A NOVEL APPROACH TO COLLISION HOTSPOT IDENTIFICATION ACCOUNTING FOR REGRESSION TO THE MEAN AND TREND

Lee Fawcett[1], Joe Matthews[1], Karsten Kremer[2], Neil Thorpe[1],
Fabio Galatioto[1], Timo Hoffman[2], Andre Muench[2]
[1]Newcastle University, Newcastle upon Tyne, UK; [2]PTV Group, Karlsruhe, Germany

1 BACKGROUND AND MOTIVATION

In this paper, we outline a new approach for collision hotspot identification based on a recent collaboration between researchers at Newcastle University, UK, and industrial partners at PTV Group in Karlsruhe, Germany. The primary aim of this work is the development of a tool for assessing the likelihood of accidents at known road safety hotspots, in the future, to enable a *proactive* – rather than *reactive* – approach to road safety scheme implementation. For example, during the roll-out of mobile road safety cameras in the UK during the late 1990s/early 2000s, a particular location was deemed worthy of treatment if the number of accidents, or in some cases casualties, exceeded some pre-determined threshold during an observation period. The effectiveness of the safety cameras was often then assessed by a before-and-after study in which attempts were made to separate any change in accident or casualty counts between the before and after periods into regression to the mean (RTM; see Section 1.1), trend and genuine treatment effect (see later). In the current paper, we focus mainly on an approach which has the potential to inform practitioners of the location of future road safety hotspots, allowing the implementation of some safety countermeasure before a high number of accidents is observed and thus potentially enabling the prevention of these accidents.

Our approach to such hotspot identification uses methods similar to those used in retrospective before-and-after studies to account for RTM, and exploits the Bayesian posterior predictive distribution to build estimates of RTM into accident predictions in future years. Incorrectly accounting for RTM and trend – or neglecting these phenomena altogether – could lead to a particular location being seemingly safer than it actually is, or perhaps unjustified investment in a site which might have an unusually high number of accidents. By construction, our method can also adjust accident counts in previous years to allow for the effects of RTM and trend, giving a more realistic assessment of the historic safety records of known hotspots.

1.1 Regression To the Mean (RTM)

The problem of *selection bias* in before-and-after studies is well known and well documented. When attempting to assess the effectiveness of a new road safety scheme, for example, sites selected for treatment are often those that have observed, over some pre-determined baseline period, an unusually high number of accidents or casualties; in any subsequent time interval, the accident/casualty count at these sites would probably reduce anyway, even if no treatment was implemented, simply because baseline counts were abnormally high. In such investigations, both ethical and economic concerns are often cited as reasons against completely randomised studies; as a result, post-treatment separation of the true causal effect from any change that would have occurred anyway, without treatment – the RTM effect – has received much attention. The main consequence of ignoring RTM is often an exaggerated treatment effect and possibly unjustified financial investment.

In the road safety literature, an empirical Bayes (EB) approach is usually employed to quantify the RTM effect; see, for example, Fawcett and Thorpe (2013), who also advocate the use of fully Bayesian (FB) techniques. Although the effect is variable, studies typically show a reduction in accident frequency owing to RTM of between 20% and 30% (Hirst *et al.*, 2004); in other words, estimates which do not account for RTM would typically be biased by 20–30%. However, some studies, as discussed in Fawcett and Thorpe (2013), have shown RTM to account for 100% of the observed accident reduction at some sites; others even point to a treatment disbenefit, where accident counts remain significantly higher than historical counts, even after accounting for RTM and trend. When using historical counts of accidents to predict future counts in a bid to identify sites which might be hotspots in future years, estimates of RTM should also be made in an attempt to 'clean' any forecasts.

As an example, we now outline the results of a before-and-after study of the effectiveness of mobile road safety cameras in the northeast of England.

1.2 Northumbria Safety Camera Partnership (NSCP) Retrospective Study

1.2.1 General findings

The Northumbria Safety Camera Partnership (NSCP) joined the national programme of UK safety camera partnerships in April 2003. In February 2004, the Partnership commissioned a team of researchers to investigate specifically the impact of operating mobile road safety cameras on the demand for secondary health care at the region's hospitals. The study group collected data from 56 such sites in the region from a before period (April 2001–March 2003) and an after period (April 2004–March 2006). Full results are reported in Thorpe and Fawcett (2012) and Fawcett and Thorpe (2013): in summary, of the observed reduction in casualties from 436 in the before period to 298 in the after period, it was estimated that 119 casualties would not have happened anyway due to RTM (111) and trend (8), leaving an estimated treatment effect of just 19 casualties out of a total observed reduction of 138. Direct savings to local NHS secondary healthcare providers as a result of these 19 casualties saved were estimated to be up to £56,000, though when attempting to account for *all* aspects of the valuation of casualties – including police costs, human costs (representing pain, grief and suffering to the casualty and their relatives/friends) and loss of output due to injury – this estimate rises to around £4.1 million.

1.2.2 Statistical modelling details

The method used to estimate the effectiveness of the mobile safety cameras initially follows an EB statistical modelling approach. Here, the casualty frequency y_j at site j in the before period is assumed Poisson-distributed with rate λ_j. The rate parameter is itself allowed to vary according to a Gamma distribution; specifically,

$$\lambda_j \sim \text{Gamma}(\text{mean} = \mu_j,\ \text{shape} = \gamma). \qquad (1)$$

Then, using Bayes' Theorem, the *posterior distribution* for the site-specific rate of casualties can also be shown to be Gamma with mean given by:

$$EB_j = \alpha_j \mu_j + (1 - \alpha_j) y_j, \qquad (2)$$

where

$$\alpha_j = \gamma / (\gamma + \mu_j).$$

The value of EB_j in Equation (2) is the empirical Bayes estimate of casualty frequency – the casualty frequency we would expect to observe if no safety countermeasure were introduced. Notice that this is a weighted sum of μ_j and y_j. Hence, we combine the unusually high observed casualty frequency in the before period, y_j, with what we would *expect* to see at this site according to our prior beliefs about the mean casualty rate μ_j. In other words, after taking our prior beliefs about the mean casualty rate μ_j into account, we would expect the observed casualty frequency to change from y_j to EB_j anyway, even if no safety countermeasure were introduced. Thus, the RTM effect is the difference between y_j and EB_j, and the treatment effect is the difference between EB_j and the observed casualty count in the after period. Of course, trend would erode the estimated treatment effect still further; see Section 4.3 of Fawcett and Thorpe (2013) for a full description of how trend can be catered for.

In order to compute EB_j in Equation (2) we need μ_j and γ. Since μ_j represents our prior beliefs about the mean casualty rate at site j, when casualty counts are not at the 'peak of a blip' (as is assumed the case with the observed y_j), it is common to use a set of reference sites to build an *accident prediction model* (APM) to find a suitable value for μ_j, that is, an estimating equation for μ_j is obtained using information on other covariates collected at the reference sites, such as average speed, traffic flow, road type etc. Of course, the reference sites must be representative of the treated sites in terms of these covariates, but *not* in terms of their casualty frequencies – the reference sites have not been chosen for treatment and so counts here are considerably and consistently lower than those at the treated sites. The authors are currently investigating the use of a technique known as *propensity score matching* (see, for example, Li *et al.*, 2013) to allow automatic selection of a set of optimal reference sites for use in this analysis. A suitable value for γ is also obtained from the method used to generate the estimating equation for μ_j; specifically, it can be shown that the APM has a negative binomial error structure with dispersion parameter γ, and the method of maximum likelihood is used to simultaneously estimate μ_j and γ – see Fawcett and Thorpe (2013) for full details.

In an EB analysis, modelling is restricted to fall within the *Poisson-gamma* structure as outlined above. As Miao and Lord (2003), Maher and Mountain (2009) and Fawcett and Thorpe (2013) discuss, a fully Bayesian (FB) analysis, making use of Markov chain Monte Carlo (MCMC) methods, allows the use of *any* statistical model for the casualty rate λ_j in Equation (1). The *deviance information criterion* (DIC; see Speigelhalter *et al.*, 2002) can then be used as a tool for selecting the 'best' model. Unlike the EB approach, a FB analysis also properly accounts for uncertainty in the estimation of the APM; as Li *et al.* (2008) and Fawcett and Thorpe (2013) show, this results in EB estimates of RTM and treatment effects with error margins that are much too small.

1.3 Aims of Current Work

The main aim of the current work is to learn from the NSCP study about optimal strategies for estimating RTM and apply these methods in a road traffic accident hotspot identification scenario. In particular, the authors are currently developing an Application for use by practitioners for predicting accident or casualty counts in future years to help inform targeted funding for road safety countermeasures, in a bid to prevent excessive accidents or casualties at known road safety hotspots before they happen. The intention is that this Application will be extremely user-friendly and the statistical modelling automatic, using sensible mathematical defaults wherever possible. However, more experienced modellers will have the option to over-ride these defaults and experiment with the modelling assumptions, should they desire. Various options will be available for exploratory data analyses, and the main results, in terms of predicted numbers of accidents or casualties in future years at known road safety hotspots, will be presented in convenient graphical and tabular forms. The Application will also feature a module to perform a road safety scheme evaluation analysis, as discussed in 1.2 for the NSCP study.

2 THE DATA

The data used in this Section, provided by PTV group, consist of annual accident counts at 734 nodes, in and around the city of Halle (Saxony-Anhalt, Germany), for the years 2004-2012 inclusive. All 734 nodes might be considered as potential hotspots, although none have, as yet, been treated with any road safety countermeasure. Accompanying these figures are observations on several covariates, including speed limit and annual traffic volumes at each node and various other indicators of node type (e.g. Urban area? Intersection? Signalised? Four-legged junction?). Figure 1 shows the location of these nodes as shown in PTV's *Visum Safety* software, as well as the location of the city of Halle in Germany and a photograph of the notorious "Riebeckplatz" hotspot, one of the greatest traffic volume intersections in Germany. Exploratory analyses reveal significant associations between many of the covariates and accident counts. For example, the Pearson product moment correlation coefficient between traffic volume and accident counts, across all sites in 2005, is 0.876; similarly, the mean number of accidents for urban areas in 2007 is 15.2, compared to just 3.7 for non-urban areas.

Figure 2 shows plots of accident counts through time at four nodes (labelled "Site 1" through to "Site 4"). From these plots it is obvious that, year-on-year, changes in accident counts are prone to both trend and RTM. For example, counts at Site 1 in 2004 and Site 3 in 2006 appear abnormally high and low, respectively, relative to counts in most other years at these sites. At the end of 2004, Site 1 might be deemed dangerous enough to warrant action in the form of a safety countermeasure – however, as we observe, counts subsequently regress towards their underlying mean level anyway, without any action. Conversely, Site 3 in 2006 appears unusually safe relative to counts in other years at this site. To some degree, Sites 1–3 also appear to exhibit some form of trend – perhaps generally increasing at Sites 1 and 3 and decreasing at Site 2. In fact, counts at Site 1 are interesting from a trend point-of-view: the outlier in 2004 seems to suggest a decreasing trend between 2004–2006. However, from 2006 onwards there is clearly a steady increase in accidents. So large is the value in 2004 that any simple approach to modelling trend at this site might indicate an overall decreasing trend, completely swamping the clear positive trend observed in more recent years. It is therefore important for any analysis to take this into account, perhaps allowing predictions for future years to give more weight to counts in more recent years and, of course, to adjust past observations, where necessary, to take RTM into account. The red question marks at 2013 and 2014 in Figure 2 illustrate the objective of our analysis: having "cleaned" for RTM and trend, can we get a handle on future counts, with a view to taking action *before* a site becomes genuinely dangerous? Or – given recent high counts – on adjusting for RTM and trend might our predictions indicate a site as being less dangerous than recently observed and so not warrant action in subsequent years?

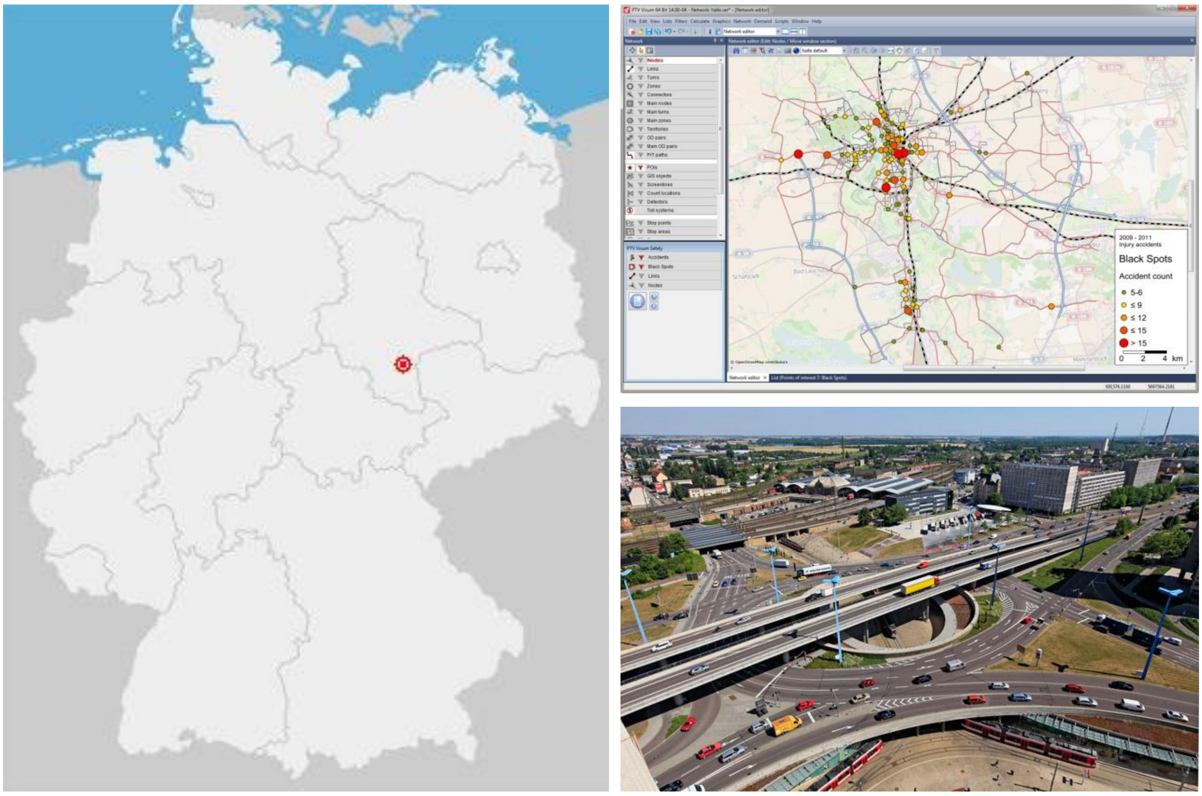

Figure 1: Left: Halle, in Saxony-Anhalt; middle: location of potential hotspots; right: 'Riebeckplatz' hotspot, site 22: greatest traffic volume intersection in Germany

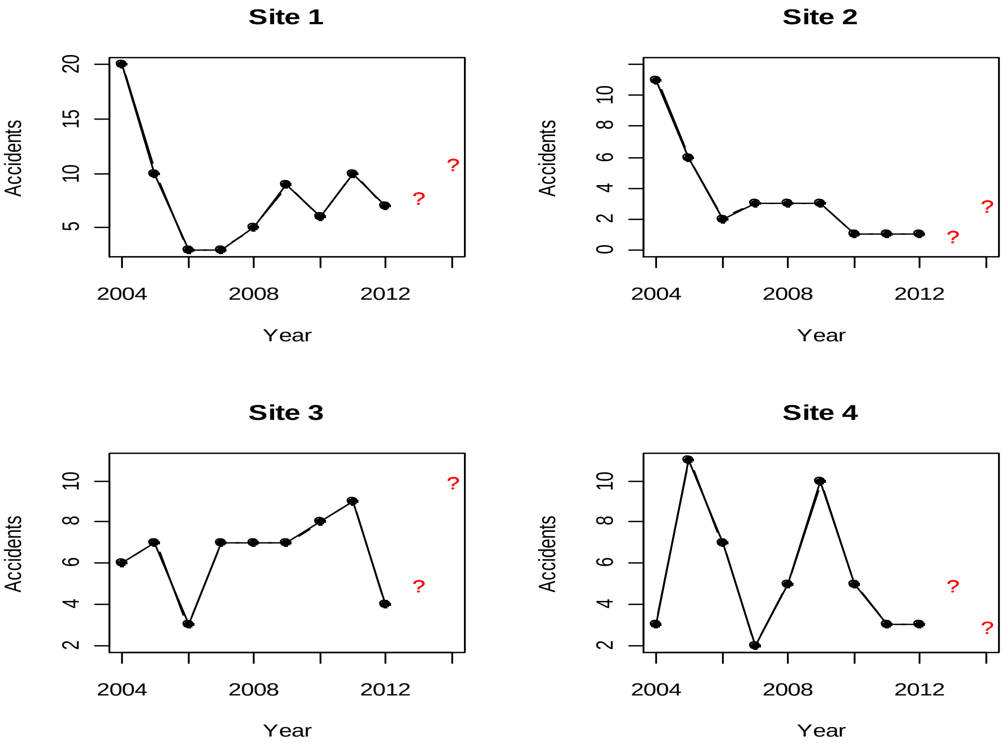

Figure 2: Observed accident counts, 2004–2012, at four of the 734 in and around Halle.

3 STATISTICAL MODELLING

For accident counts in the current year, we adopt the same modelling assumption as used in the NSCP study in Section 1.2.2; that is, we assume counts y_j at site j follow a Poisson distribution with rate λ_j. However, to allow historical values to inform our predictions for future years, whilst at the same time giving more weight to values in more recent years, we relax the Poisson assumption of equal mean and variance for observations in past years by assuming a negative binomial model here, with mean $\lambda_j(t)$ and variance $\lambda_j c(t)$, where $c(t)$ is a variance inflation factor dependent on time t, and $t = 0$ for the current year and –1, –2, ... for years in the past. The exact form of $c(t)$ is chosen such that the variance of y_j increases as t decreases, giving more variability to observations further in the past and allowing more recent counts more weight in how they inform future predictions.

As with the NSCP study in Section 1.2.2, observed accident totals in each year are adjusted to take into account our expectations for each site according to a global APM estimated using counts, and covariate information (e.g. traffic volumes, speed limit etc.), collected across all 734 nodes as a whole. We do this by constructing the APM for $\mu_j(t)$, considered to be our prior belief about likely values for $\lambda_j(t)$ at site j, time t. Hence, abnormally low / high counts are inflated / deflated accordingly, these adjustments representing the RTM effect. The time indicator t is included as a covariate in this APM to allow the inclusion of global trends in these adjustments, although any trends deemed not substantial (e.g. as in the years 2005–2007 at site 1 in Figure 2) are penalised, with trends observed over longer time periods having greater importance. We also allow changes in the infrastructure at each node to inform our APM via *crash modification factors*, these being known multiplicative factors depending on the type of intervention used at the node in question. To fully capture all source of variability in the data, we proceed with a fully Bayesian analysis as discussed in Section 1.2.2. Although this has the effect of giving greater uncertainty to our predictions of accident counts in future years, at least all sources of variability – including in the estimation of the APM itself – are properly acknowledged (unlike in an EB approach). The *Bayesian posterior predictive distribution* is exploited (see, for example, Lee (2012) for details) to incorporate all of these features in our predictions of counts in future years. A more complete mathematical exposition will be available in Fawcett *et al.* (2015).

4 RESULTS

Figure 3 shows some results of our analysis, returning to Sites 1–4 from Figure 2 for convenience. For example recall that, in Section 2, we discussed that accident counts at Site 1 in 2004 seemed abnormally high. The blue line represents the APM, and shows what we would expect to see at each site, in terms of accident counts, ‘ordinarily’. We then combine both the observed value, and the APM value to obtain our full Bayes estimate of accident frequency shown by the red line (with 95% confidence bounds shown by the red dashed lines; notice that, in general, these become narrower as we move from left to right, with increasing certainty in our fully Bayes estimates of accident frequency as more data comes online). Thus, the shift from the observed value to the red line is taken as the RTM effect, and this effect is considerable for Site 1 in 2004. Note also that the abnormally high accident count in 2004 has been penalised so as not to unduly influence the overall estimated trend in the APM. We also commented that site 3 in 2006 seems to have an abnormally low accident count. Adjusting form both trend and RTM gives an inflated value pulled towards the APM, as shown by the red line. Generally, the closer the red and blue lines are together, the lesser the role of RTM at a particular site.

Shown in grey are the predictive values for the year 2013; the vertical lines are the 95% confidence intervals around these predictions and represent our uncertainty here. These represent our beliefs about future values of accidents at each site after taking into account trend and RTM, as well as uncertainty in the estimation of the APM. Of course, predictions might be sensitive to the modelling choices outlined in Section 3; as discussed in Section 1.2.2, the DIC can be used to automatically judge many different models, and results such as those shown in Figure 3 could be produced for the 'best' model. This is our intention for the software tool we are developing, as discussed in Section 5. It is also our intention to run various model validation checks, including a full comparison of 2013/2014 predictions with the raw observed values in these years.

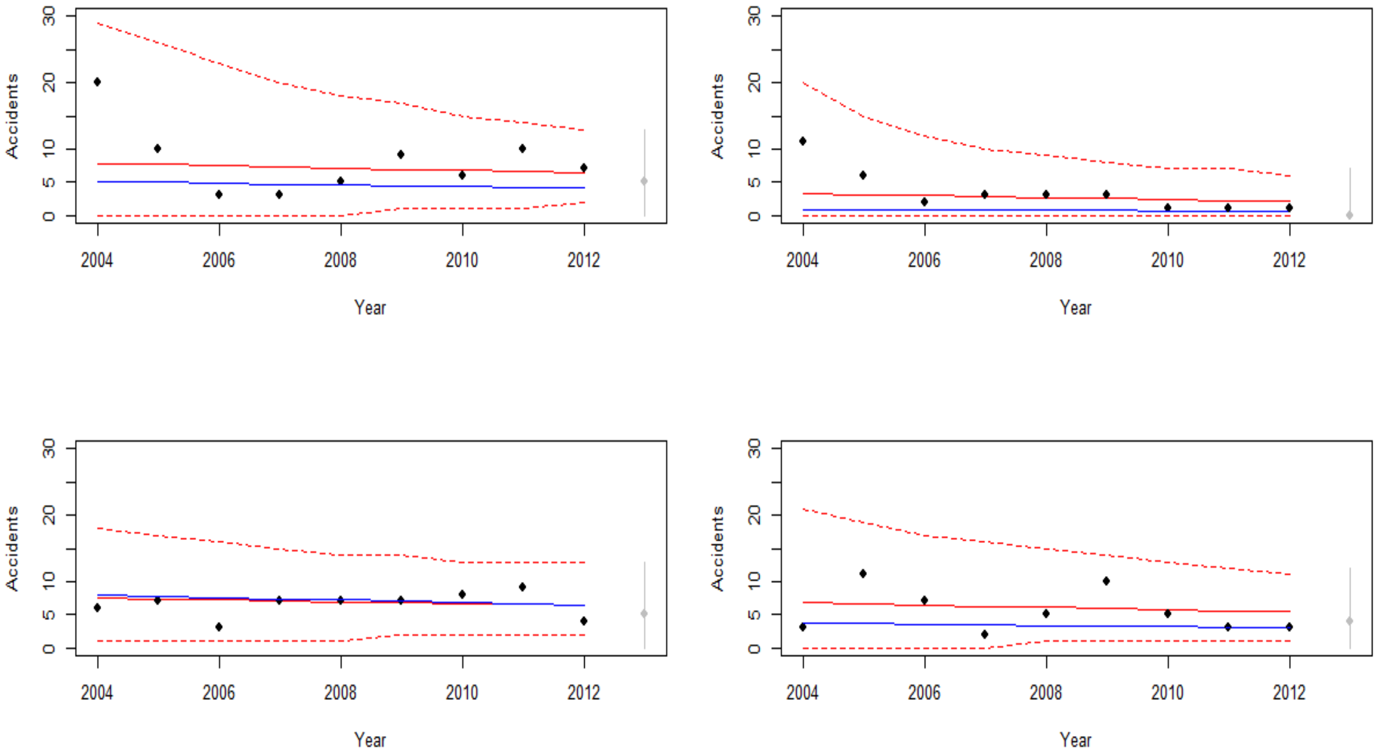

Figure 3: Observed accident counts (black points) for four nodes in and around Halle. The blue line represents expected counts from the APM, and the red line shows full Bayes estimates of these counts, taking RTM and trend into consideration. The red dashed lines are the 95% confidence limits around the full Bayes estimates. The grey points/lines show our predictions for the year 2013, with associated uncertainty.

5 A TOOL FOR SAFETY SCHEME EVALUATION / HOTSPOT IDENTIFICATION

The Newcastle University authors of this paper were recently the recipients of a University research grant aimed at "pump-priming" research-based impact work. The main focus of this work is to produce an easy-to-use tool for performing before-and-after safety scheme evaluation (e.g. Section 1.2) or hotspot identification (e.g. Sections 2–4), accounting for trend and RTM as we do in this paper. The tool uses the *Shiny* application, a web-based application framework for the statistical software package R. Both R and *Shiny* are free to download, and R has a huge range of built-in intrinsic functions for statistical analysis, as well as many add-on packages. In order to use the tool for safety scheme evaluation/hotspot identification, the user will need R installed on their machine; once the *Shiny* application has been installed, the user is presented with a simple graphical user interface. For safety scheme evaluation, the user must upload data from the treated sites and the reference sites, and a wide variety of file types (e.g. comma-separated, tab-separated etc.) are supported. Unless the user wants to change some of the default modelling settings, a simple click of the button will then perform the analyses described in Section 1.2, with a results page being produced giving a breakdown of changes in accident/casualty figures after having implemented some road safety scheme, separated into estimates of RTM, trend and genuine treatment effect. The user will be able to sort results by various attributes – e.g. sites will be able to be sorted by increasing/decreasing RTM effect or treatment effect, or perhaps sites having observed a genuine treatment effect exceeding a certain target might be mapped geographically. We are currently in consultation with colleagues at the Tyne and Wear Traffic Accident Data Unit, and other Local Authorities, to explore the various options for presenting results from hotspot identification analyses as described in Sections 2–4, although we have been told that graphics such as those shown in Figure 3 would be informative. An early prototype of this Application will be demonstrated during our presentation at the TPM on July 1st.

REFERENCES

Fawcett, L., Matthews, J.T., Kremer, K., Thorpe, N., Galatioto, F., Muench, A. and Hoffmann, T. (2015). A Novel Approach to Collision Hotspot Identification accounting for Regression To the Mean and Trend. *In preparation.*

Fawcett, L. and Thorpe, N. (2013). Mobile safety cameras: estimating casualty reductions and the demand for secondary healthcare. *Journal of Applied Statistics,* **40**, 11, pp. 2385-2406.

Hirst, W.M., Mountain, L.J. and Maher, M.J. (2004). Sources of error in road safety scheme evaluation: A quantified comparison of current methods. *Acci. Anal. Prev.* **36**, pp. 705–715.

Lee, P.M. (2012). *Bayesian Statistics: An Introduction.* Wiley, London.

Li H., Graham D.J., Majumdar A. (2013). The impacts of speed cameras on road accidents: An application of propensity score matching methods. *Acci. Anal. Prev.* **60**, pp. 148-157.

Li, W., Carriquiry, A., Pawlovich, M. and Welch, T. (2008). The choice of statistical models in road safety countermeasure effectiveness studies in Iowa. *Acci. Anal. Prev*. **40**, pp. 1531–1542.

Maher, M.J. and Mountain, L.J. (2009). The sensitivity of estimates of regression to the mean. *Acci. Anal. Prev*. **41**, pp. 861–868.

Miaou, P. and Lord, D. (2003). Modeling traffic crash flow relationships for intersections: Dispersion parameter, functional form and Bayes versus empirical Bayes methods. *Transp. Res. Rec.* **1840**, pp. 1310–1340.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc. B* **64**, pp. 583–616.

Thorpe, N. and Fawcett, L. (2012). Linking road casualty and clinical data to assess the effectiveness of mobile safety enforcement cameras: a before and after study. *BMJ Open*, **2**: e001304.