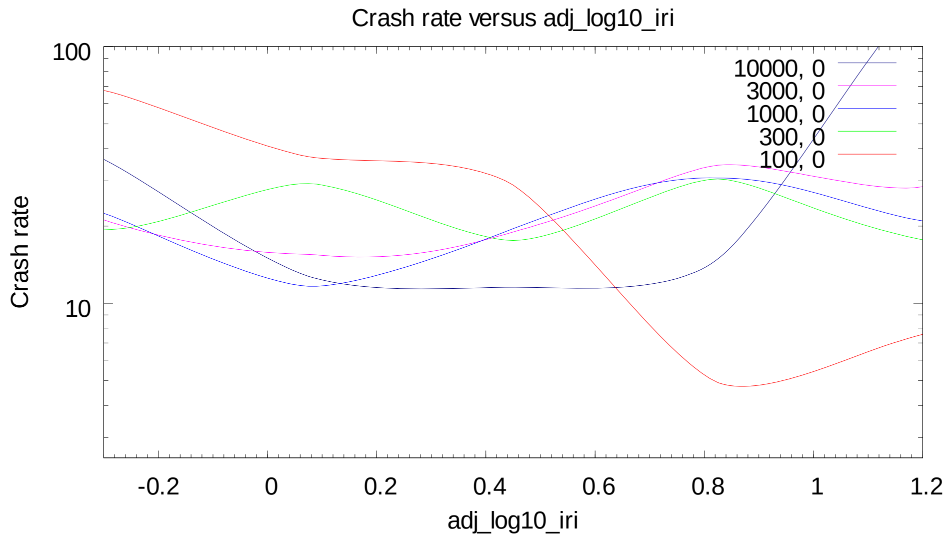
The graph is for radii of curvature 10,000, 3,000, 1,000, 300, and 100 all with *OOCC* = 0. This graph takes the *adjusted* log_{10} *iri* from -0.3 rather than 0 as in the first graph in this section. There is very little data with *adjusted* log_{10} *iri* less than 0 so the increasing values here are possibly a quirk of the polynomial fit.

This graph suggests that

- for straight or near straight roads roughness is not very relevant.
- for modestly curved roads roughness increases crash risk – the contour plot suggests a range of radii of curvature of 500 – 5000 metres for roughness causing increased crash risk.
- for very curved roads roughness is associated with decreased crash risk – but one should be very wary about interpreting this result.
- possibly, very low roughness may be associated with an increase in crash risk (possibly low roughness encourages drivers to go faster or possibly low roughness is associated with other hazards – for example, bridge decks).

5.3.12 All casualty crashes – adjusted iri with tps interaction

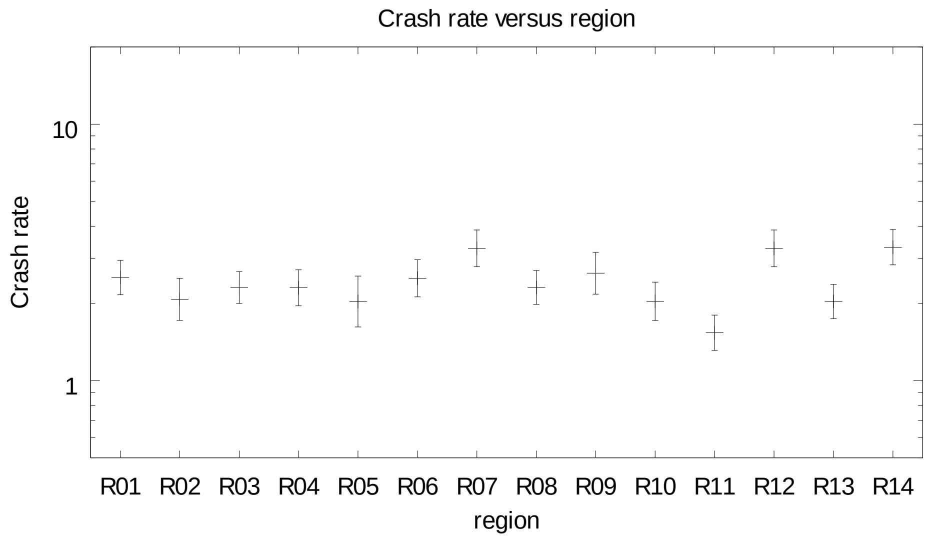
I want to see whether the modelling of the effect of curvature and *adjusted iri* was being constrained by the use of polynomial functions. So I have replaced the *curvature* and *adjusted iri* terms by a thin plate spline (tps) function of these two terms. A thin plate spline is like a higher dimensional analogy to the one dimensional spline used for fitting curves. I have used a thin plate spline based on a 5 × 5 array of knots covering the range of log_{10} *curvature* from 2 to 4 and *adjusted* log_{10} *iri* from -0.3 to 1.2. This gives 24 degrees of freedom for the combined log_{10} *curvature* and *adjusted* log_{10} *iri* term as opposed to the 9 assigned to log_{10} *curvature* and *adjusted* log_{10} *iri* and their interaction in section 5.3.11. So there is a lot more flexibility at the expense of a lot more random error. Here is the *crash risk* versus *adjusted iri* plot.



Most of the *adjusted* log_{10} *iri* data is between 0 and 0.8 (adjusted iri between 1 and 6.3). In this range the results are similar to those using the polynomial functions.

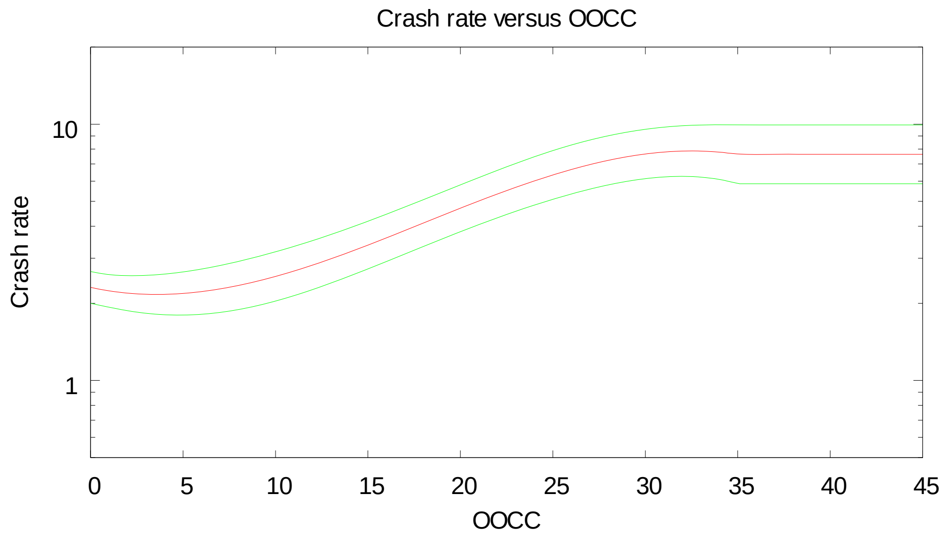
There is a suggestion that above this range there is increased risk in near straight roads. Other analyses do not support this and it seems likely that this is *not* a real effect.

5.3.13 Wet casualty crashes – region



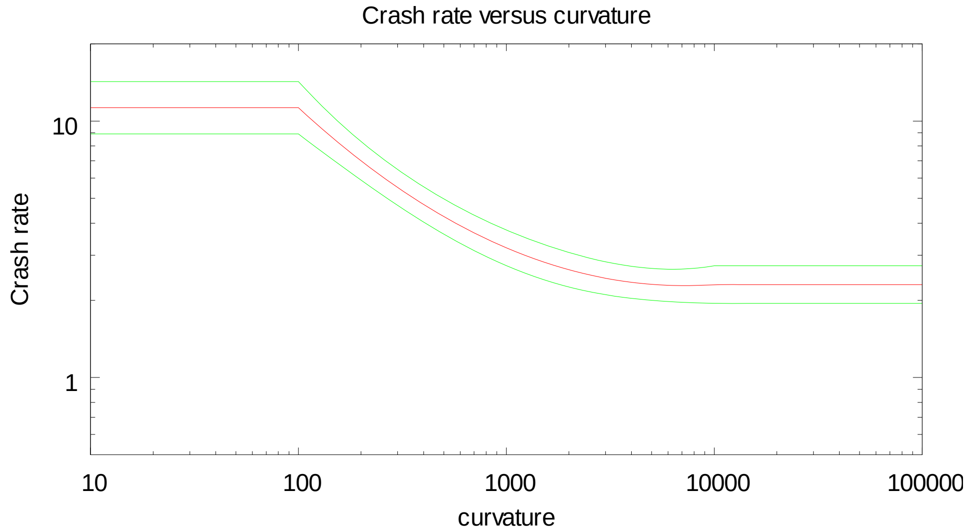
This is similar to the corresponding plot for all crashes in section 5.3.2. The confidence intervals are longer reflecting the smaller number of crashes. The value for the South Island West Coast in R12 is somewhat higher, presumably because of the higher rainfall in the West Coast.

5.3.14 Wet casualty crashes – OOCC



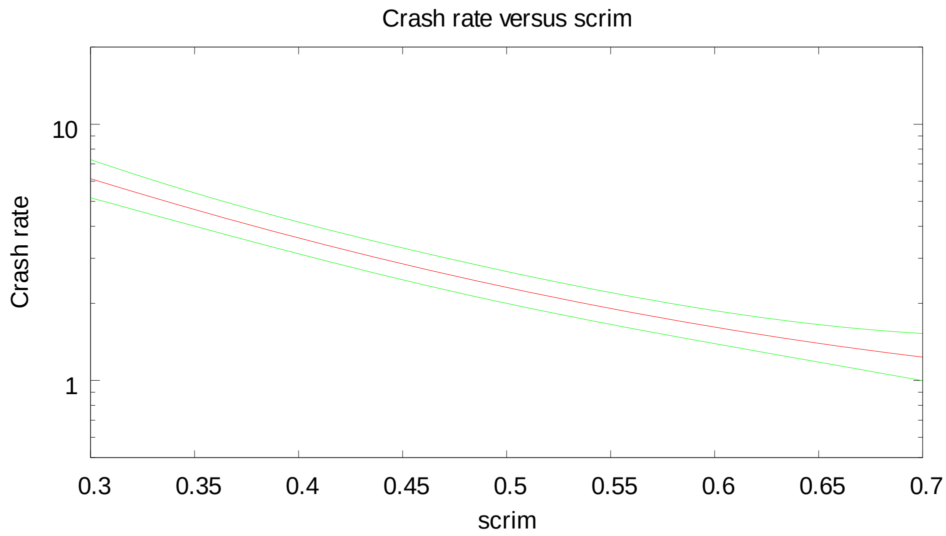
This is roughly similar to the corresponding graph for all casualty crashes.

5.3.15 Wet casualty crashes – curvature



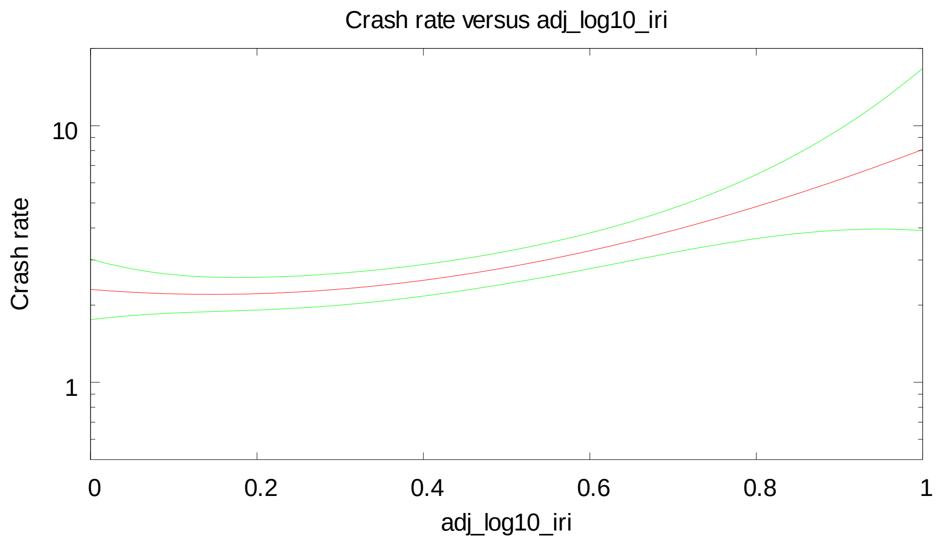
The effect is larger when compared with all casualty crashes.

5.3.16 Wet casualty crashes – scrim

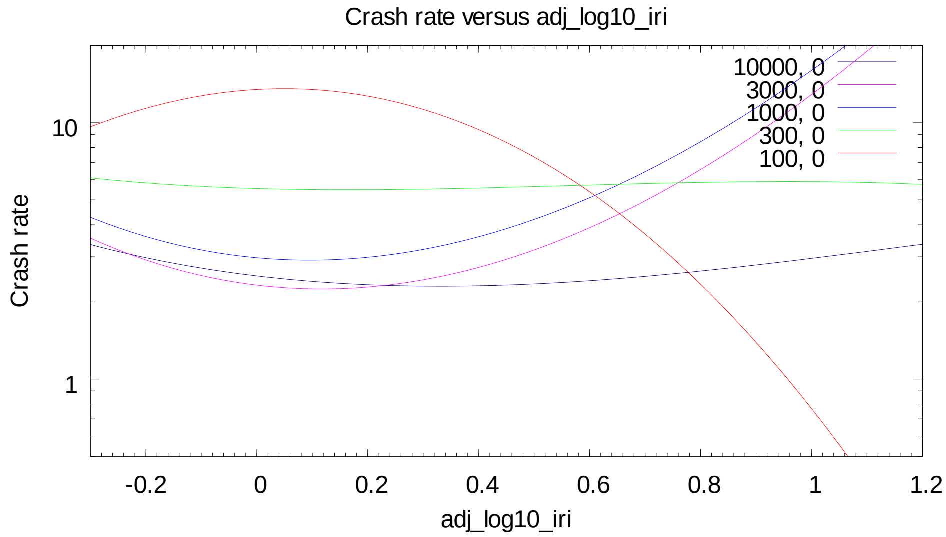


The effect is larger when compared with all casualty crashes.

5.3.17 Wet casualty crashes – adjusted iri

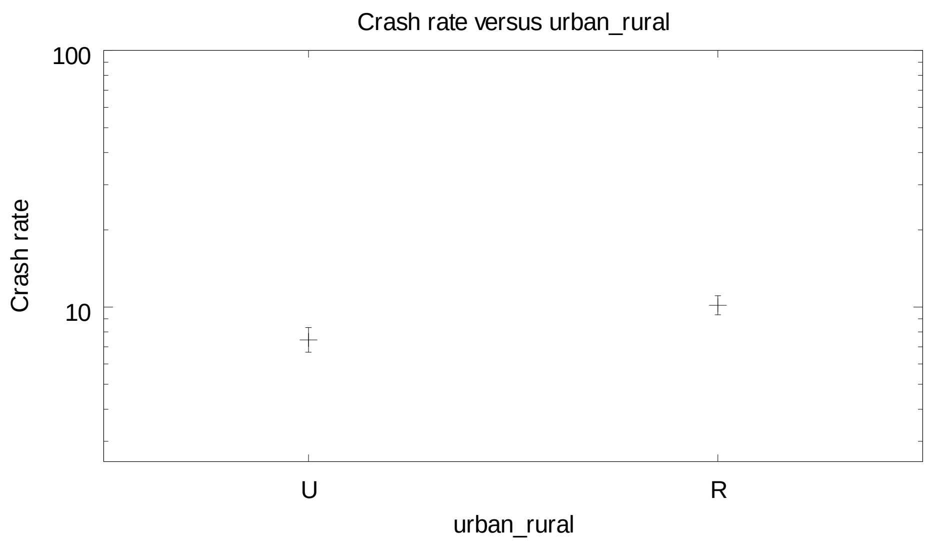


The effect seems a little bigger when compared with the all casualty crashes.



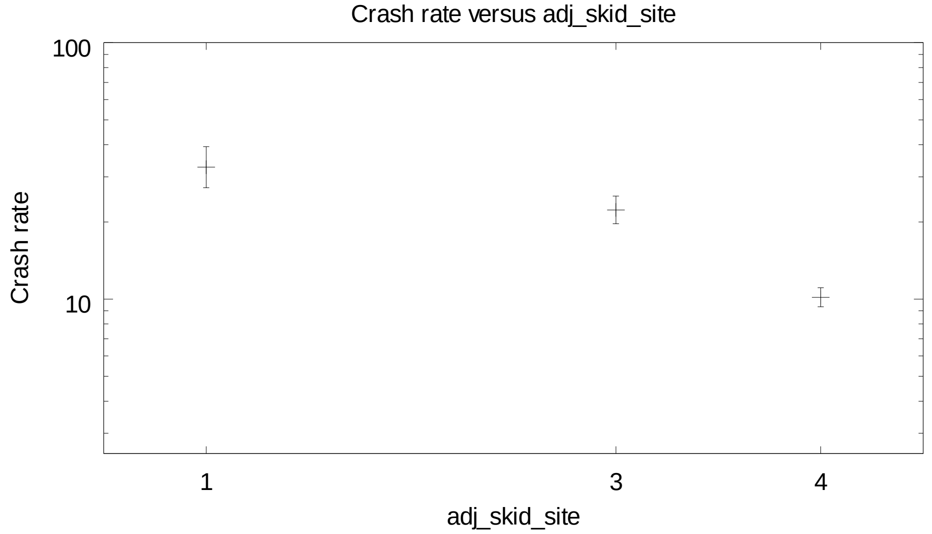
The effects in the graph for different curvature values are similar to that for all casualty crashes except the effects seem a little larger. However with the effects only marginally statistically significant one shouldn't read too much into this graph.

5.3.18 Selected casualty crashes – urban/rural



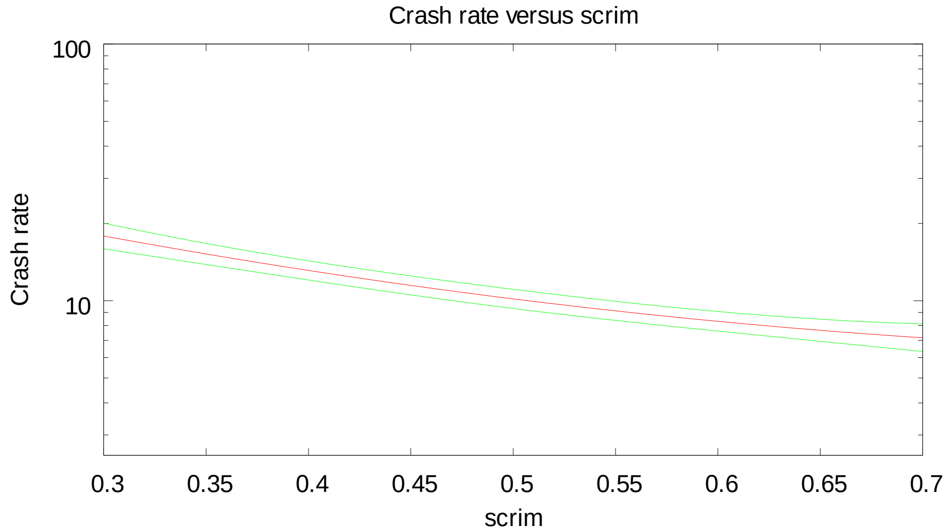
This shows urban slightly lower than rural – possibly because of lower speed limits, but there might be other factors such as better roads and better policing.

5.3.19 Selected casualty crashes – adjusted skid-site



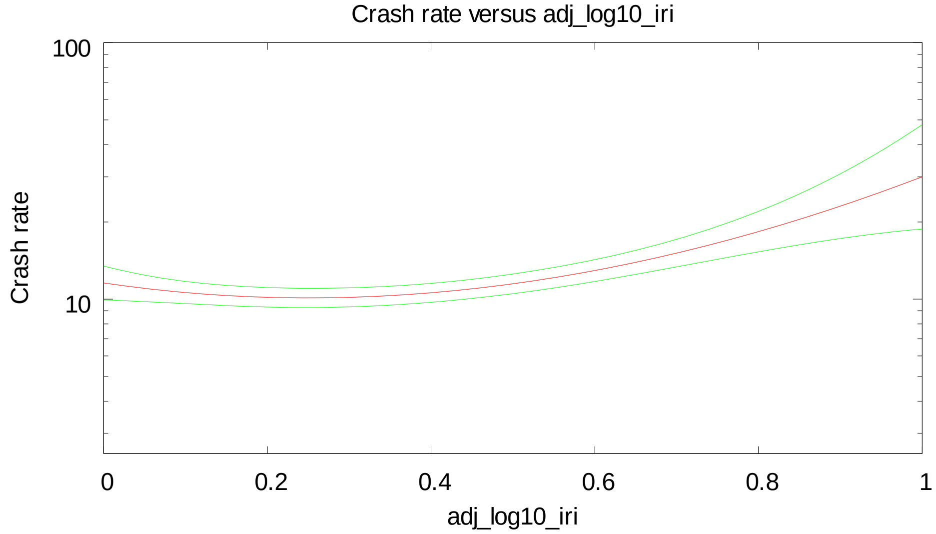
The shows a smaller effect than for all casualty crashes – presumably because some types of intersection crashes are excluded by the selection criteria.

5.3.20 Selected casualty crashes – scrim

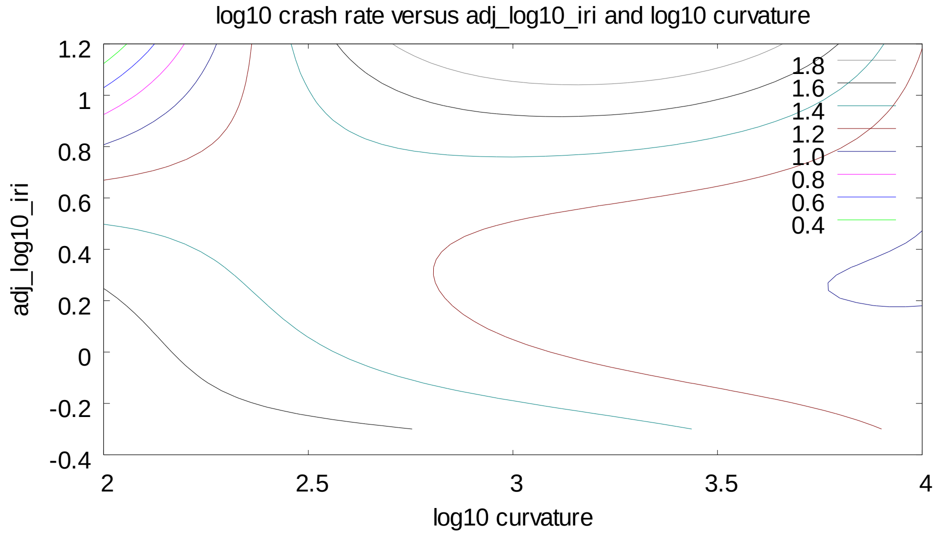


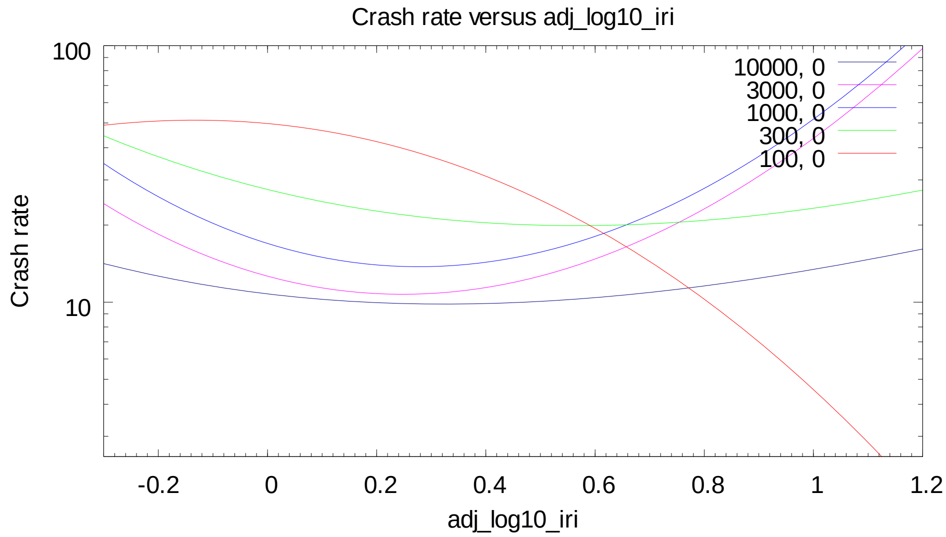
The shows a larger effect than for all casualty crashes – presumably because some types of non-skid related crashes are excluded by the selection criteria.

5.3.21 Selected casualty crashes – adjusted iri



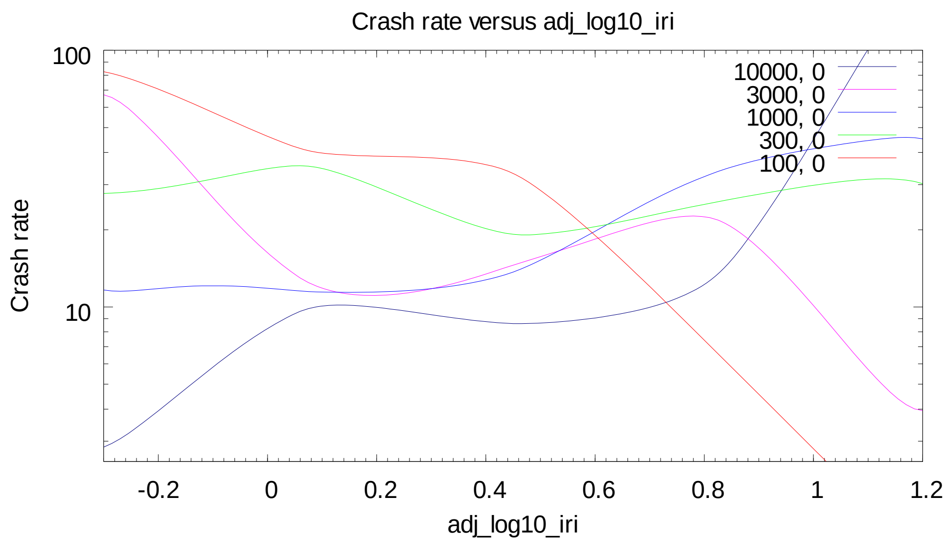
The plot is similar to that for all casualty crashes as are the following two graphs showing the interaction with curvature.



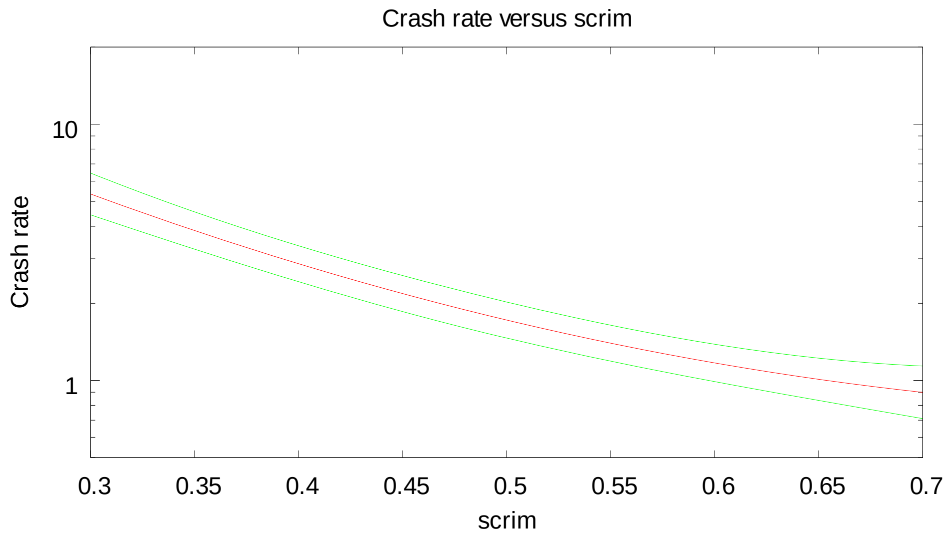


5.3.22 Selected casualty crashes – adjusted iri with tps interaction

I have repeated the analysis with the 5×5 this plate spline and again the results are similar to those for all casualty crashes.

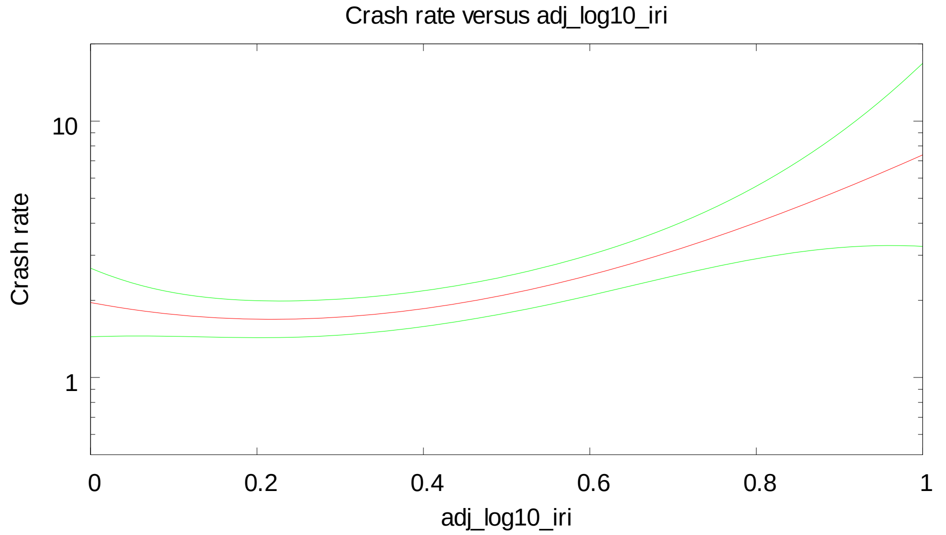


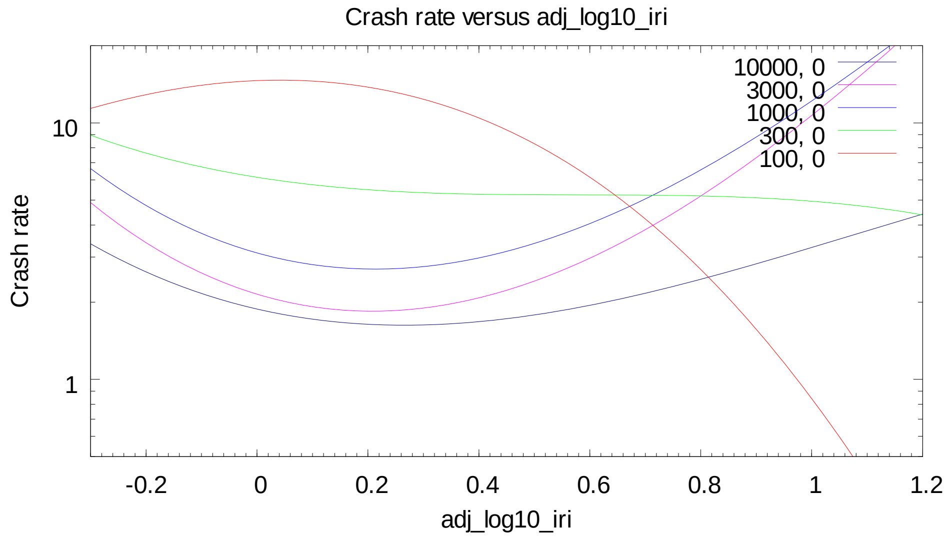
5.3.23 Wet selected casualty crashes - scrim



The effect is similar to that for the wet casualty crashes.

5.3.24 Wet selected casualty crashes – adjusted iri





The graphs are similar to those for the wet casualty crashes.

5.4 The coefficients and calculation of predicted crash rates

The coefficients from the regression are in the worksheet *Coeff* in the spreadsheets *fitted_all.xls*, *fitted_wet.xls*, *fitted_sel_xls* and *fitted_wet_sel.xls* distributed with this report.

The spreadsheets demonstrate how the predicted crash rates are calculated. There are comments in the spreadsheets explaining the calculations.

The road data is entered into the yellow region in the worksheet "Values". If an invalid value is entered into cells C9 to C12 the word "Error" will appear in column D. Remember that if you enter the year 2009 you will get low results because we have only partial data for 2009. The value in cell C25 should be 'y' unless you don't want to adjust the IRI value – for example if the adjusted IRI value has been entered into cell C18. The column "default" shows the default values used in my graphs.

The adjustment to the IRI is carried out in worksheet "Adj IRI" and the results reported in rows 23 and 24 in worksheet "Values". Cell C23 shows the amount to be subtracted from the $log_{10}(IRI)$ and C24 shows what *IRI* would be multiplied by.

Lines 29 to 40 in worksheet "Values" calculate the transformed and bounded values of the predictor variables.

The actual calculation of the predicted crash rate is carried out in worksheet "Calculation".

The result is reported in B42 on the "Values" sheet.

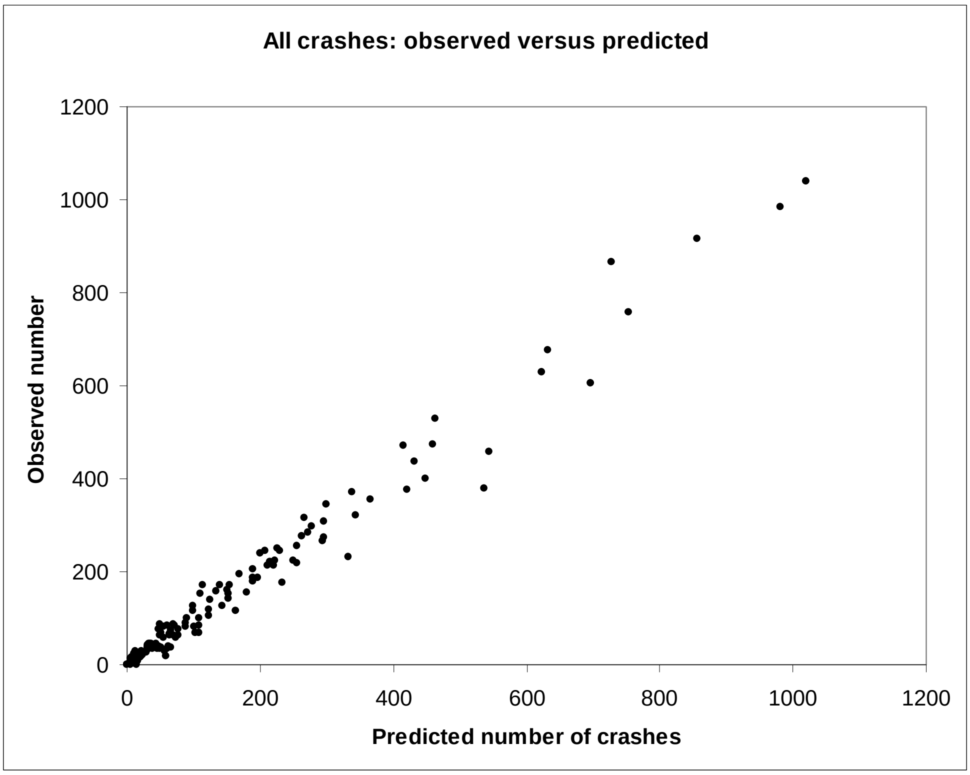
5.5 Comparison of fitted and observed counts

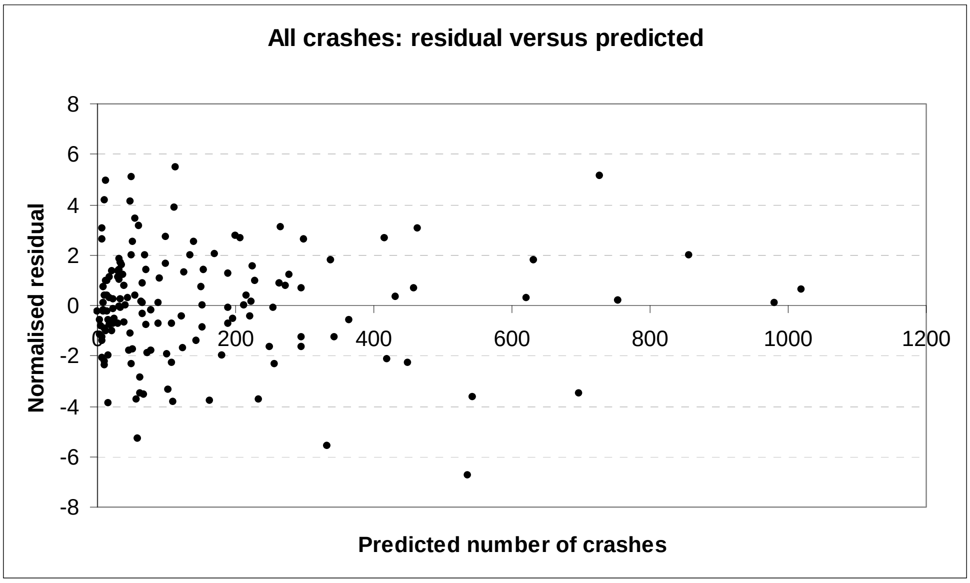
This is an attempt to see how well the model fits.

The highway network was divided into segments by partitioning by carriageway area and state highway number. (The network is divided into 24 *carriageway areas*). That is two sections of road are in the same partition if they are in the same carriageway area and on the same state highway. This gave 148 partitions. Then the model was used to predict the number of crashes in each of these. The observed numbers were compared with the predicted number. The first graph shows the comparison for all crashes and the second graph shows the normalised residual

$$\frac{\text{Observed} - \text{Predicted}}{\sqrt{\text{Predicted}}}$$

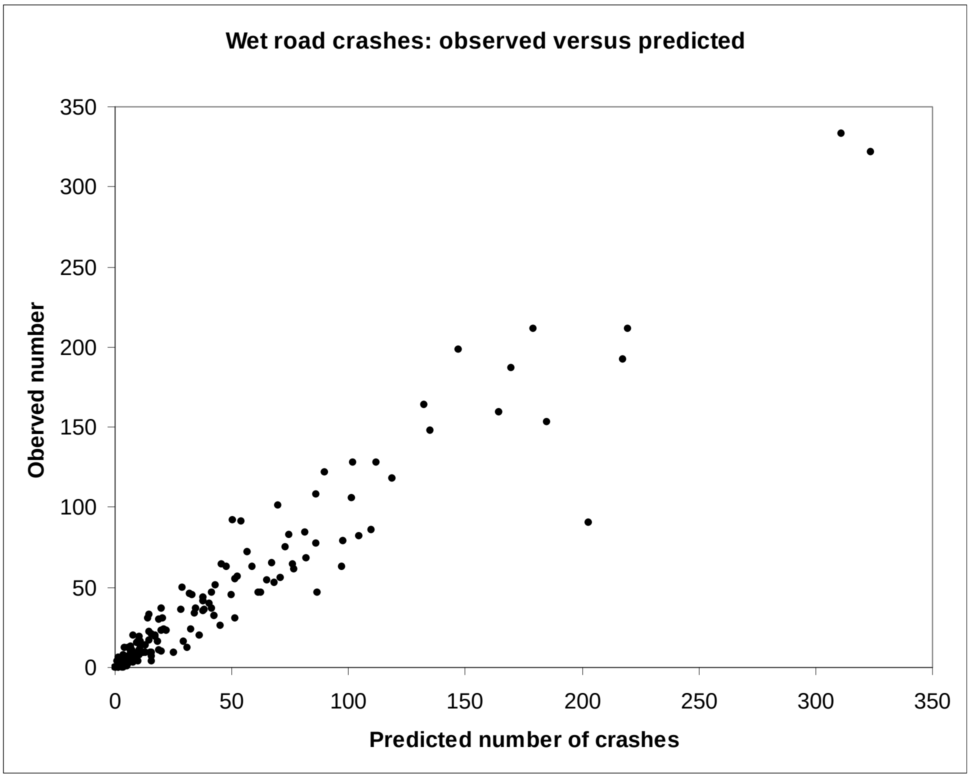
in terms of *predicted*.

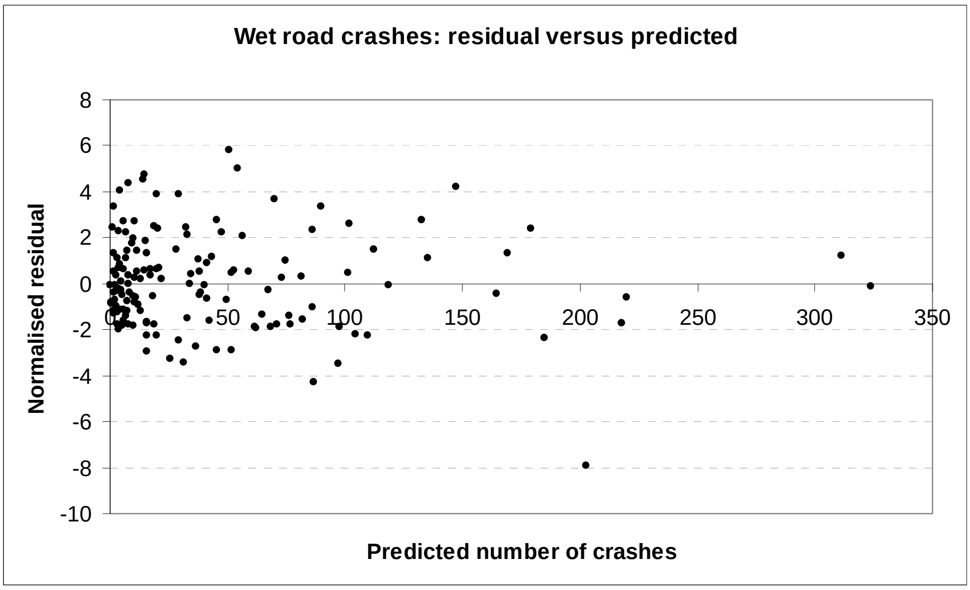




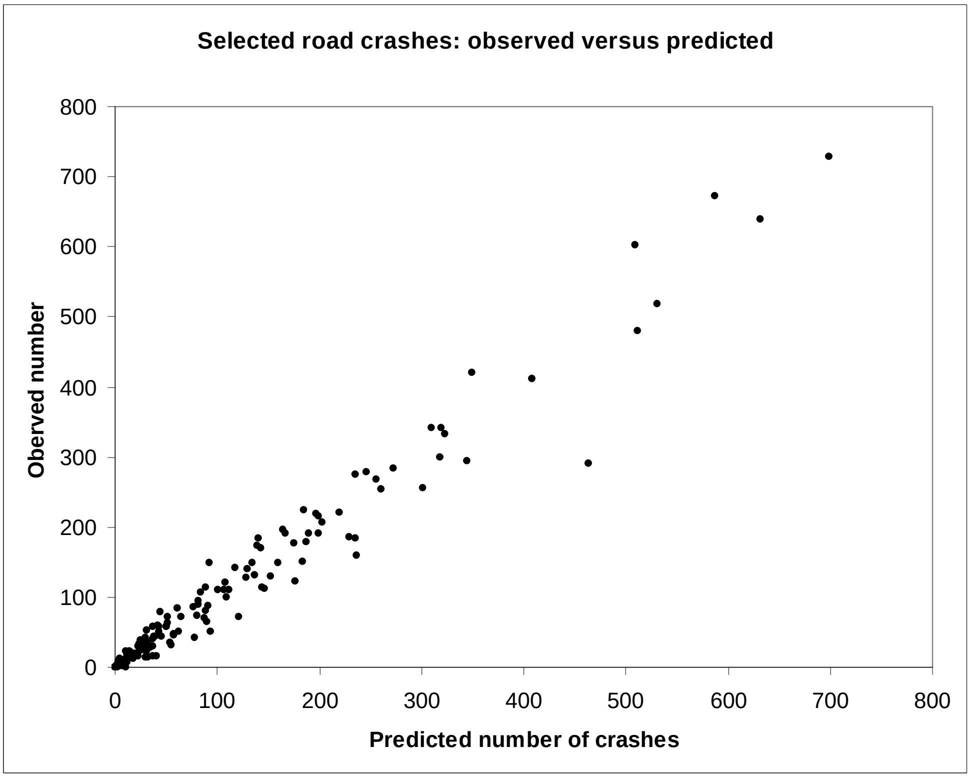
If the model was fitting perfectly there would be few points outside the range –2 to 2. The actual range of points is more like –4 to 4 with a few outside this range (particularly to the left of the graph). So the model doesn't fit perfectly.

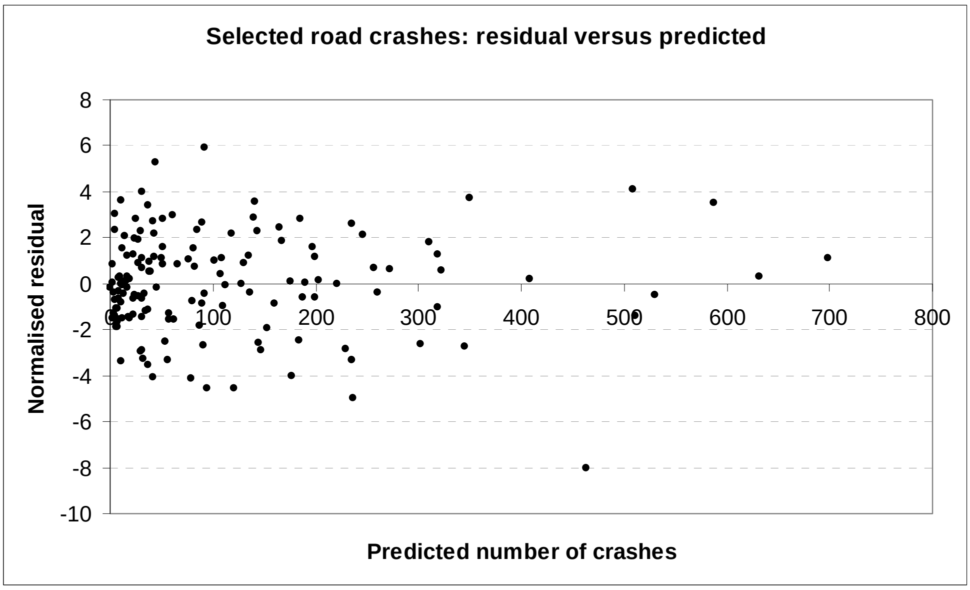
Here are the corresponding graphs for the wet road crashes.





The quality of the fit is about the same as for all crashes as it is for selected crashes below.





One could identify the particular points where the fit is bad and see if there was a data problem or where there were special risk factors. There is one point that possibly stands out as being low in the wet and selected graphs and, to a lesser extent in the all crashes graph. This corresponds to highway 25 in the East Waikato carriageway area. I haven't investigated the possible reasons for this low value.

The following table shows the numbers of crashes in each of the 3 categories of crashes I have just considered and the sum of squares of the normalised residuals. If nothing had been fitted and the Poisson model was true these would have chi-squared distributions with 148 (the number of categories) degrees of freedom. So the values would be close to 148. Because we have fitted parameters the number of degrees of freedom must be reduced. We have fitted 45 parameters (including the constant) but it would be wrong to reduce the number by 45 since there has been a lot of amalgamation of data. Choose 23 as a ball-park figure. So we have 125 degrees of freedom. So in the case of all crashes we have a value that is too large by a factor of 5.4. This is where my suggested adjustment to the significance tests comes from.

	All crashes	Wet crashes	Selected crashes
Number of crashes	22870	6476	16516
Chi-squared value	675	624	693

It is a little surprising that the chi-squared values remains high for the wet-crashes. We have a lot less crashes so we would expect the randomness in the crashes to begin to mask whatever is causing the lack of fit.

The situation changes very little when we include a year × region interaction, see section 7.4. So problem is probably not one of fluctuation of weather patterns in different regions.

One possibility is that traffic patterns over a day and over a year are rather different from rainfall patterns over a day or year and this might introduce additional variation into the wet crash data.

5.6 Effect of the averaging

The model fitted in section 5 averages the prediction from the log Poisson model over 21 adjacent 10 metre segments, that is the segments within a range of 10 segments of the one in which we are actually making the estimate. This section investigates the effect of altering the number of segments being averaged. The following table shows the log-likelihood and the type I chi-squared values for the all crashes model for various values of averaging length. I have shown the maximum values in each line in bold.

Effect	Averaging length (metres)				
	410	210	110	50	30
Log likelihood	-152161	-151293	**-150680**	-150803	-150917
year	**528**	526	525	521	519
region	**678**	665	655	657	657
urban rural	383	486	**552**	546	523
adjusted skid-site	4318	6289	**7319**	6541	6053
OOCC	**5871**	5400	4419	3401	3108
curvature	263	460	752	1125	**1228**
ADT	430	518	590	632	**640**
scrim	250	**265**	237	249	256
gradient	62	66	**69**	64	62
adjusted IRI	40	109	160	**277**	275
curvature × adj IRI	85	107	102	117	**120**

The analyses in the report use an averaging length of 210. If we just look at log-likelihood, then we might decide that an averaging length of 110 in more appropriate. The different chi-squared values don't show any consistent pattern and it is hard to make much sense of them.

6 Application of the model

6.1 What-if study for skid resistance

This is an illustration of the use of the model for seeing the effect of improving skid resistance.

Consider increasing the skid resistance of the sections of road classified as skid-site 2 (curve with less than 250m radius or gradient greater than 10%; investigatory SCRIM level 0.5). Choose one of the following minimum levels for scrim: 0.4, 0.5, 0.6. If the actual scrim value is below this the road surface is upgraded to raise the scrim value to this amount; otherwise it is left the same. We can choose to do this for all the roads that fall short of the minimum scrim level or only those with ADT above some prescribed level.

Our what-if study is for 2008. The following table shows the actual road and crash data for 2008 for the road segments classified as skid-site 2.

Length of road sides	2056
Number of crashes	567
Number of wet crashes	207

The next table shows the reduction in the predicted number of crashes 2008 if the upgrade had been done before 2008.

			All crashes		Wet crashes	
minimum scrim	fix for ADT≥	fix length	predicted crashes	saved crashes	predicted crashes	saved crashes
0	0	0	565	0	219	0
0.4	0	440	547	18	194	24
0.5	0	1292	506	59	159	60
0.6	0	1911	452	113	124	95
0.4	1000	350	548	17	195	23
0.5	1000	955	509	56	162	57
0.6	1000	1336	460	105	130	88
0.4	5000	69	558	7	210	9
0.5	5000	179	541	24	196	22
0.6	5000	228	521	44	184	35

The first line is for no upgrade, the rest are for the values of minimum scrim value and ADT shown in the first two columns. It is supposed that the two sides of the road are handled independently and the column *fix length* shows the length of side that needs to be upgraded. The analysis is carried using the model for all crashes and the model for all wet crashes. The table shows the predicted number of crashes for each model and then the reduction in the number of crashes compared with the first line.

The table suggests one should consider raising minimum scrim to 0.6 on skid-site 2 roads with ADT ≥ 1000.

The numbers of saved crashes are a lot higher than in the 2004 study, when I looked at saved crashes for 2001. There are several reasons for this. I am classifying more crashes as wet than in the 2004 study and most crashes have been located as opposed to about 75% in the 2004 study. However, the main reason seems to be that much more road needs fixing in 2008 than in 2001 (or has there been a change in the calibration of the SCRIM readings).

This example should be regarded as somewhat tentative and I have not tried to estimate accuracy.

6.2 What-if study for IRI

This is an illustration of the use of the model for seeing the effect of reducing roughness.

Consider decreasing roughness on roads with absolute radius of curvature greater than 500 and less than 5000 metre. Choose one of the following maximum levels for IRI: 2.00, 3.98, 7.94. If the actual IRI value is above this the road surface is upgraded to lower the IRI value to this amount; otherwise it is left the same. We can choose to do this for all the roads that are above the maximum IRI level or only those with ADT above some prescribed level.

Our what-if study is for 2008. The following table shows the actual road and crash data for 2008 for the road segments in the curvature range 500 to 5000 metre.

Length of road sides	7348
Number of crashes	1229

The next table shows the reduction in the predicted number of crashes 2008 if the upgrade had been done before 2008. The table is using adjusted IRI, but for this curvature range the adjustment has little effect.

maximum adjusted IRI	fix for ADT≥	fix length	predicted crashes	saved crashes
	0	0	1182	0
7.94	0	30	1180	2
3.98	0	815	1148	33
2.00	0	4212	1048	134
7.94	1000	20	1180	1
3.98	1000	602	1150	31
2.00	1000	3242	1055	126
7.94	5000	5	1181	1
3.98	5000	169	1163	19
2.00	5000	953	1106	76

The first line is for no upgrade, the rest are for the values of maximum adjusted IRI value and ADT shown in the first two columns. It is supposed that the two sides of the road are handled independently and the column *fix length* shows the length of side that needs to be upgraded. The analysis is carried using the model for all crashes. The table shows the predicted number of crashes for each model and then the reduction in the number of crashes compared with the first line.

Again this example should be regarded as somewhat tentative and I have not tried to estimate accuracy.

7 Variations on model

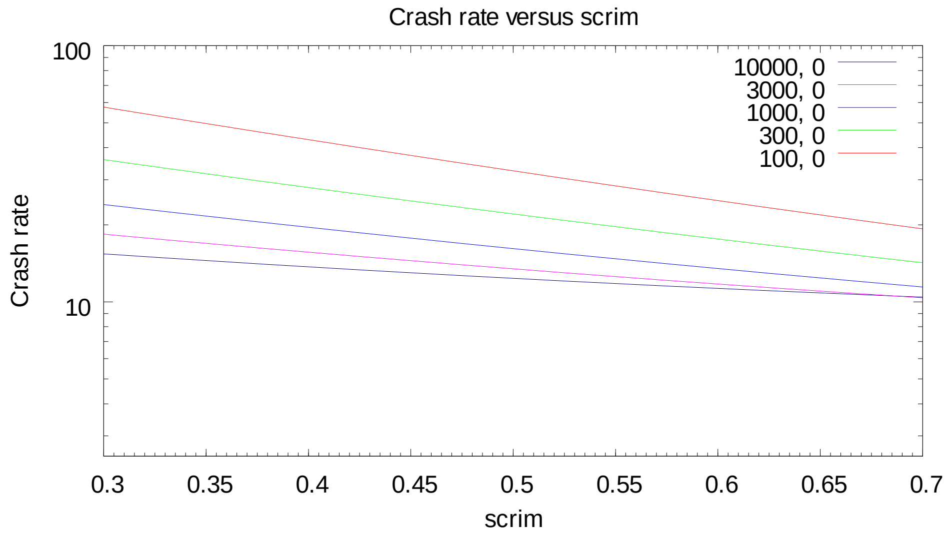
The following sections explore a number of minor variations of the model of section 5.

7.1 Include curvature × SCRIM interaction

In this section I look for an interaction between curvature and SCRIM. That is, does the effect of SCRIM change according to curvature. In the 2004 study, I did not find any effect. I this report I am fitting a very simple $\log_{10}$(curvature) × scrim interaction for both the all-crash data and the wet-crash data. In both cases, it is fairly marginal whether the results are statistically significant. I have given the analysis of variance tables with the last line corresponding to the interaction term. I expect the amount we need to inflate the significance levels to be rather less the 5.4 suggested in section 4.2 since curvature is a rapidly changing predictor variable. So, probably, we should regard the effect as statistically significant. Both graphs of the effect of SCRIM on crash rate at various curvatures suggest that the effect of SCRIM is higher on curves rather than on straight or near straight roads.

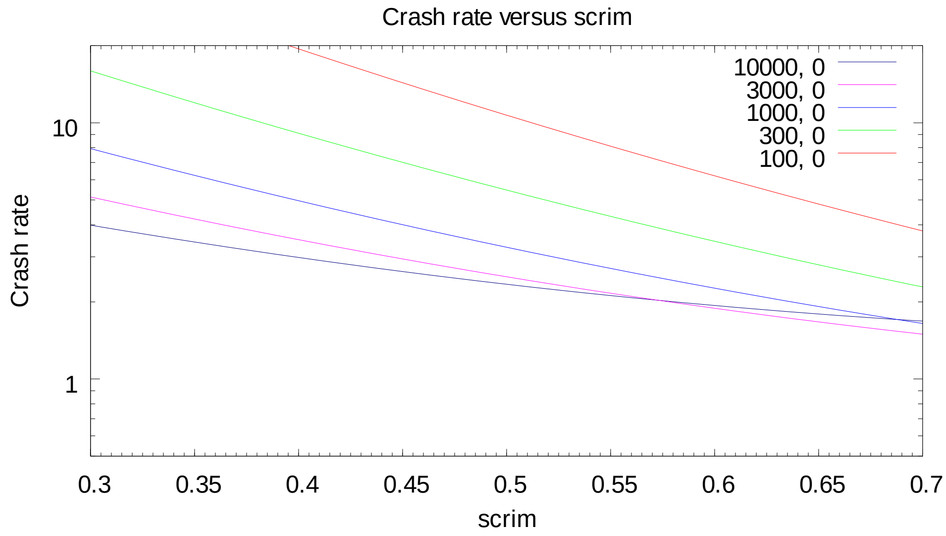
7.1.1 All crashes

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	514.67	533.09
region	13	27.7	301.76	673.14
urban_rural	1	6.63	27.538	493.38
adj_skid_site	2	9.21	4693.1	6298.4
poly3_bound_OOCC	3	11.3	455.22	5494.1
poly2_bound_log10_abs_curvature	2	9.21	107.4	470.42
poly2_log10_ADT	2	9.21	575.29	519.67
poly2_scrim-0.5000	2	9.21	61.688	262.21
poly3_bound_abs_gradient	3	11.3	49.306	66.255
poly3_bound_adj_log10_iri	3	11.3	82.228	109.61
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	110.42	106.86
bound_log10_abs_curvature × scrim-0.5000	1	6.63	22.107	22.107



7.1.2 Wet crashes

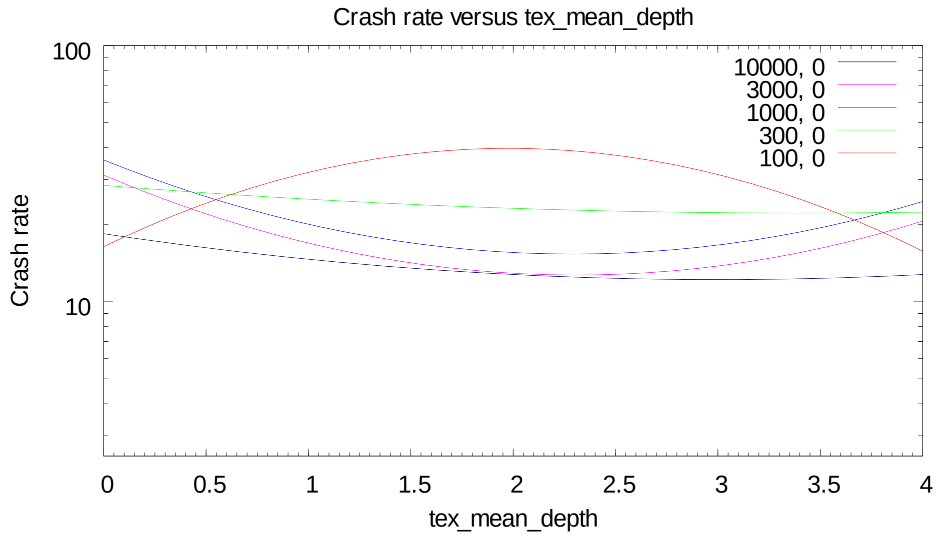
Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	191.62	138.25
region	13	27.7	228.46	508.26
urban_rural	1	6.63	39.01	6.7804
adj_skid_site	2	9.21	687.16	1014.3
poly3_bound_OOCC	3	11.3	246.82	4436.6
poly2_bound_log10_abs_curvature	2	9.21	70.925	464.02
poly2_log10_ADT	2	9.21	137.65	104.25
poly2_scrim-0.5000	2	9.21	104.52	416.16
poly3_bound_abs_gradient	3	11.3	73.606	84.394
poly3_bound_adj_log10_iri	3	11.3	32.22	42.277
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	46.788	42.724
bound_log10_abs_curvature × scrim-0.5000	1	6.63	28.369	28.369



7.2 Include texture and interaction

I repeat the analysis for all crashes with additional terms for mean texture depth and its interaction with curvature and show the graphs of crash rate versus mean texture depth for various curvatures. I show the analysis of variance table for just the terms involving the mean texture depth.

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
poly2_bound_tex_mean_depth	2	9.21	29.009	75.174
poly2_bound_log10_abs_curvature × poly2_bound_tex_mean_depth	4	13.3	58.924	58.924



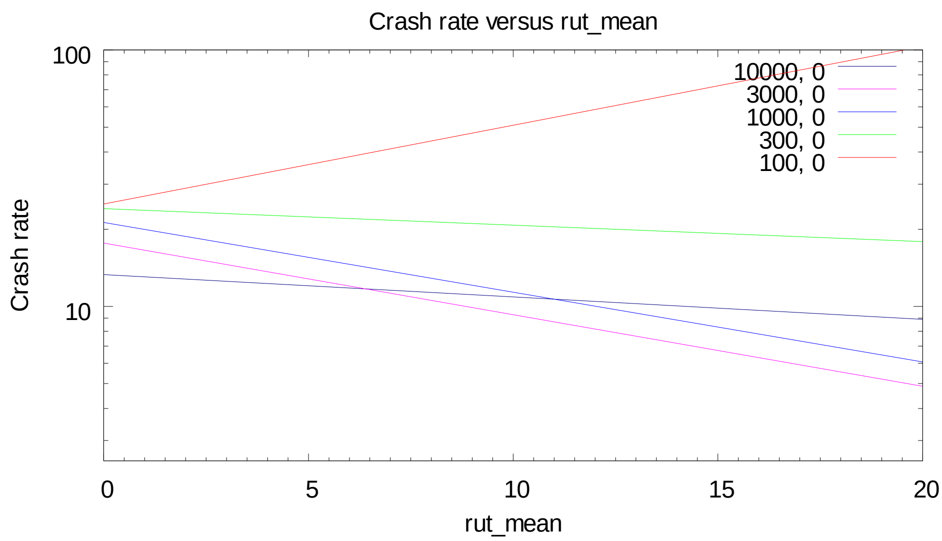
These terms do appear to be statistically significant. The graph for the 100 radius curvature and the upwards slope of the graph for high values of mean texture depth may be a result of polynomial fit.

However, I am very suspicious of any results concerning texture for the reasons stated in section 3.17.

7.3 Include mean rut depth plus interaction

I repeat the analysis for all crashes with additional terms for the mean rut depth and its interaction with curvature and show the graphs of crash rate versus mean rut depth for various curvatures. I show the analysis of variance table for just the terms involving the mean rut depth.

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
poly2_sqrt_bound_rut_mean	1	6.63	37.159	39.661
poly2_bound_log10_abs_curvature × bound_rut_mean	2	9.21	58.43	58.43



The interactions terms do appear to be statistically significant. However, the results don't make much sense and some more analysis is required to see what is going on.

7.4 Include year × region interaction

The previous analyses are robust against variations in crash rate due to changes in reporting rates, policing, changes in traffic volume, weather provided they were the same across the whole network. Likewise, variations from region to region would not be a problem provided they remained the same from year to year. However, if the year to year changes varied from region to region then there could be a problem.

We can, at least partially, compensate for year to year changes that vary from region to region by including a year × region interaction term in the analysis. This substantially increases the computer time required, but is feasible.

I have run this analysis for all crashes and wet crashes. The estimates are almost identical to the corresponding ones in section 5. Here is the analysis of variance table for all crashes.

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	96.854	521.06
region	13	27.7	44.891	675.18
urban_rural	1	6.63	27.946	484.56
adj_skid_site	2	9.21	4705.6	6291.9
poly3_bound_OOCC	3	11.3	452.66	5402.2
poly2_bound_log10_abs_curvature	2	9.21	110.14	460.85
poly2_log10_ADT	2	9.21	579.81	517.45
poly2_scrim-0.5000	2	9.21	225.5	266.24
poly3_bound_abs_gradient	3	11.3	45.924	65.845
poly3_bound_adj_log10_iri	3	11.3	81.861	108.67
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	105.56	107.61
year × region	117	155.5	274.61	274.61

Apart from year and region the entries are almost the same as in the corresponding table in section 5.2.1.

We can also repeat the fit test of section 5.5.

	All crashes	Wet crashes
Number of crashes	22870	6476
Chi-squared value	672	619

The results are almost unchanged.

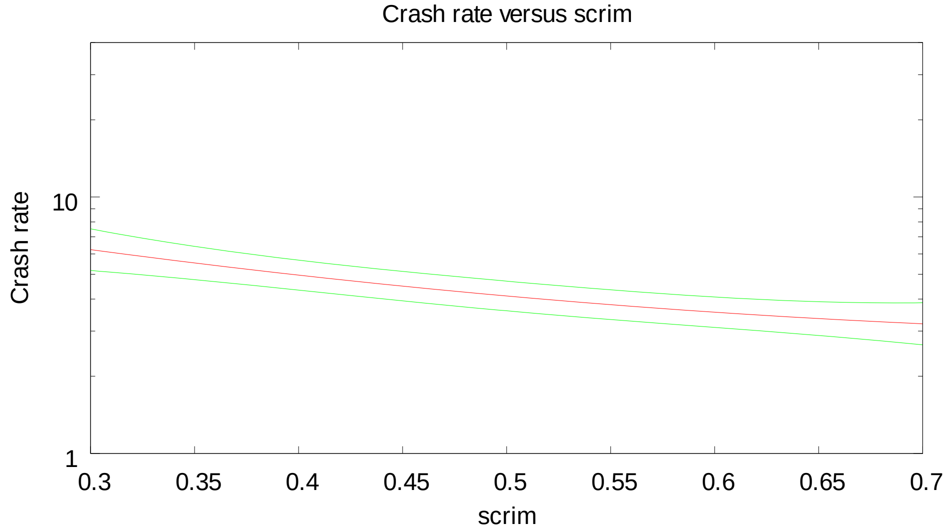
So changing patterns of reporting rates, policing, changes in traffic volume, weather do not seem to be a problem. There could of course be more subtle changes which these analyses would not detect. See section 5.5.

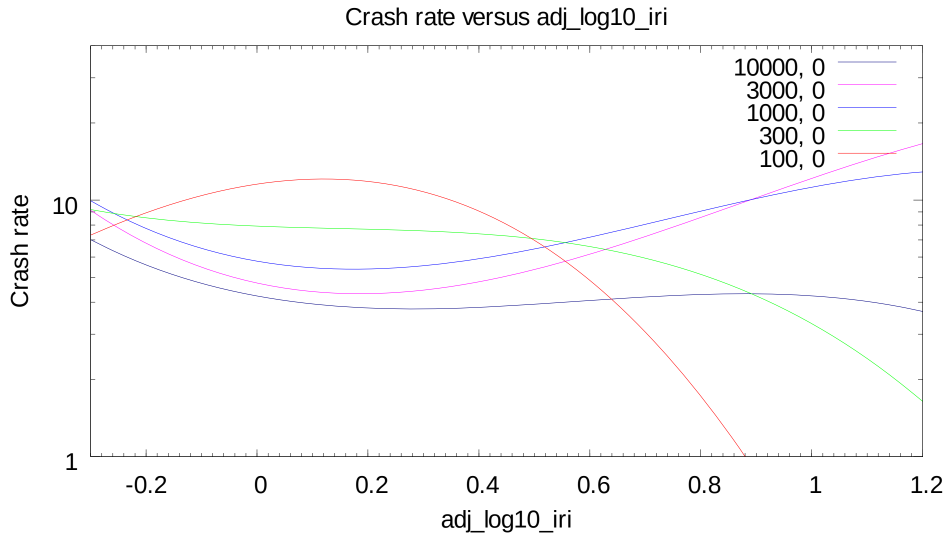
7.5 Serious and fatal crashes

The previous analyses have used all reported crashes that have involved at least one fatality, serious injury or minor injury. Minor injury crashes have a low reporting rate and this is may be variable over time and over the country. With ten years data it is possible to carry out some of the analyses with only fatal and serious injury data. Crash numbers were given in section 3.4. I have repeated the analysis for all crashes and wet crashes using only the serious and fatal data. The following two sections give the analysis of variance tables and the predictor graphs for SCRIM and IRI.

7.5.1 *All serious/fatal crashes*

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	154.58	127.38
region	13	27.7	69.341	196.97
urban_rural	1	6.63	64.973	1.3263
adj_skid_site	2	9.21	1167.7	1379.8
poly3_bound_OOCC	3	11.3	147.83	1835.6
poly2_bound_log10_abs_curvature	2	9.21	22.536	153.84
poly2_log10_ADT	2	9.21	275.36	250.43
poly2_scrim-0.5000	2	9.21	61.674	67.457
poly3_bound_abs_gradient	3	11.3	9.0996	11.754
poly3_bound_adj_log10_iri	3	11.3	13.212	8.7159
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	21.354	21.354

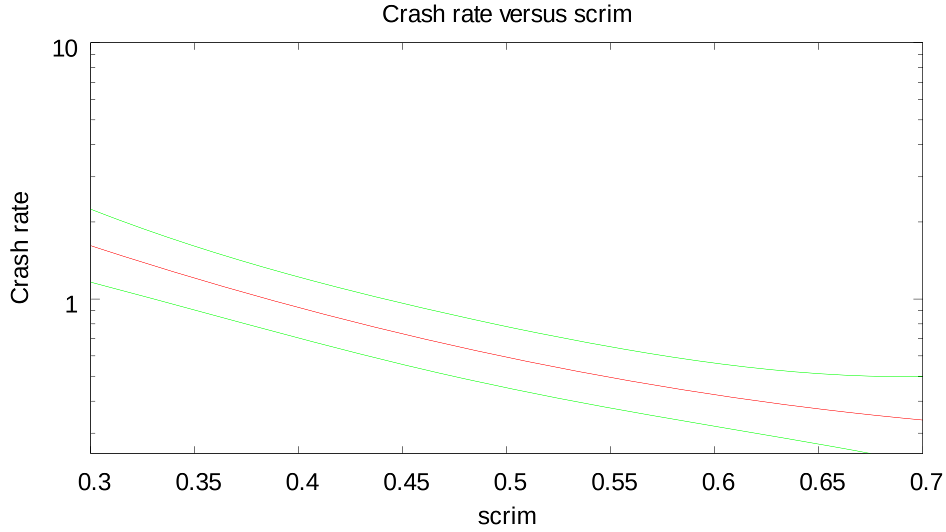


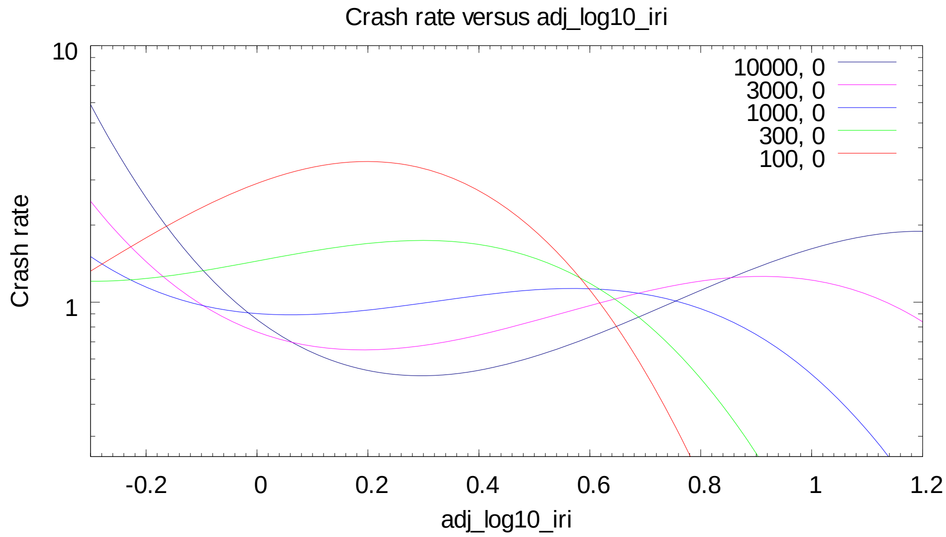


The gradient and IRI terms are barely statistically significant, even if we don’t apply the adjustment factor discussed in section 7.5.3 below. However, the predictor graphs are remarkably similar to those for all crashes in sections 5.3.9 and 5.3.11. The other graphs, not shown here, are also similar.

7.5.2 Wet serious/fatal crashes

Predictor variable	**df**	**1% pt.**	**Chi-squared**	
			Type III	Type I
year	9	21.7	82.281	47.539
region	13	27.7	52.372	132.35
urban_rural	1	6.63	39.751	33.754
adj_skid_site	2	9.21	196.88	235.06
poly3_bound_OOCC	3	11.3	66.906	1160.8
poly2_bound_log10_abs_curvature	2	9.21	11.109	111
poly2_log10_ADT	2	9.21	43.878	26.218
poly2_scrim-0.5000	2	9.21	111.01	115.24
poly3_bound_abs_gradient	3	11.3	13.693	15.982
poly3_bound_adj_log10_iri	3	11.3	4.6432	2.0679
poly2_bound_log10_abs_curvature × poly2_bound_adj_log10_iri	4	13.3	12.69	12.69





The IRI terms are now not statistically significant even without any adjustment term. The graph for SCRIM is similar to the corresponding one in section 5.3.16 and the graph for IRI somewhat similar to the one is section 5.3.17.

7.5.3 Fit test

We can repeat the fit test of section 5.5.

	All crashes	Wet crashes
Number of crashes	7088	1853
Chi-squared value	287	332

The chi-squared values are now substantially reduced. I expect this to be because the random variation of the crash data is partially masking whatever is causing the lack of fit.

This suggests an adjustment factor of 2.3 for all crashes and a slightly higher value for the wet crashes.

8 IRI variance terms

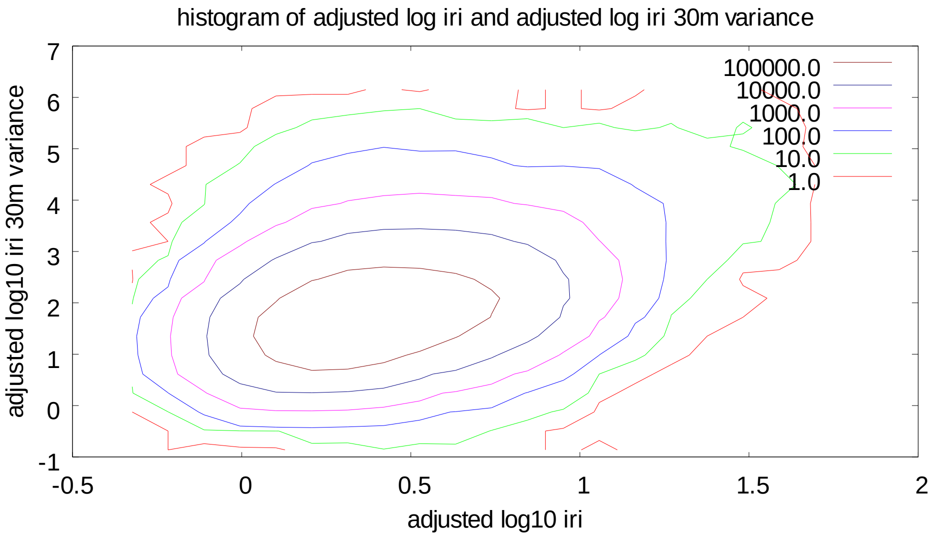
The IRI variance measurements measure the IRI at various wavelengths. These are available for 2006 to 2009 for wavelengths of 3 metres, 10 metres and 30 metres.

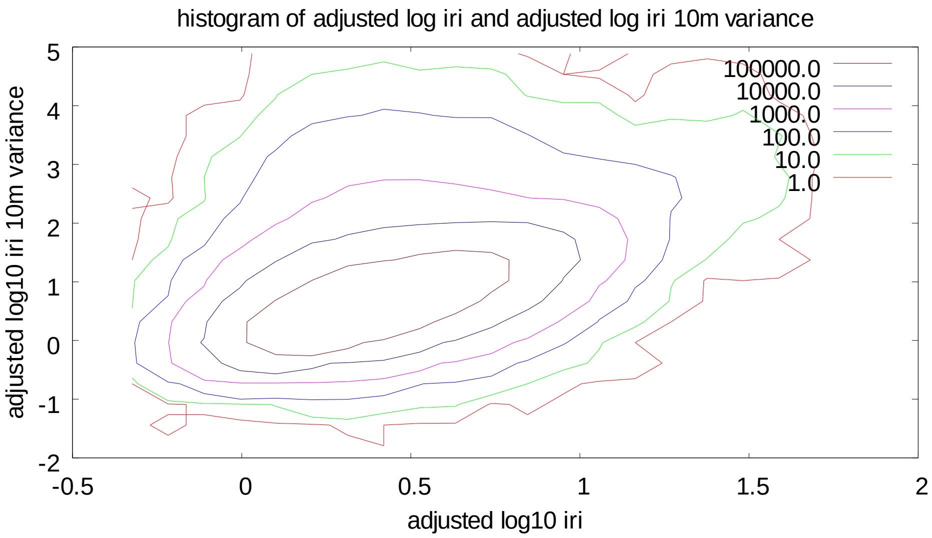
I repeated the analyses for all crashes with the IRI term being replaced, in turn, by each of these IRI variance measurements. I also carried out the analysis with the regular IRI measurement for the same years as a comparison. In each case I adjusted for curvature and gradient as before. To simplify the analyses I used the spline model fitted separately for the increasing and decreasing sides of the road as in the 2004 analyses.

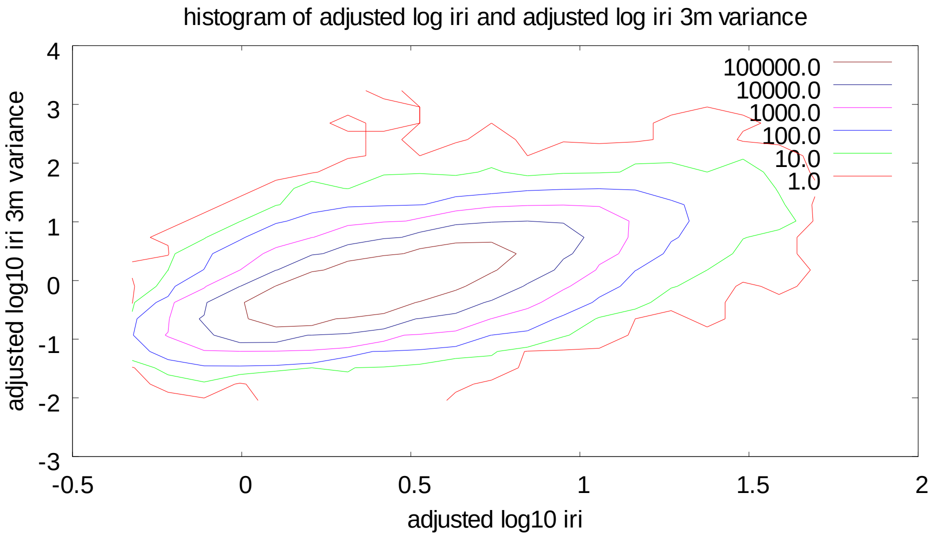
The following table shows the R-squared values for the fit of the dependence of the IRI variance on curvature and gradient. The dependence is strongest for the 10 and 30 metre wavelength data and weakest for the 3 metre wavelength data and the usual IRI.

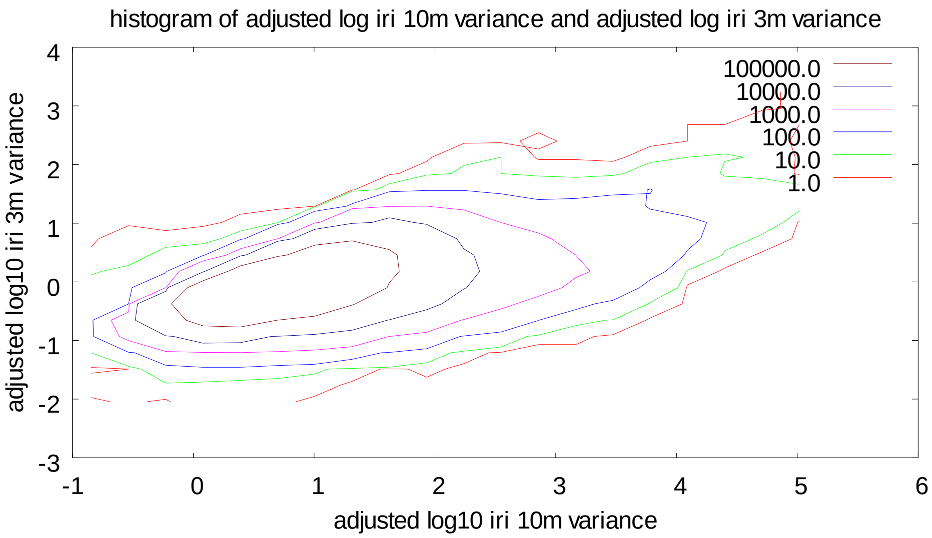
	Wavelength	3 metre	10 metre	30 metre	IRI
R-squared for curvature/gradient adjustment	increasing	6%	16%	18%	7%
	decreasing	7%	16%	19%	8%

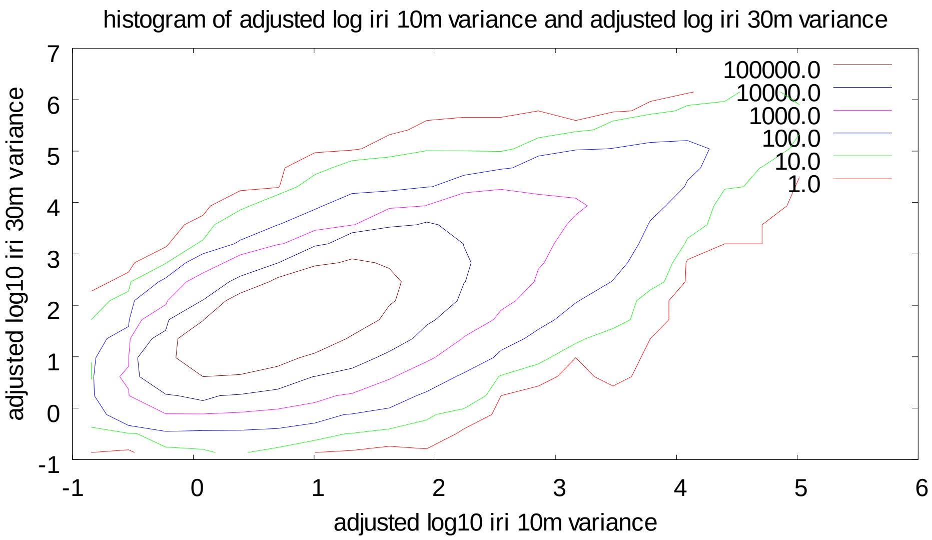
I plotted two dimensional histograms of each set of pairs of the adjusted IRI measurements. Here are the pair-wise plots.

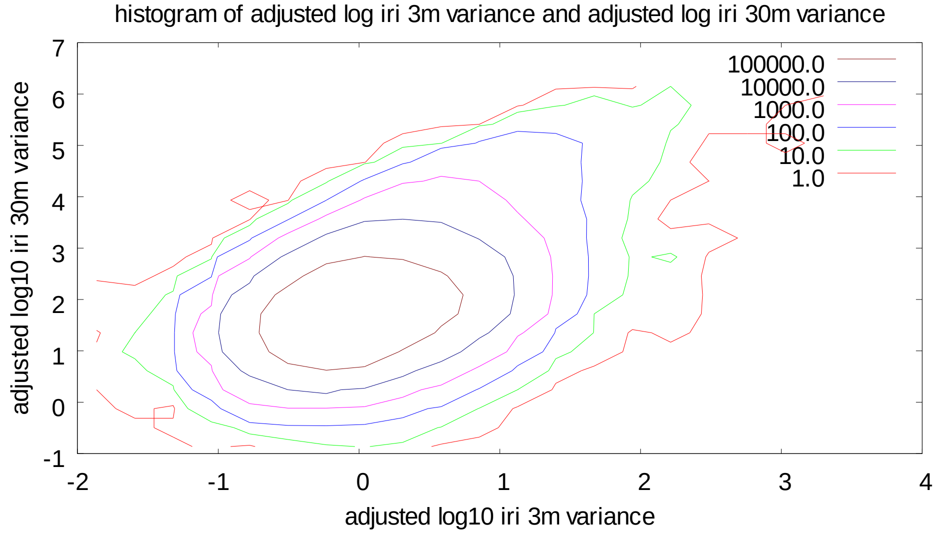





histogram of adjusted log iri and adjusted log iri 3m variance
100000.0
10000.0
1000.0
100.0
10.0
1.0
adjusted log10 iri 3m variance
adjusted log10 iri
4
3
2
1
0
-1
-2
-3
-0.5
0
0.5
1
1.5
2


histogram of adjusted log iri 10m variance and adjusted log iri 3m variance
100000.0
10000.0
1000.0
100.0
10.0
1.0
adjusted log10 iri 3m variance
adjusted log10 iri 10m variance
4
3
2
1
0
-1
-2
-3
-1
0
1
2
3
4
5
6


histogram of adjusted log iri 10m variance and adjusted log iri 30m variance
100000.0
10000.0
1000.0
100.0
10.0
1.0
adjusted log10 iri 30m variance
adjusted log10 iri 10m variance
7
6
5
4
3
2
1
0
-1
-1
0
1
2
3
4
5
6



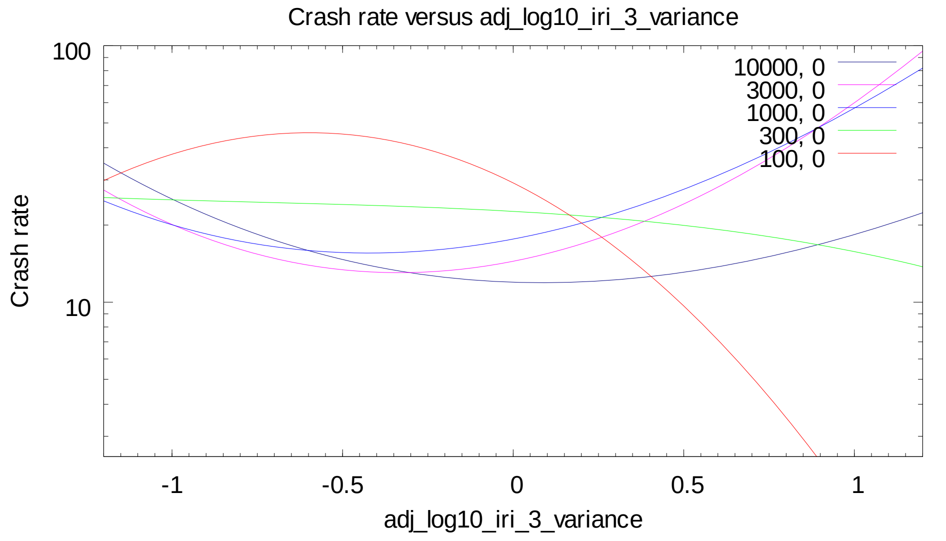
In all cases there is an obvious relationship between each member of the pair.

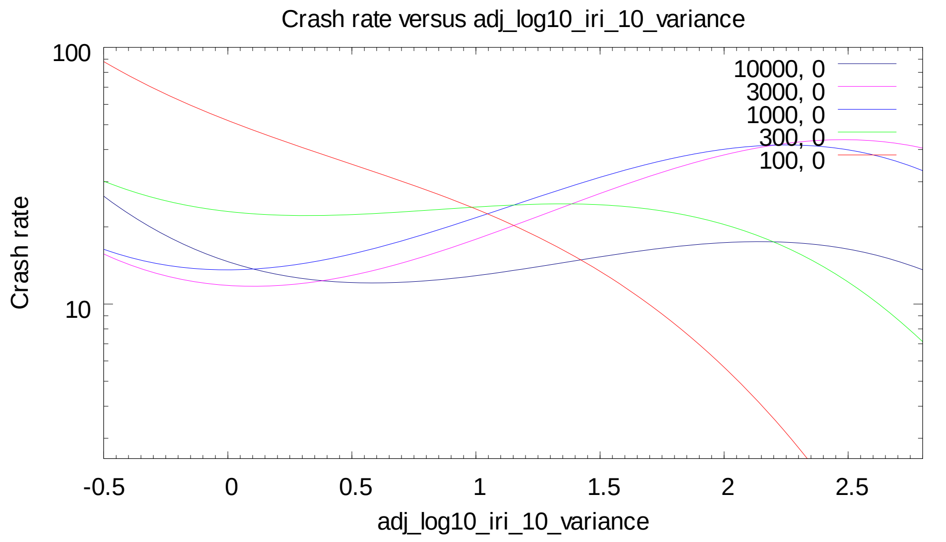
I repeated the analysis of variance with the usual IRI and with the usual IRI replaced, in turn, by each of the three IRI variance terms. The following table shows the log-likelihood for each of the four analyses relative to the value for the usual IRI and the type I chi-squared value for the IRI term and for its interaction with curvature.

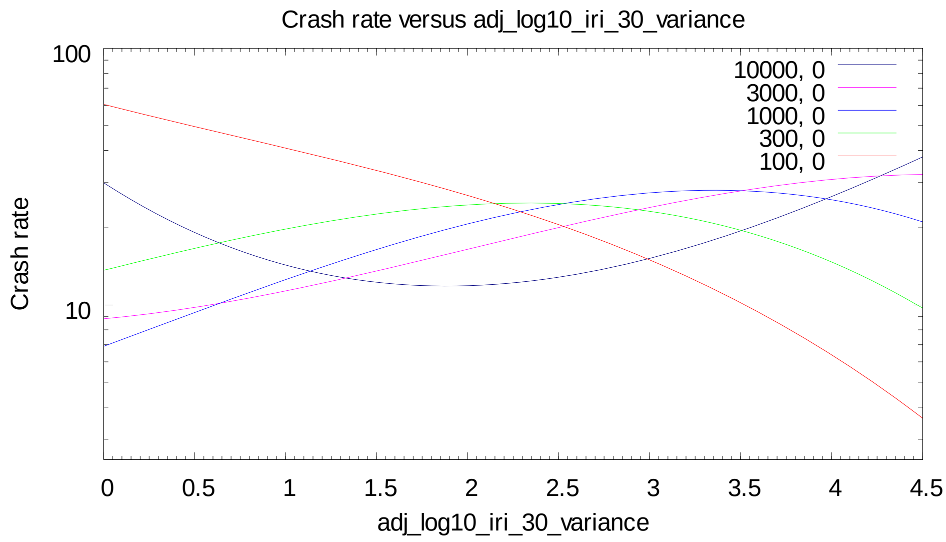
	Wavelength	3 metre	10 metre	30 metre	IRI
Log-likelihood		-9.4	1.7	-17.1	0
Chi-squared in ANOVA table	IRI term	71	71	34	78
	interaction	50	64	44	57

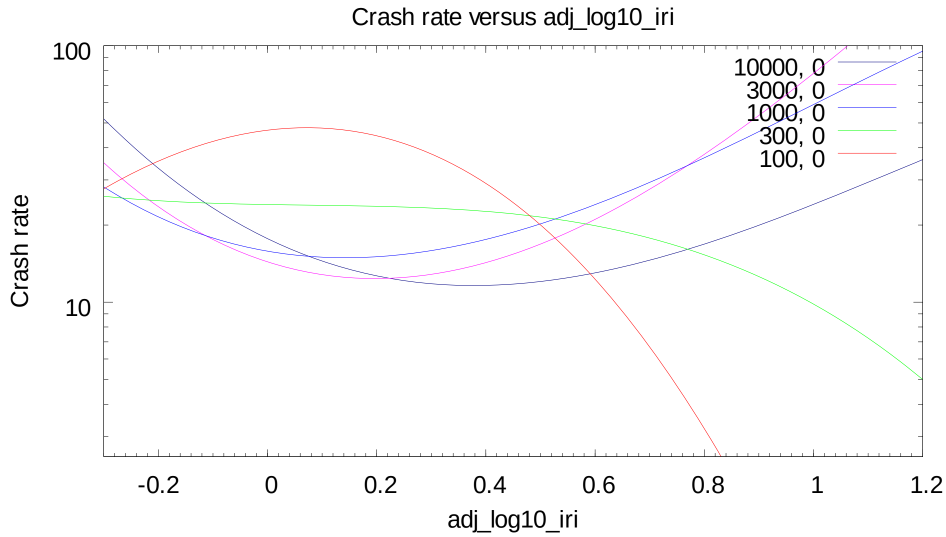
The model with the 10 metre wavelength IRI fits best but is not statistically significantly better than the usual IRI. The 30 metre wavelength IRI fits worst.

Here are the graphs of the crash rate predictions for the various versions of the IRI. They are all quite similar.


Crash rate versus adj_log10_iri_3_variance
Crash rate
adj_log10_iri_3_variance
10000, 0
3000, 0
1000, 0
300, 0
100, 0
100
10
-1
-0.5
0
0.5
1


Crash rate versus adj_log10_iri_10_variance
Crash rate
adj_log10_iri_10_variance
10000, 0
3000, 0
1000, 0
300, 0
100, 0
100
10
-0.5
0
0.5
1
1.5
2
2.5


Crash rate versus adj_log10_iri_30_variance
Crash rate
adj_log10_iri_30_variance
10000, 0
3000, 0
1000, 0
300, 0
100, 0
100
10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5



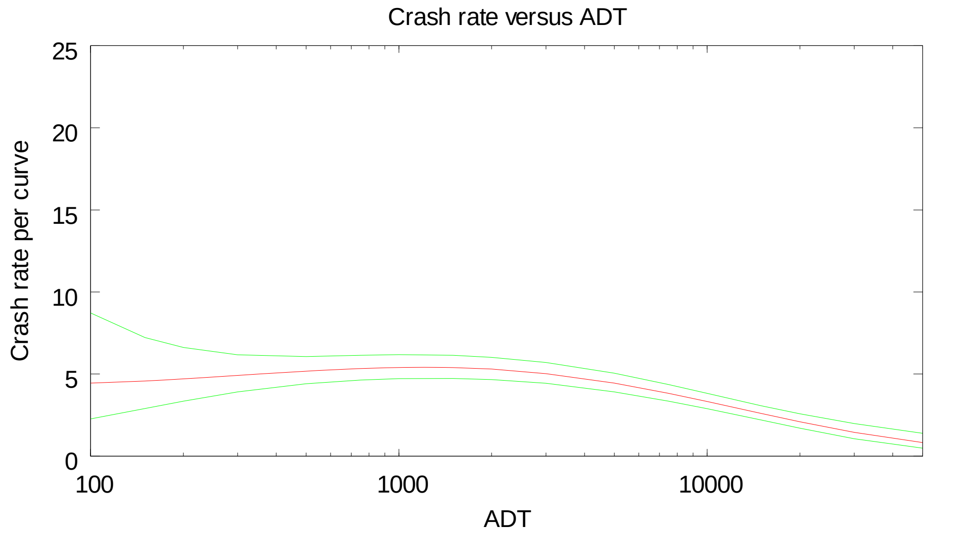
9 Curve model

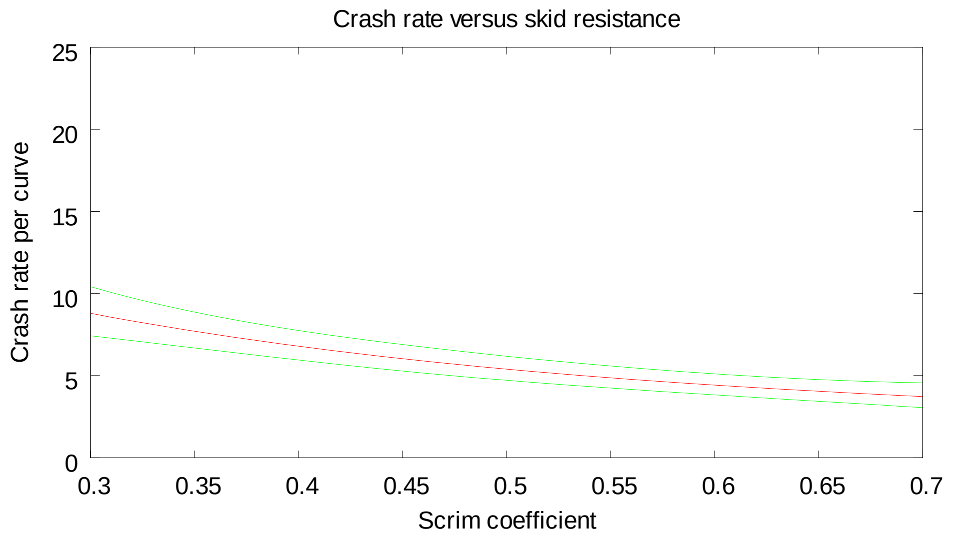
See http://www.robertnz.net/pdf/report_curves3.pdf for details of the curve model. I have run this with the current set of data. The model is the same as before except that I am now including a term for roughness plus its interaction with advisory speed.

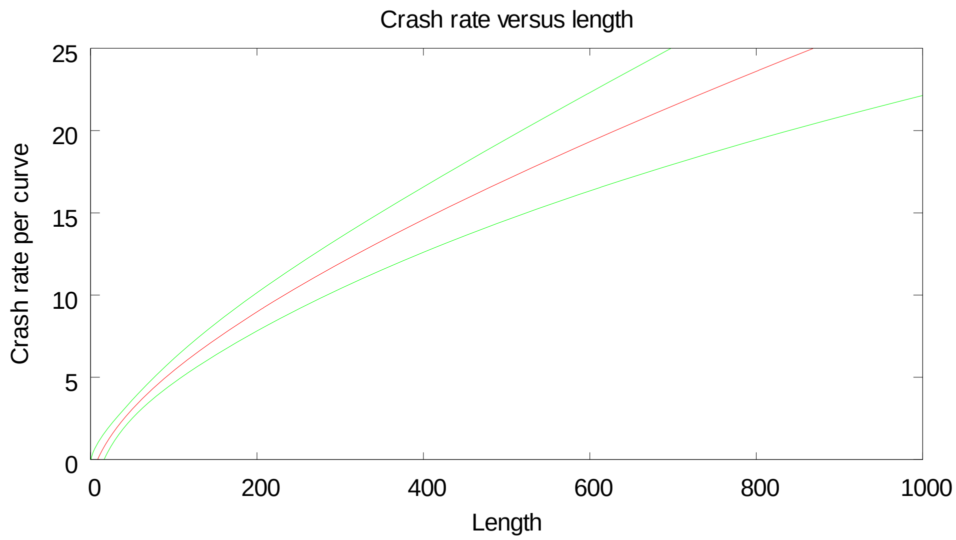
The results are almost the same as in the previous study. Here is the analysis of variance table

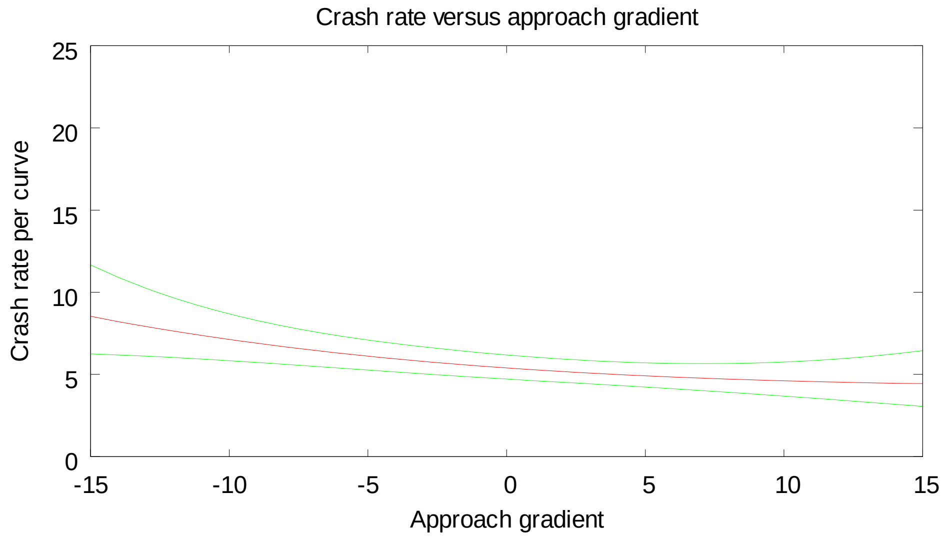
Predictor variable	df	1% pt.	Chi-squared	
			Type III	Type I
year	9	21.7	151.33	150.66
region	13	27.7	140.91	201.76
poly3_OOCC-30.0000	3	11.3	337.62	1127.1
poly3_AS-50.0000	3	11.3	4.6604	9.327
poly2_scrim-0.5000	2	9.21	107.78	75.253
poly3_log10_ADT-3.0000	3	11.3	129.19	122.62
poly2_gradient_app	2	9.21	17.856	19.841
poly3_adj_log10_iri	3	11.3	37.886	32.753
poly2_adj_log10_iri × poly2_AS-50.0000	4	13.3	39.654	39.654
poly2_sqrt_lengthR-15.0000	2	9.21	36.573	36.573

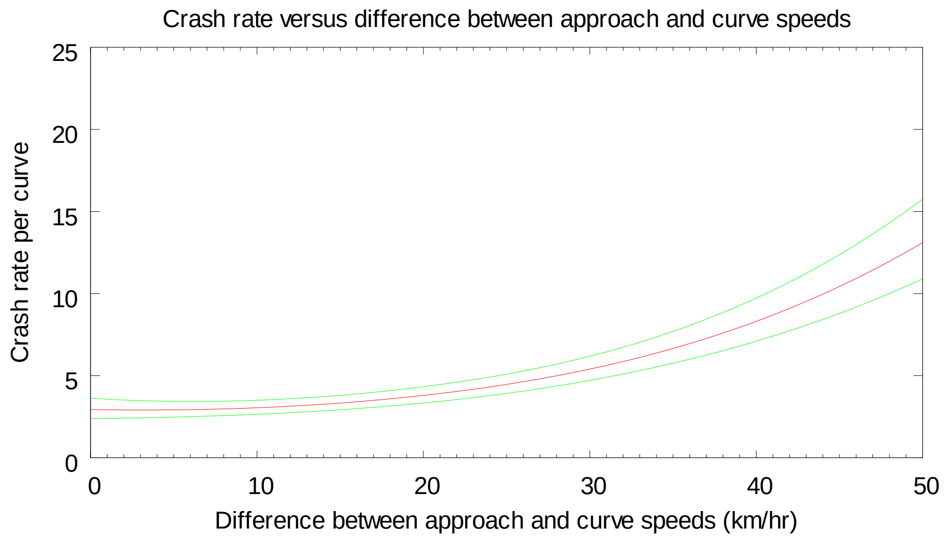
Here are the predictor plots. Note that in line with the previous study, the crash rate scale is linear rather than log.

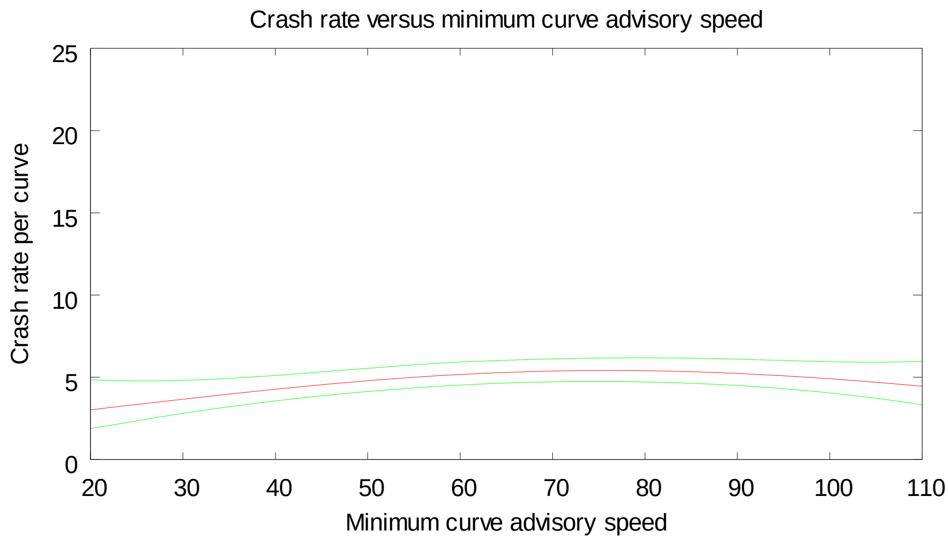

Crash rate versus ADT
Crash rate per curve
25
20
15
10
5
0
100
1000
10000
ADT

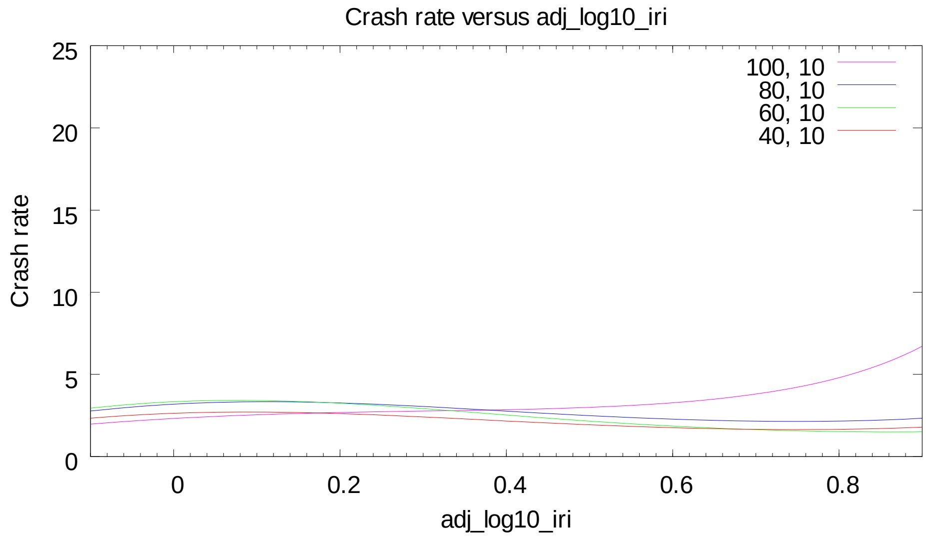

Crash rate versus skid resistance
Crash rate per curve
25
20
15
10
5
0
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
Scrim coefficient


Crash rate versus length
Crash rate per curve
25
20
15
10
5
0
0
200
400
600
800
1000
Length


Crash rate versus approach gradient
Crash rate per curve
25
20
15
10
5
0
-15
-10
-5
0
5
10
15
Approach gradient


Crash rate versus difference between approach and curve speeds
Crash rate per curve
25
20
15
10
5
0
0
10
20
30
40
50
Difference between approach and curve speeds (km/hr)


Crash rate versus minimum curve advisory speed
Crash rate per curve
25
20
15
10
5
0
20
30
40
50
60
70
80
90
100
110
Minimum curve advisory speed



Apart from the graph of adjusted IRI, which is new, the graphs are very similar to those of the previous report.

It is somewhat marginal whether the IRI should have been included. However, its message seems to be that roughness is important only for the curves with high minimum advisory speeds. And that is in line with the rest of this report.

10 Discussion

The most striking result is the close agreement with the analyses of the 1997-2002 data.

The inclusion of the interaction term between roughness and curvature suggests that roughness is a factor for curves where traffic is going at close to full speed but there still is some curvature.

There is a suggestion that skid-resistance is more important on curves than on straight roads. This makes sense. At present I am not able to quantify the dependence very precisely.

The agreement between the analyses when we look at all casualty crashes and those when we consider only serious/fatal crashes suggests that the low reporting rates associated with minor injury crashes is not a serious problem. Similarly there is little change when we include the year × region interaction and this also suggests that reporting rates are sufficiently consistent for the analyses to be valid.

We still find more variability in the data than the Poisson model would predict. Of course, we can’t expect the model to fit exactly as there are numerous things not included and we may be getting the best fit than one can reasonably expect. However it is possible that the problem lies in the estimates of average daily traffic and it may be worth investigating this further. Ideally the analysis should be improved to take account of this variability. One approach is through adding an additional random effects term. Another is using a statistical technique known as *The Jack-Knife* to find improved significance tests and confidence intervals.

The presence of the ADT (average daily traffic volume) term in the regression part of the model is also slightly worrying. It suggests that there are other road characteristics that are not included in the model that are present in low ADT roads. It is easy to suggest a number of these – for example, lower standard of driving, less policing, poorer signage, more hazards on the edge of the roads, and so on. It is also possible that there is a bias in the measurement of ADT on low ADT roads, but I would not have thought it sufficient to cause the effect we see.

In needs to be remembered that this is a *retrospective analysis* as opposed to a *designed experiment*. So it is not possible to be sure that the predictor variables used in the regression analysis are really the ones affecting the crash rates. I have already suggested that this is the case with the ADT effect. Roughness, in particular, could be a surrogate for a number of characteristics associated with roads in need of repair.