Freeway rear-end collision risk for Italian freeways.

An extreme value theory approach

Gregorio Gecchele  
Federico Orsini  
University of Padova
Outline

Freeway rear-end collision risk
- Main concepts
- Focus of the paper

The study
- Proposed Approach
- Experimental Analysis
- Analysis of Results

Theory and Practice

Conclusions
- Concluding Remarks
- Future Developments
Freeway rear-end collision risk

- Main concepts
- Focus of the paper
Motivations

• Safety analysis is a relevant activity in transportation planning
• Road safety analysis traditionally is based on crash data.
• However, the use of crash data has several shortcomings:
  – under-reporting and frequent problems due to poor data quality
  – often used in an aggregated manner, limiting the opportunity of describing in detail heterogeneous crash causes and driver crash avoidance behavior
  – crash data must be collected for long periods of time (about 3 years) to accumulate sufficient data for analysis, since crashes are comparatively rare events
  – the use of such data is a reactive approach, since a significant number of collisions must occur before action can be taken

Objectives

• Testing Extreme Value Theory (EVT) approach based on surrogate safety measures instead of crash data for freeway rear-end collisions
• First step of a more comprehensive research activity
The study

Theory and Practice

- Proposed Approach
- Experimental Analysis
- Analysis of Results
Basic concepts

• Use of surrogate safety measures instead of road crash data
• For transportation safety applications surrogate measures should:
  – be based on observable non-crash events, physically related predictably and reliably to crashes
  – provide a practical method for converting the non-crash events into corresponding crash frequencies and/or severity data (Tarko et al., 2009).

A surrogate event must be affected by the engineering treatment if the treatment indeed affects safety.
Basic concepts

Adopting Extreme Value Theory (EVT) approach the crash frequency could be predicted given the estimated relationships between safety and road and other conditions $X$:

$$F(C|X) = F(H|X) \cdot P(C|H, X)$$

$F(C|X) =$ frequency of crashes under a set of conditions $X$

$F(H|X) =$ count model of surrogate event $H$ frequency

$P(C|H, X)$ extreme value model of risk of crash given surrogate event $H$ and conditions $X$

• The EVT abandons the assumption of a fixed coefficient converting the surrogate event frequency into the crash frequency

• The risk of crash given the surrogate event is estimated for any conditions based on the observed variability of crash proximity without using crash data

• The crash proximity measure precisely defines the surrogate event.
Block Maxima with Generalized Extreme Value distribution

Extreme observations are aggregated into time (or space) intervals, called blocks. In each block the event with the highest value, i.e., the block maximum, is considered as an extreme event and is sampled.

Generalized Extreme Value (GEV) function:

\[ G(x) = \exp \left\{ - \left[ 1 + \frac{x - \mu}{\sigma} \right]^{-\frac{1}{\xi}} \right\} \]

- \(-\infty < \mu < \infty\) is the location parameter
- \(\sigma > 0\) is the scale parameter
- \(-\infty < \xi < \infty\) is the shape parameter:
  - \(\xi > 0\) type II = Fréchet distribution
  - \(\xi = 0\) type I = Gumbel distribution
  - \(\xi < 0\) type III = Weibull distribution
Block Maxima with Generalized Extreme Value distribution

Block interval size is a compromise, to avoid poor fit:

• small interval may result in sampling events which are not really extremes
• large interval may result in an overall sample size with not enough extremes

For a reliable fit, the sample should contain at least 30 maxima

In cases in which blocks contain enough observations, $r$-largest order statistic is usually recommended, as it allows to sample more extreme events to fit the GEV.

In EVT applied to road safety estimation it is often required to study the minima within the blocks. The BM approach can also be used to study minima, by considering the maxima of the negated values.
The Study.
Proposed approach

Modeling details

The surrogate safety measure considered is time-to-collision (TTC), calculated by dividing the distance between consecutive vehicles by the difference in their speed. In this case study, events are extreme when TTC is low; therefore, the model is fitted using the variable “negated-TTC” (-TTC).

\[
TTC = \frac{R}{RR} = \frac{S}{(V2 - V1)}
\]

The cumulative probability of \(-TTC \geq 0\) represents the probability of an accident to occur in any given time block. It is then possible to calculate the expected number of accidents in a larger period of interest (e.g., one year), by multiplying this probability to the number of blocks in the period of interest.

\[
ACC_T = N_T \Pr\{-TTC \geq 0\} = N_t [1 - G(0)]
\]

- \(T\) period of interest (e.g., one year, five years);
- \(ACC_T\) number of estimated accidents in the period of interest \(T\);
- \(N_T\) number of time blocks contained in the period of interest \(T\);
- \(TTC\) time-to-collision;
- \(G\) extreme value distribution function.
Main steps

1. Traffic data and incident data treatment
   - Time-To-Collision (TTC) calculation
   - Rear-end observed accidents selection

2. Calibration of EVT models with Matlab and R package “extRemes”
   - BM based on hourly values
     - Full dataset calibration (BM-hour full)
     - Week datasets calibration (BM-hour week)
   - BM based on r-largest daily values (BM-day)
     - Full dataset calibration (BM-day full) – STATIONARY and NON STATIONARY
     - Week datasets calibration (BM-day week)

3. Analysis of results
   - Analysis of model fitting
   - Comparison with rear-end observed accidents
Northern Italy freeway case study

Straight road segment:
- Length: 5 km
- Road section: 3-lanes
- Lane width: 3.75 m

Detailed traffic data (Vehicle by Vehicle)
- For each lane vehicle length and speed, time headway

<table>
<thead>
<tr>
<th>Week</th>
<th>Period</th>
<th>ADT weekdays</th>
<th>ADT weekends</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14-20/01/2013</td>
<td>35140</td>
<td>23160</td>
</tr>
<tr>
<td>2</td>
<td>11-17/03/2013</td>
<td>38190</td>
<td>25550</td>
</tr>
<tr>
<td>3</td>
<td>15-21/07/2013</td>
<td>46890</td>
<td>41110</td>
</tr>
<tr>
<td>4</td>
<td>14-20/10/2013</td>
<td>40660</td>
<td>30200</td>
</tr>
</tbody>
</table>

Detailed incident dataset
- 6-years incident data: 2011-2016
- 100m-precision position, incident type, weather conditions, vehicle involved, consequences
Block of maxima based on hourly values (BM-hour full)

**DATA**
- Dataset contains the highest negated-TTC value for each hour block.
- Only daytime hours considered (7:00 - 19:00).
- Only working days considered (Monday - Friday).
- 4 weeks data.

**60 maxima for each week, 240 in total.**

**CALIBRATION**
- Fitting the model to the whole dataset gives poor results.
- Apparently, blocks are not large enough, and the dataset contains negated-TTC values which are not really extremes.
- The solution is to set a “limit value”. All data below it are not considered extremes and are excluded from the dataset.

**How to choose the limit value?**
Model calibrated multiple times, increasing the limit value progressively.

Limit values above -0.9 are not acceptable due to an erratic shape parameter trend.

For safety reasons, the limit value chosen is the one (lower than -0.9s) that gives the highest number of predicted accidents; this value is **-1.04 s**.
Block of maxima based on hourly values (BM-hour full)

Parameters estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit value [s]</td>
<td>-1.04</td>
</tr>
<tr>
<td>N° of records</td>
<td>138</td>
</tr>
<tr>
<td>Shape parameter $\xi$ (S.E.)</td>
<td>0.0053 (0.0892)</td>
</tr>
<tr>
<td>Scale parameter $\sigma$ (S.E.)</td>
<td>0.1361 (0.0106)</td>
</tr>
<tr>
<td>Location parameter $\mu$ (S.E.)</td>
<td>-0.8640 (0.0139)</td>
</tr>
<tr>
<td>Neg. Log-Likelihood</td>
<td>-56.207</td>
</tr>
</tbody>
</table>

Annual accident estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N° of accident</td>
<td>3.4789</td>
</tr>
<tr>
<td>2.5% confidence</td>
<td>0.0000</td>
</tr>
<tr>
<td>97.5% confidence</td>
<td>25.588</td>
</tr>
</tbody>
</table>
Block maxima based on r-largest daily values (BM-day full)

DATA
- The block is defined as the whole day (00:00-23:59).
- Only working days considered (Monday-Friday).
- 4 weeks data.

5 blocks for each week, 20 in total.

CALIBRATION
- The blocks are large enough so that the highest negated-TTC value for each block really is an extreme.
- However, the model cannot be fitted on a dataset that contains only 20 values.
- R-largest negated-TTC values are selected from each block; the model is fitted to this dataset.

How to choose the R-value?

The model is fitted in both stationary (without covariates) and non-stationary (with covariates) conditions.
Experimental Analysis.
Analysis of results

Block maxima based on r-largest daily values (BM-day full)

- Model calibrated multiple times, increasing the R-largest value progressively.
- Parameters show a regular trend with R higher than 20.
- Negative log-likelihood minimum when R is between 20 and 22.
- The selected R-largest value is 21; it also gives the highest number of predicted accidents.
Experimental Analysis.
Analysis of results

Block maxima based on r-largest daily values (BM-day full)

STATIONARY MODEL

<table>
<thead>
<tr>
<th>Parameters estimation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R-largest value</td>
<td>21</td>
</tr>
<tr>
<td>N° of records</td>
<td>420</td>
</tr>
<tr>
<td>Shape parameter $\xi$ (S.E.)</td>
<td>0.1727 (0.0424)</td>
</tr>
<tr>
<td>Scale parameter $\sigma$ (S.E.)</td>
<td>0.1361 (0.0072)</td>
</tr>
<tr>
<td>Location parameter $\mu$ (S.E.)</td>
<td>-1.0527 (0.0097)</td>
</tr>
<tr>
<td>Neg. Log-Likelihood</td>
<td>-71.412</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annual accident estimation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N° of accident</td>
<td>0.5790</td>
</tr>
<tr>
<td>2.5% confidence</td>
<td>0.0733</td>
</tr>
<tr>
<td>97.5% confidence</td>
<td>1.6854</td>
</tr>
</tbody>
</table>
Block maxima based on $r$-largest daily values (BM-day full)

NON-STATIONARY MODEL - The model was fitted considering as a covariate the daily traffic volume.

The fitted non-stationary model was tested against the stationary model using the likelihood ratio test, resulting in a $p$-value of 0.0083, significantly smaller than alpha=0.05. R value is 21 in this case too (which results in 420 records).

<table>
<thead>
<tr>
<th>Parameters estimation</th>
<th>Annual accident estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape parameter $\xi$ (S.E.)</td>
<td>0.0188 (0.0484)</td>
</tr>
<tr>
<td>Scale parameter $\sigma$ (S.E.)</td>
<td>0.1696 (0.0072)</td>
</tr>
<tr>
<td>Location parameter $\mu$ (S.E.)</td>
<td>-1.1943(0.0098)+ +3.54959(0.0000)*10^{-6}*VOL</td>
</tr>
</tbody>
</table>
Comparison between single-week and 4-weeks approach

BM-hour and BM-day model are fitted with the full (4 weeks) and the partial dataset (only single weeks) datasets. Estimated number of accidents (and respective confidence intervals) are compared.
Experimental Analysis.
Analysis of results

Comparison between predicted and observed accidents

- Accident data in the case-study road segment are available in the 2011-2016 period (six years).
- Only rear-end accidents that occurred on working days are considered.
- The total number of observed accidents is 16; which is **2.67 observed accidents per year**.
- Three of these accidents did not occur during the daytime (7:00-19:00): they cannot be predicted with the BM-hour approach, which is fitted only with daytime data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predicted annual accidents [CI]</th>
<th>Annual mean of observed accidents [min;max]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM-hour full</td>
<td>3.48 [0.00;25.8]</td>
<td>2.17 [1;3]</td>
</tr>
<tr>
<td>BM-day full (stationary)</td>
<td>0.58 [0.07;1.69]</td>
<td>2.67 [1;4]</td>
</tr>
<tr>
<td>BM-day full (non-stationary)</td>
<td>0.71 [0.08;2.18]</td>
<td>2.67 [1;4]</td>
</tr>
</tbody>
</table>
Experimental Analysis.
Analysis of results

Comparison between predicted and observed accidents

The number of observed accidents depends on the distance from the section.

What is the distance from the section inside which we can predict accidents?

Further tests will be carried out in different locations.
Conclusions

- Concluding Remarks
- Future Developments
This study proposes an Extreme Value Theory approach to estimate freeway rear-end collision risk for Italian freeways.

The proposed BM approach (based on hourly and daily blocks) shows:

- overestimation of road accidents for BM-hour approach
- underestimation of road accidents for BM-day approach
- better performance using 4-weeks dataset for both approaches
- more precise estimates for BM-day approach

Future Research:

- test the introduction of new covariates for the BM-day model
- test Peak Over Threshold (POT) approach
- apply the models to other road sections
- apply the models to other types of accidents
Thank you for your attention...