Crash-risk model for curves

12 October 2009

1 Introduction

This is an update on my 17 June and 12 August, 2009 reports. My definition of a curve is now more closely aligned with the Opus definition. I have included *approach gradient* in the model and have included *length* in a more satisfactory way.

The aim of the model is to estimate the risk of a crash for each curve on the State Highway network. The input parameters to the model are those found by the Opus curve identification program.

Originally, I had proposed to modify my 2004 crash-risk model which is based on 10 metre segments. This would have involved assigning a risk to each 10 metre segment based on the curve parameters. But this wasn't going to work well since it was going to require a rather arbitrary assignment of the location of the risk over the curve rather than assigning a risk to the curve as a whole.

I decided I needed to modify the program to treat each curve as a unit. Now, the model tries to fit the number of crashes in each curve (and in each year) based on the curve parameters.

One problem with this approach is that it is difficult to allow for the error in the location of the crashes. The risk averaging method used in the 10 metre model is not applicable here. A partial solution has been to assign each crash that occurs within 50 metres of a curve to that curve. Where this would result in a crash being assigned to two curves, it is assigned to the nearest one.

The data used in this study is from 1997 to 2002 as used in the crash-risk model. See <http://www.robertnz.net/pdf/crashrisk8.pdf> for more details. Much of the analysis used in this paper is similar to that used for that crash-risk model. In what follows I refer to this as the *2004 paper*.

As in the 2004 paper, only undivided roads are included in the data. I have used the “all crashes” data is this study. In the August report I also looked at “selected crashes” as defined in the 2004 paper, but this made little difference to the results so I have not repeated that analysis here.

Section 2 of this paper describes the curve identification process; section 3 shows histograms of the curve data; section 4 gives the description and numerical results of the statistical analysis; section 5 shows the results in

graphical form; section 6 describes some possible further work and section 7 contains an appendix.

2 The curve identification process

Here is the process for identifying the curves.

Each year is regarded as a separate set of data in the sense that there is no attempt to align or identify curves found in one year with those found in another.

Each state highway is regarded as a continuous block of data. So running means can be carried out on this block of data but will terminate at each end of a state highway.

These are the steps:

2.1 Advisory speed

Calculate advisory speed. I use the version that includes crossfall if it is in the correct direction and excludes gradient. I have an upper bound of 110 km/hr. See section 7.1.

2.2 Running mean of advisory speed

Calculate the running mean of three adjacent measurements of the advisory speed. If less than three measurements are available, for example, because we are at the end of a state highway or because there are missing observations, take the mean of those that are available. If no measurements are available, assume 110km/hr. This default is used so that occasional missing values are unlikely to affect the statistics for a curve. We should reject curves with a lot of missing values, but usually this will happen automatically since no curve will be identified.

2.3 Running mean of absolute value of curvature

Calculate the running mean of the *absolute* values of three adjacent measurements of the curvature for each side of the road. Note that curvature is measured as a "radius of curvature" so sharp bends have small values. A straight road is assigned a value of 100,000. Assign a sign to the curvature as that of the running *harmonic mean* of the observations. Again, if less than 3 measurements are available, use those that are available. If none are available assume 100,000 as in a straight road.

2.4 Harmonic mean and sign of the curvature

The *harmonic mean* is the reciprocal of the mean of the reciprocals. Using the harmonic mean of the curvature rather than the ordinary mean avoids the possibility of an awkward cancellation when we have an alternating curvature in a near straight road – see the next paragraph. Using the harmonic mean also reduces the smoothing out of a very sharp bend that affects only a single measurement.

If we have a nearly straight road the radius of curvature will fluctuate around 100,000 (100,000 is used for an exactly straight road instead of infinity) but with the sign being positive or negative depending on the direction of the curve. So where a road is almost straight but the actual curvature changes sign we can get a small value of the running mean through cancellation, and so, apparently, a sharp bend, even though the road is almost straight. The harmonic mean doesn't have this problem. However, Opus prefers to use the arithmetic mean. So I have used the arithmetic mean of the absolute values but used the harmonic mean to set the sign of the curve. This avoids the cancellation problem and will usually give the same answer as the ordinary arithmetic mean.

2.5 Initial identification of curves

Make an initial identification of curves by supposing that a 10 metre segment belongs to a curve if the 30 metre running mean of the absolute curvature for that segment is less than **800** on at least one side of the road.

The direction of curvature of each segment is determined by the side of the road with the minimum value of the running mean of the absolute curvature. Change the sign when the direction is determined by the right hand side.

Treat adjacent segments that have been identified as belonging to a curve as in the same curve if they have the same direction of curvature.

2.6 Deletions

Delete any curves that are less than **30** metres long (i.e. only 1 or 2 ten metre segments). I think this is slightly less restrictive than the Opus criterion.

Delete any curves that do not have any points with running mean of absolute curvature less than **500** metres. That is, curves must have at least one segment with running mean of absolute curvature less than **500** metres.

2.7 Connect curves

Connect two curves if they have the same curvature, have no identified curve between them and have a gap between them of **20** metres or less. I do not check the direction of curvature in this gap.

2.8 Delete long curves

Delete any curves longer than **1000** metres.

2.9 Curve apex and associated curve parameters

Find the apex of the curves. This is the point on the curve at which the running mean of the advisory speed is a minimum. The advisory speed of the curve is the value of the running mean at this point. *I calculate separate values for the left and right hand side*.

Also calculate the 30 metre running mean of the skid resistance at the apex. Where there are no scrim values available in the 30 metres, 0.5 is used as the default. This is the skid resistance of the curve. Again there are separate values for the left and right hand sides.

Choose the apex with the lowest advisory speed and find the estimated average daily traffic, ADT, at this point. This is the ADT of the curve. Also identify the region at this point. *P1* is combined into *R2*. We have seven regions altogether.

2.10 Approach speeds

Calculate the *approach speeds*. This is the average of the advisory speed over the **500** metres before the curve on the left hand side and after the curve on the right hand side (i.e. before the curve as a driver would see it). In this case, I truncate the advisory speed at 70km/kr for segments in an urban area. If there are at least 40 observations (400 metres) I use the average. If there are less than 40 observations I reject the curve.

2.11 Approach gradients

Calculate the *approach gradients*. This is the average gradient over **100** (10 observations) metres preceding (as a driver would see it) the curve. Delete the curve if we don't have at least 8 non-missing values.

2.12 Out of context curve effect

The OOCC (out-of-context-curve) effect is the approach speed less the curve advisory speed. Replace by zero if negative.

2.13 Catchments and crash counts

Calculate the *catchments* for each curve. This includes the curve plus **50** metres at each end. Where catchments overlap reduce them equally so that they just don't overlap.

Count the number of crashes in the catchments for each curve.

Delete all curves for which the catchment includes an urban segment, a skid site 1 section or an intersection or more than 2 lanes.

This completes the identification of the curves used in the analysis.

3 The curve data

This section shows histograms of the data. The data is divided into 100 bins for most of the plots. But when this leads to very jagged graphs because of the discrete nature of the data, a lower number of bins is used.

3.1 The number of curves and crashes by region and year

The following tables show the number of curves and the number of crashes identified each year and in each region.

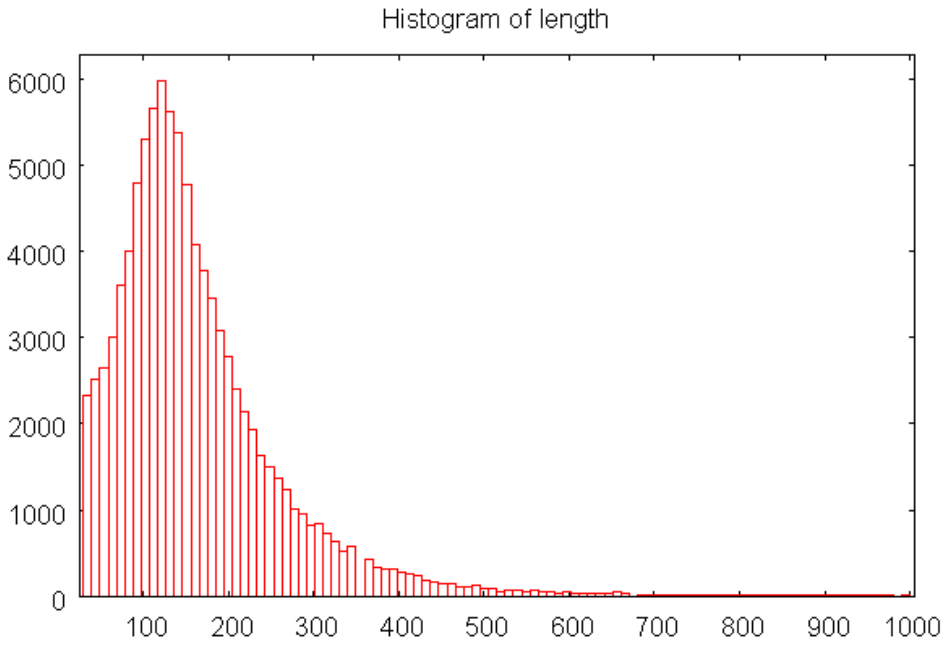
Year	Number of curves	Number of crashes
1997	15352	508
1998	15762	522
1999	16031	505
2000	16416	506
2001	15841	589
2002	16033	614

Region	Number of curves	Number of crashes
R1	9292	345
R2	23457	986
R3	9957	317
R4	10293	429
R5	8620	339
R6	19790	406
R7	14026	422

As noted in section 2, region *P1* is combined into *R2*.

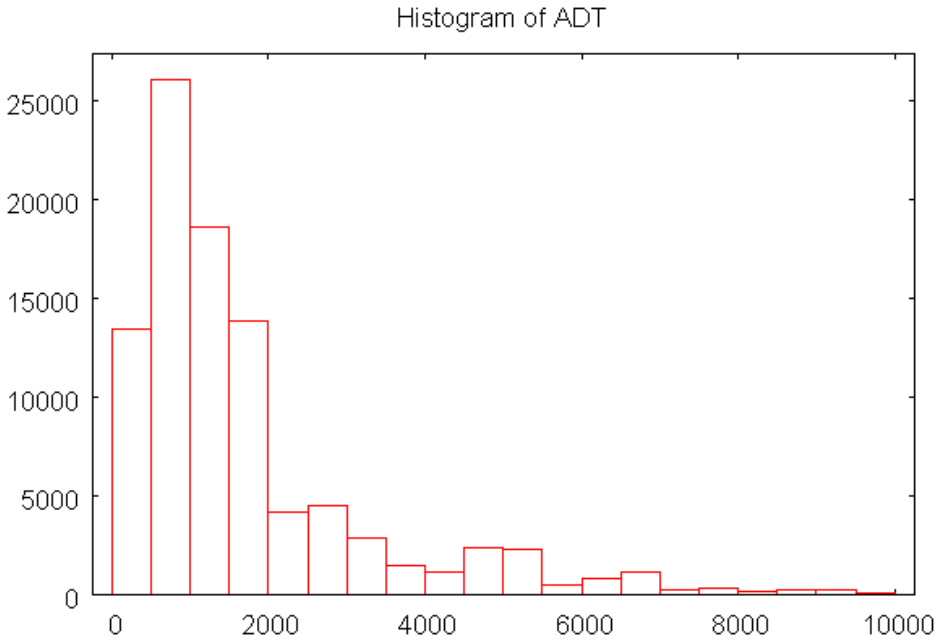
3.2 Length

Here is the histogram of length. It shows that while most of the curves are less than 500 metres in length, they can stretch out to 1000 metres. Any curves longer than 1000 metres have been deleted.



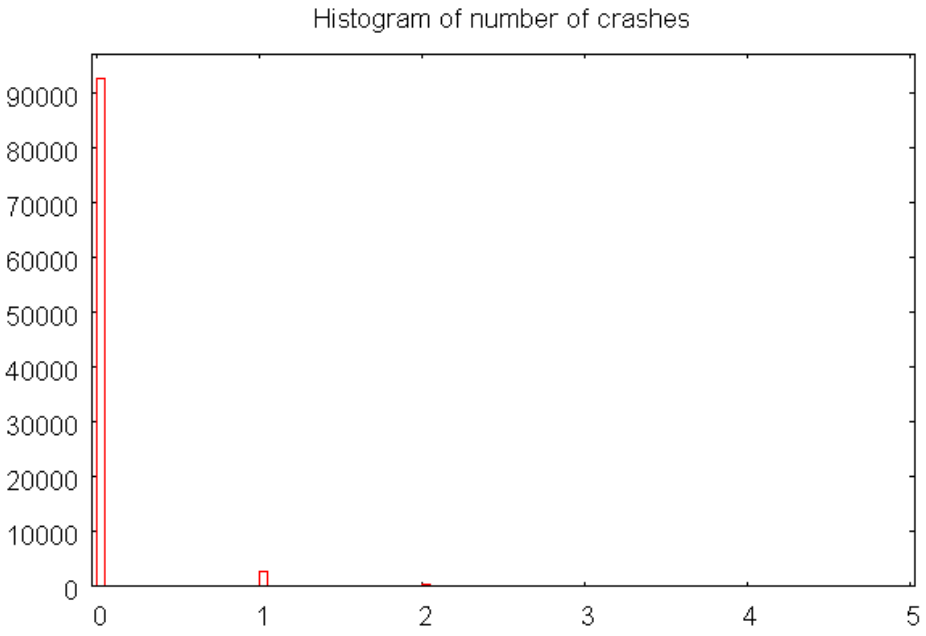
3.3 ADT

The histogram is of the ADT. The graph has been truncated at 10,000. The most common values are in the range 0 to around 2,000. However the roads in the upper end of the graph are important because of their higher ADT.

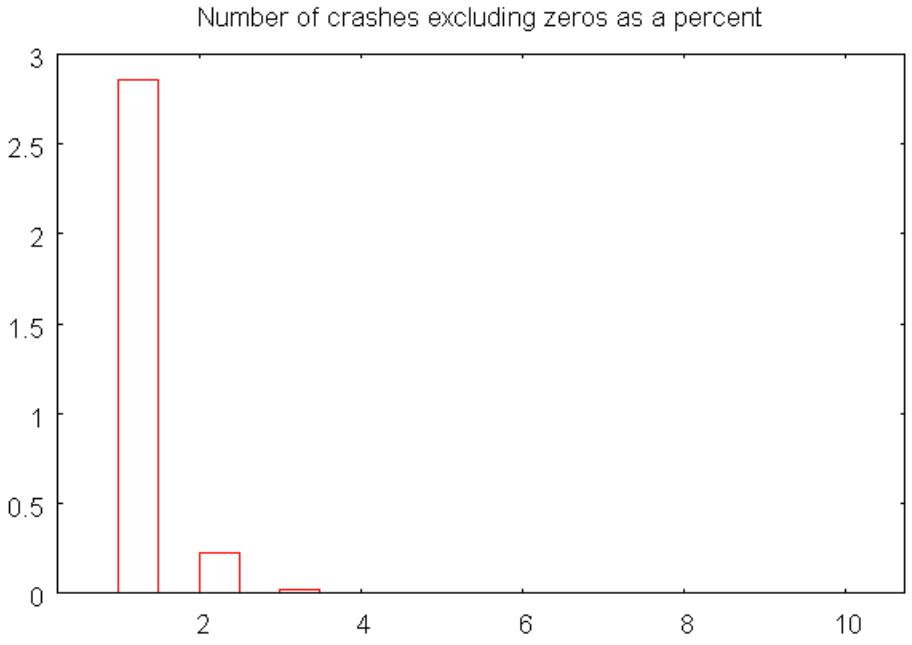


3.4 Crashes

Most of the curves have no crashes with a maximum of around 5. (Each curve generates a line of data for each year so, for example, each year a curve has no crashes contributes to the zero column).

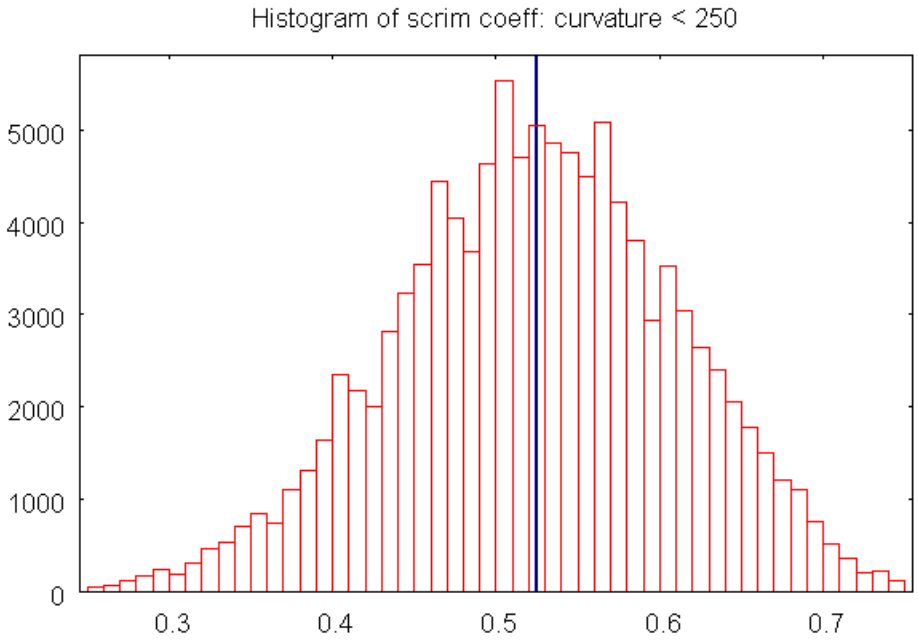


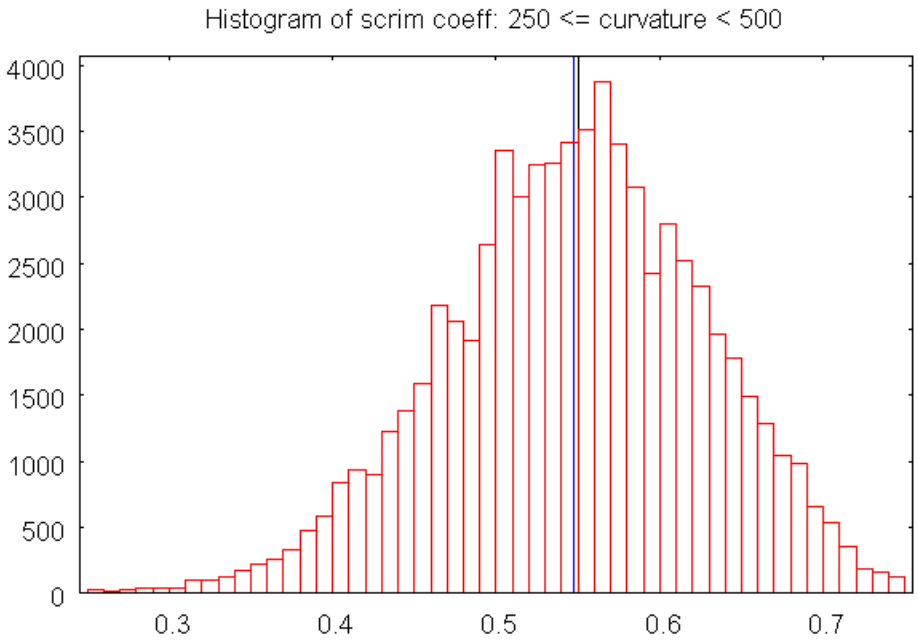
The following graph shows the same data with the results expressed as a percentage and column corresponding to zero crashes omitted.



3.5 Scrim coefficient

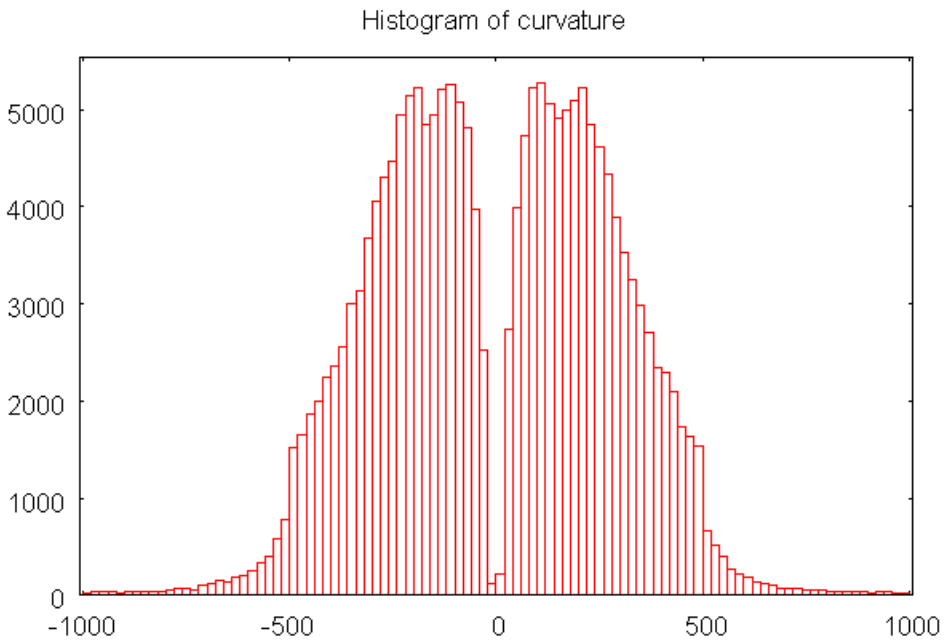
The graphs show the histograms for the scrim coefficient at the curve apex curves with curvature less than 250 and curves with curvature between 250 and 500. The jagged nature of the graph may be due to an interaction between the rounding of the data and the sampling into bins. However, there is an extra peak at 0.5 due to this being used as a default value. The vertical black line shows the median and the blue line shows the mean. The histograms are truncated at 0.25 below and 0.75 above.





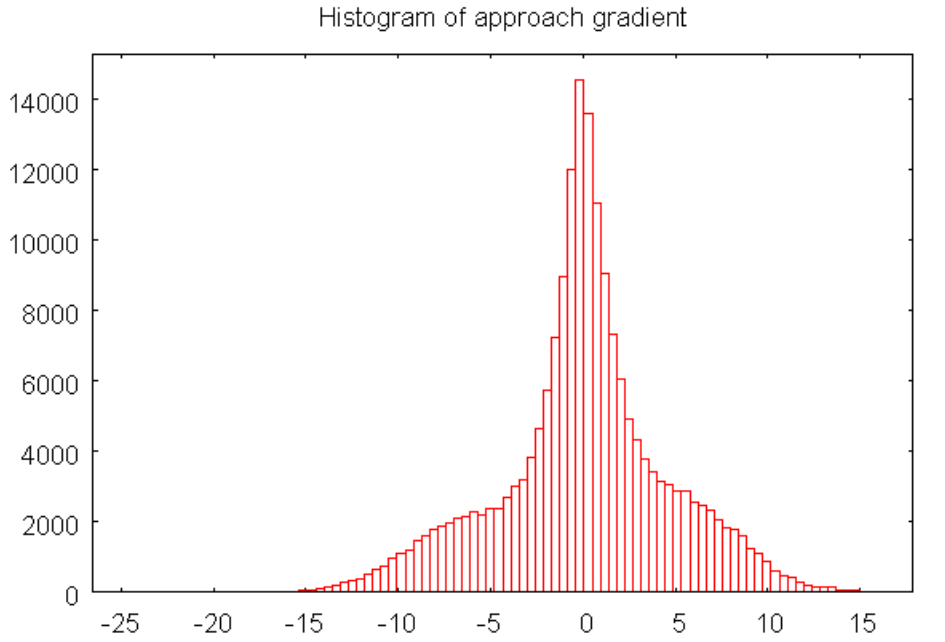
3.6 Curvature

These graphs are of the 30 metre running harmonic mean of the absolute curvature calculated at the apex with signs assigned as in section 2. They have been truncated at plus and minus 1000.



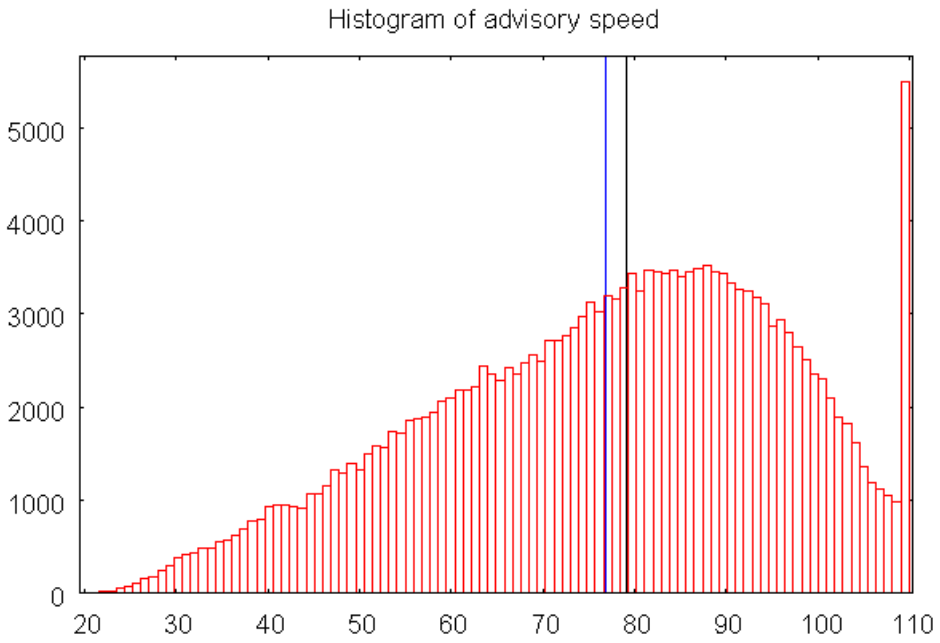
3.7 Approach Gradient

This is the average of the gradient over 100 metres before the curve.



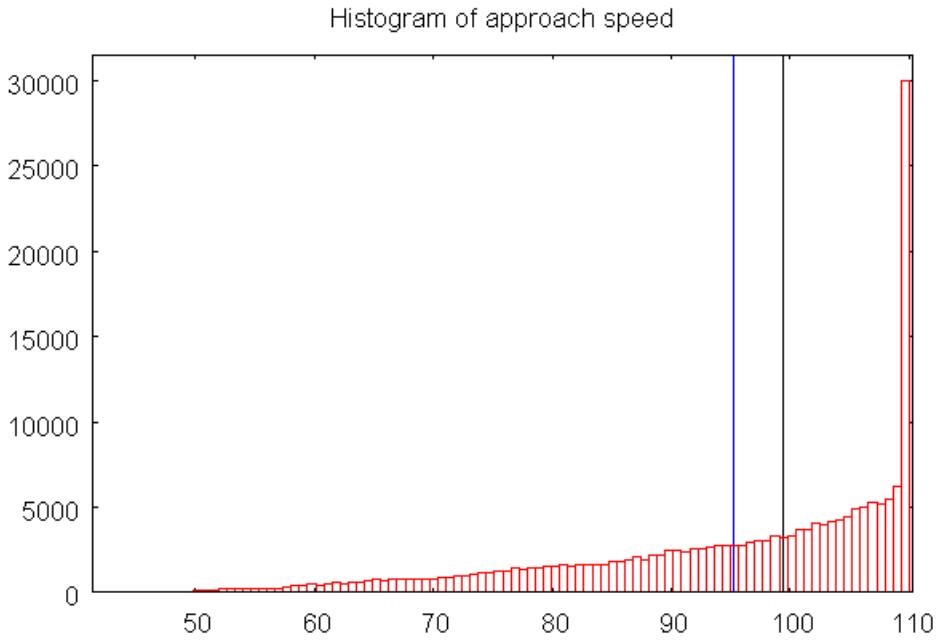
3.8 Curve (advisory) speed

This is the running mean of the advisory speed at the apex. There is a wide range of values. The most common values are the peak at the 110 km/hr and the range 80-90 km/hr. The black and blue lines show the median and mean.



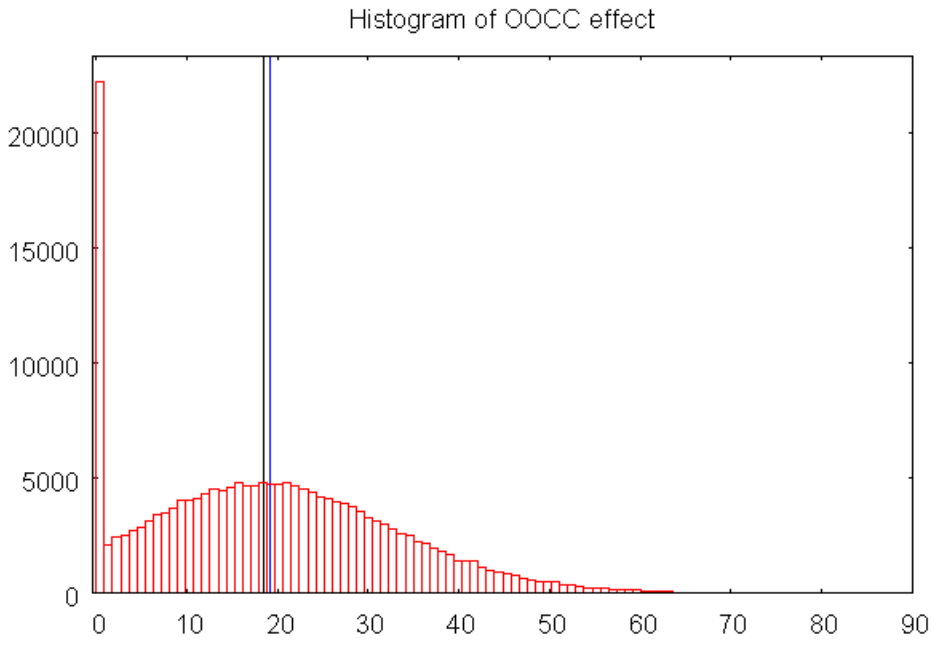
3.9 Approach speed

This is the average of the advisory speed over 500 metres before the curve. There is a peak at 110 km/hr and most values are greater than 70 km/hr. The black and blue lines show the median and mean.



3.10 Out of context curve (OOCC) effect

This is the approach speed less the curve speed, truncated below at zero. There is a peak at zero, otherwise most values are less than 50 km/hr. The black and blue lines show the median and mean.



3.11 Approach speed and OOCC

The ranges of each variable were divided into 25 bins and a two dimensional histogram calculated. The results were fed to a contour plot program. There were no points left of the red line – OOCC must be less than the approach speed; no more than 10 points per histogram column left of the green line; no more than 100 points per column left of the blue line; and no more than 1000 points per column left of the violet line.

4 The statistical analysis

I use a modification of a Poisson linear/log-linear model. I suppose each side of each curve of road can *generate* crashes at the rate (per year)

$$aL_1 \exp(L_2) \qquad (1)$$

where a is the ADT (per side) and L_1 and L_2 are linear combinations of the road characteristics including (polynomial functions of)

- a constant
- square root of curve length

for L_1 and

- OOCC
- curve speed
- skid resistance
- approach gradient
- $\log_{10}$(ADT)
- year
- region

for L_2.

The model is a combination of the linear model (the L_1 part) and a log-linear model (the L_2 part). Having the curve length in the linear part gives a more satisfactory looking fitted curve – see below. However, the model involving both linear and log components seems little more problematic than a purely log version.

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

The values calculated from the model for each side of the road need to be added together to get the value for the curve.

To find the number of crashes per 100 million vehicles passing through the curve one needs to calculate

$$\frac{10^8}{365} L_1 \exp(L_2) \qquad (2)$$

for the side of the road one is interested in. Average the values for the two sides of the road to get an overall curve effect.

Here is the analysis of variance table for the model fit. I am taking third degree polynomials of the (OOCC − 30), (advisory speed − 50), and {$\log_{10}$(ADT) − 3}; second degree polynomial of (scrim − 0.5), approach gradient and (square root of length – 15) . To help the fitting process I have subtracted a middle value from most of the variables. For the length term take the square root of the length, subtract 15 and then apply the polynomial transform.

term	df	SS(3)	SS(1)
year	5	35.889	28.447
region	6	61.351	89.667
poly3_OOCC-30.0000	3	189.66	457.41
poly3_AS-50.0000	3	28.029	14.163
poly2_scrim-0.5000	2	63.426	47.629
poly3_log10_ADT-3.0000	3	35.484	35.991
poly2_gradient_app	2	13.806	13.806
poly2_sqrt_lengthR-15.0000	2	81.969	81.969

The column SS(3) shows the chi-square value when the corresponding term is the last one added to the analysis and SS(1) shows the chi-square value

when the terms are included sequentially. See the 2004 paper from more details of these quantities. When calculating the SS(1) values, for each set of terms (i.e. the L_1 term or the L_2 terms) I assume that the other set has already been fitted. I think the SS values for the length underestimate the significance of that term.

The significance points are shown in the following table

df	1	2	3	4	5	6
5% point	3.84	5.99	7.81	9.49	11.07	12.59
1% point	6.63	9.21	11.34	13.28	15.09	16.81

Comparing the SS(3) with the figures in this table, all are statistically significant at the 1% level, if the Poisson model was valid. In the 2004 paper it appeared that the significance points should be raised by a factor of 4. If this was the case here, most of the effects, would be significant at the 5% level and some would be significant at the 1% level.

The OOCC effect is less significant than in the July version of the report and this is partly due to extending the curves to include the 800 metre curvature parts rather than just 500 metre – see section 2.5.

In my initial testing of the model I also tried the average gradient between the beginning of the curve and curve apex. This was less statistically significant than the approach gradient when only one of average gradient and approach gradient was included. It was not statistically significant when approach gradient was included first and so was not included in the final model.

Here is the table of effects

effect	estimate	s.d.	ratio
year:1997	0		
year:1998	-0.023517	0.062334	-0.37728
year:1999	0.043604	0.063351	0.68829
year:2000	0.020113	0.063313	0.31767
year:2001	0.19874	0.061102	3.2526
year:2002	0.25136	0.061086	4.1148
region:R1	0		
region:R2	0.13161	0.063507	2.0724
region:R3	0.38803	0.080129	4.8425
region:R4	0.40065	0.073579	5.4452
region:R5	0.28962	0.078518	3.6886
region:R6	0.33949	0.078584	4.3201
region:R7	0.43579	0.076388	5.7049
OOCC-30.0000**1	0.043873	0.0041339	10.613
OOCC-30.0000**2	0.00039063	0.00010672	3.6605
OOCC-30.0000**3	-1.241e-05	4.9702e-06	-2.4969
AS-50.0000**1	0.015698	0.0034107	4.6026
AS-50.0000**2	-9.4268e-05	0.00017091	-0.55157
AS-50.0000**3	-9.8667e-07	2.6517e-06	-0.37208
scrim-0.5000**1	-2.1705	0.27257	-7.9631
scrim-0.5000**2	-1.1439	2.1587	-0.5299
log10_ADT-3.0000**1	-0.059041	0.093727	-0.62993
log10_ADT-3.0000**2	-0.17294	0.20598	-0.83962
log10_ADT-3.0000**3	-0.08039	0.15472	-0.5196
gradient_app**1	-0.02628	0.0077375	-3.3964
gradient_app**2	0.00034872	0.00079037	0.44121

```
                constant     1.7707e-05     1.7967e-06         9.8552
sqrt_lengthR-15.0000**1      1.6081e-06     1.9194e-07         8.3785
sqrt_lengthR-15.0000**2      6.8419e-09     1.2095e-08         0.5657
```

I could have reduced the degree of the polynomial in a number of cases, but it seems to make sense to allow a little more flexibility in the curves than the model strictly requires. Values of the ratio greater than 2 are statistically significant if we believe the Poisson model and this should be raised to 4 if we follow the 2004 study. These tests should be applied only to the highest degree term in each polynomial.

The column labelled estimate is the one to use in a spreadsheet for estimating risk of a curve.

5 Graphs of the effects

The following graphs show the effect of the different variables varied one at a time, or in the case of OOCC and AS, two at a time.

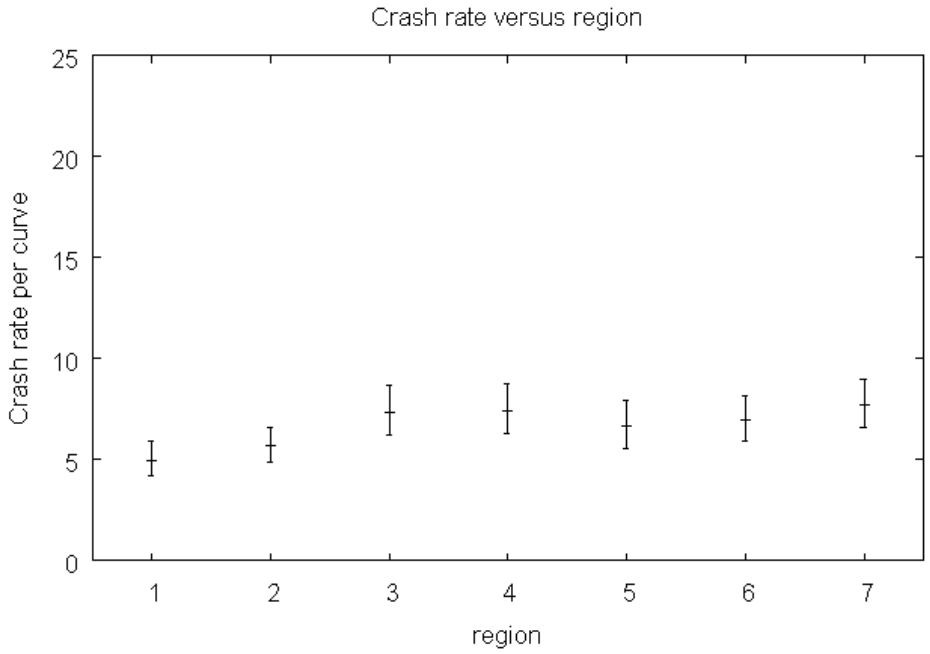
The variables not being varied have the following values:

```
Year:      2002
Region:      R2
OOCC:        30
AS:          80
scrim:      0.5
ADT:       1000
gradient:     0
length:     100
```

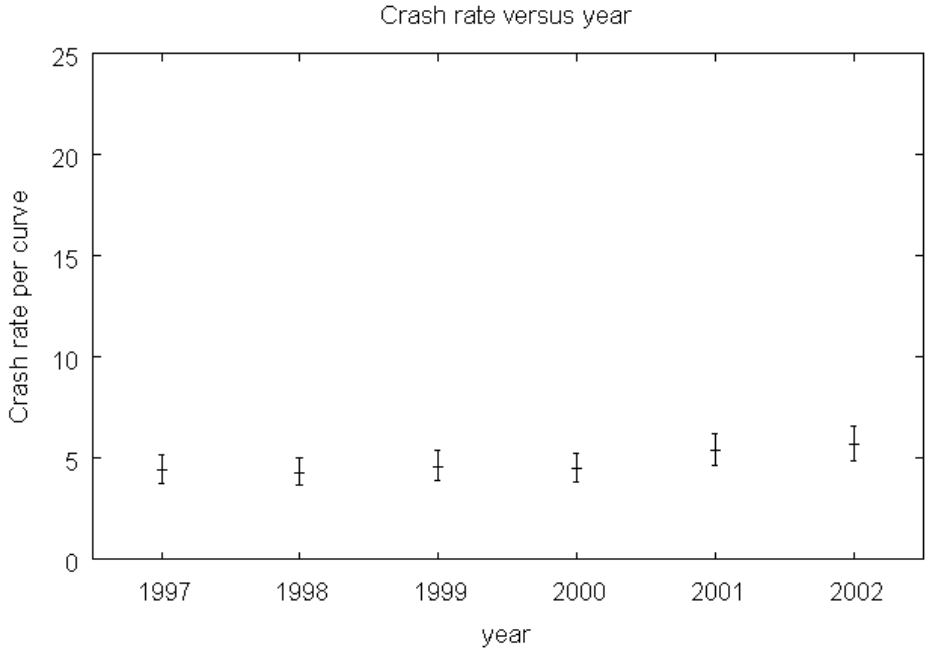
The quantity being modelled is crashes per 100 million vehicles. That is, the graphs show the predicted crashes per 100 million vehicles with the values in the table above except for the one or two being varied.

The red line in the line graphs showing curves is the estimate and the green lines are 5% confidence intervals. There is no adjustment for a possible increase in the error.

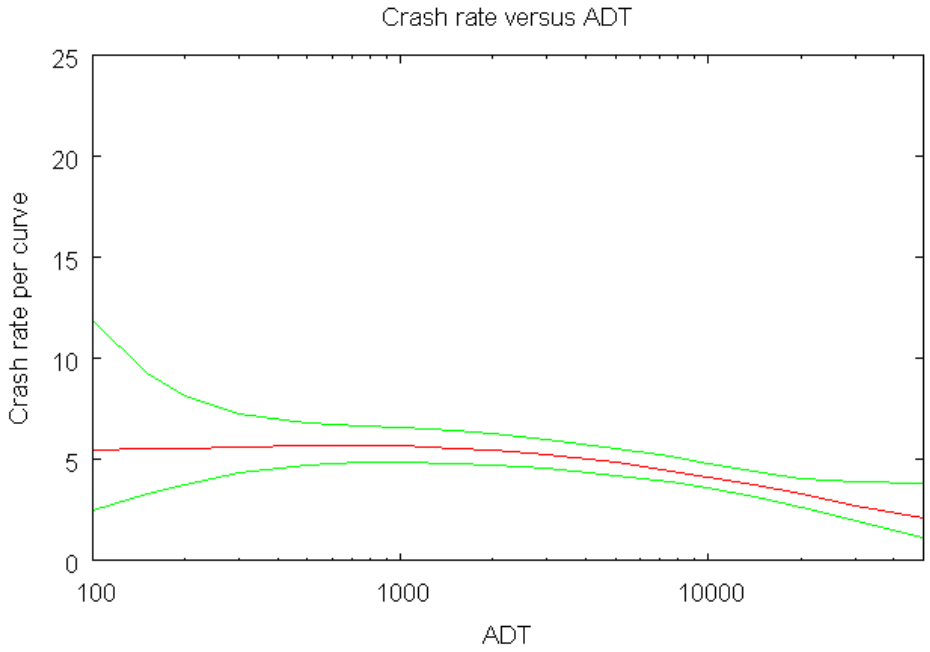
5.1 Crash rate versus region



5.2 Crash rate versus year

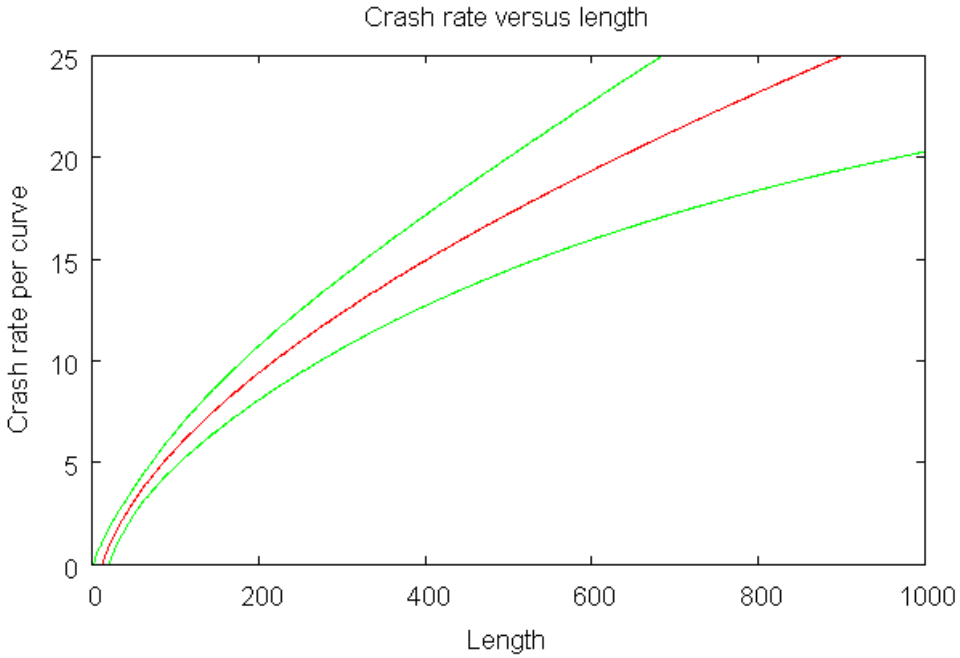


5.3 Crash rate versus ADT

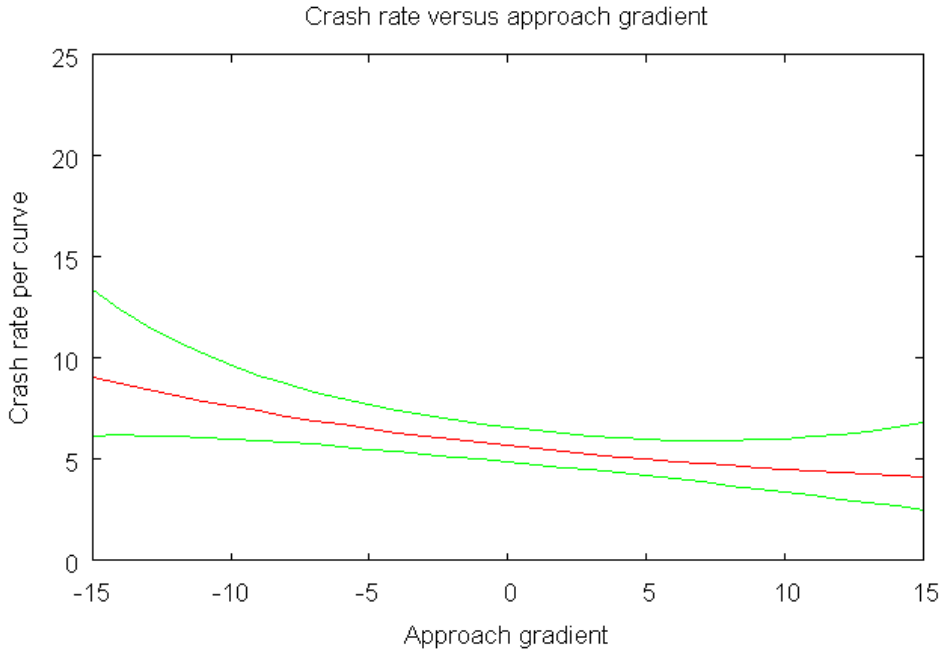


5.4 Crash rate versus skid resistance

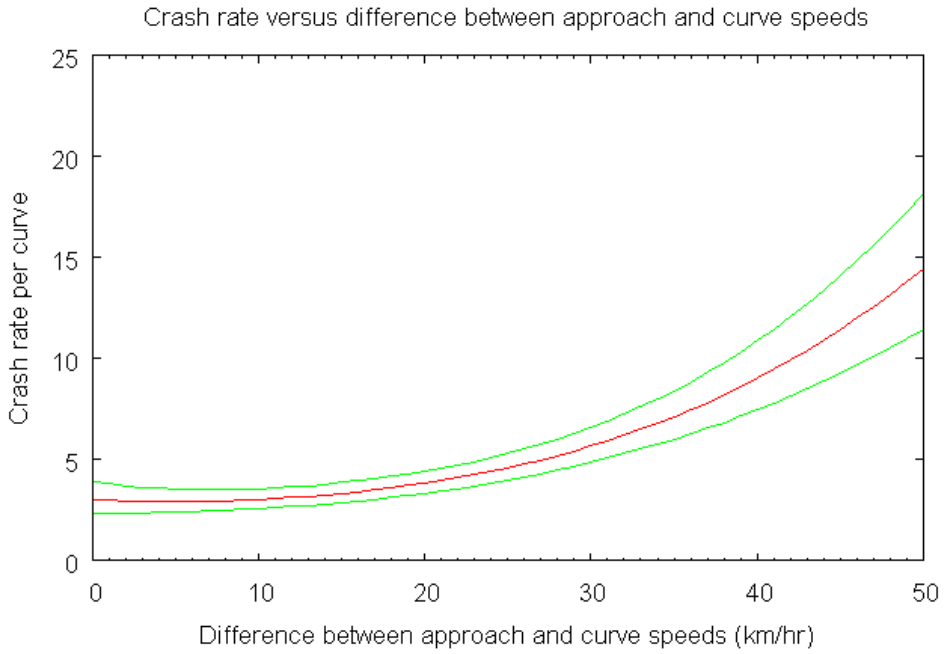
5.5 Crash rate versus length



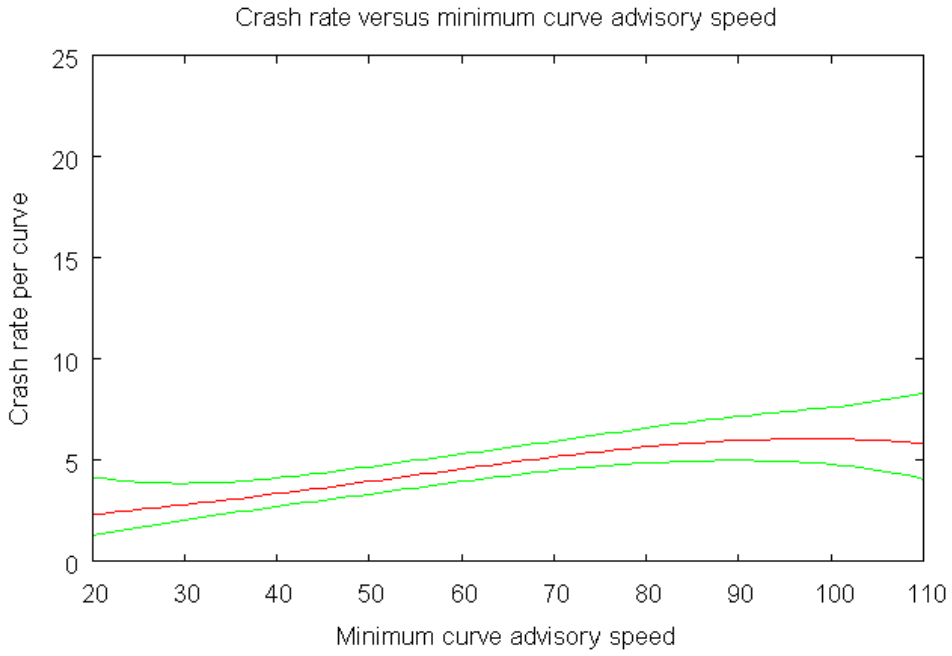
5.6 Crash rate versus approach gradient



5.7 Crash rate versus OOCC



5.8 Crash rate versus advisory speed



At a first glance, this graph doesn't make sense. Crash-rate seems to be going up as the road becomes less curved (AS increases). The reason is that OOCC is being held constant so the approach advisory speed is also increasing.

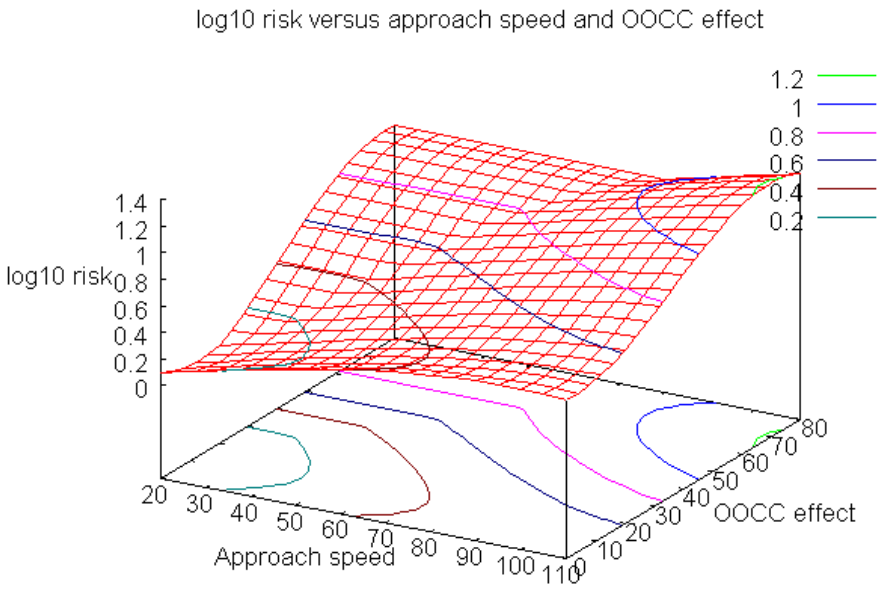
5.9 Crash rate versus approach speed and OOCC

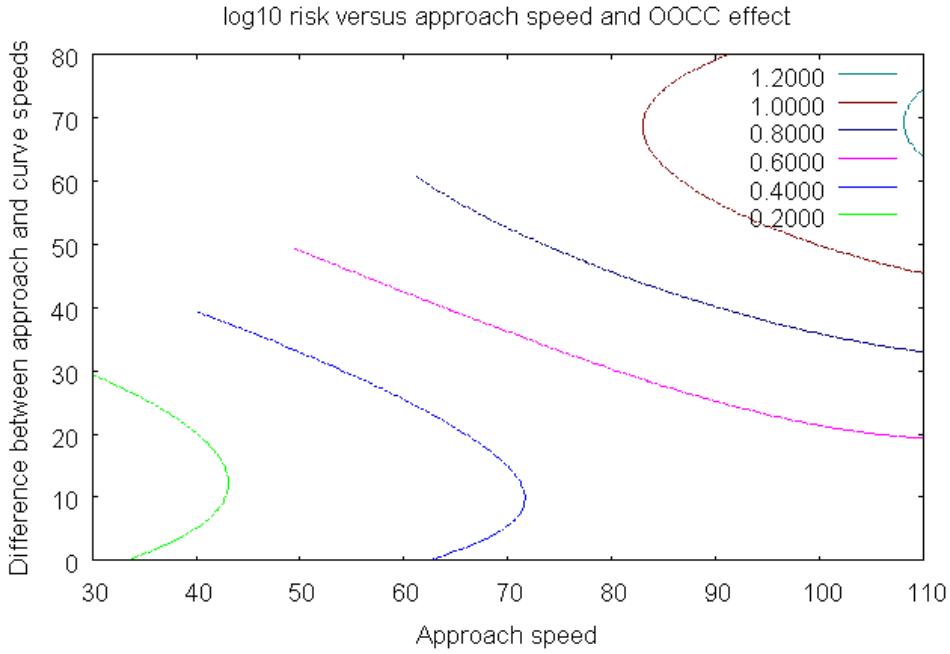
The graphs in this section look at approach speed and OOCC together. It seems a little easier to understand if we use these two variables rather than curve advisory speed and OOCC.

I provide two versions of the graphs. The first is as a three dimensional plot to show the overall appearance and a contour plot to show the results more accurately. The three dimensional plot also shows a few contours and note that they use different colour coding from the contour plot.

The values at the rear of the three dimensional plot don't have any meaning because OOCC is greater than the approach advisory speed. These points are omitted from the contour plot.

See the figure in section 3.11 for the numbers of curves contribution to the different parts of these graphs.





The general appearance is an increasing risk as either approach advisory speed or OOCC increase. But for lower OOCC and the approach speed, the approach speed is more important. For higher OOCC both are important, but OOCC tends to play a larger role.

6 Possible further work

Investigate some kind of goodness of fit measure.

Compare results with the 2004 paper.

7 Appendix

7.1 Advisory Speed

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 = % Crossfall (sign relative to curvature)

H = Absolute Curvature (radians/km) = (1000m / R)

R = Absolute radius of curvature

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 on rural roads and 70 km/hr on urban roads.