Crashes Insights & Modelling

Lina Berbesi

May, 2024

Unsplash images Taken from: https://unsplash.com/photos/

Situation

Insights on crashes can assist in a better understanding of the effect of various factors on crashes, leading to developing effective countermeasures. [1]

For government purposes crashes represent both social and financial costs in terms of people injured, life lost and claims logged through ACC.

For research purposes crashes represent an interesting opportunity where a variety of quantitative methods could be used. From classification models based on the type of crash or crash severity, to regression models and density analysis based on the number of crashes.

[1] Sage https://journals.sagepub.com/doi/10.1177/036119812110378882

Action/ Results

Accidents data was obtained from CAS [1]

Spatial analysis and modelling started by doing a point-in-polygon allocation from the crash geo-locations to spatial units, in this case both the territorial authorities and regional council administrative boundaries.

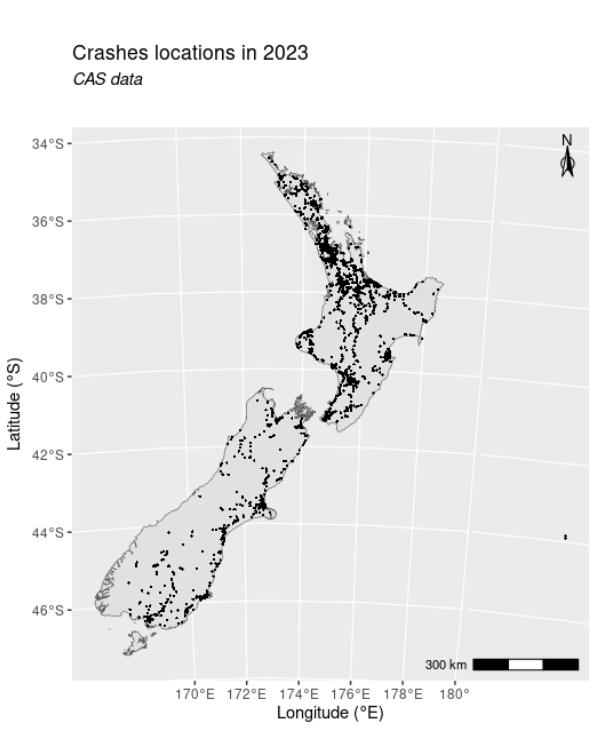
Point pattern analysis and density analysis for accidents that occurred in 2023 were carried out to identify how they behave related to the degree of spatial aggregation.

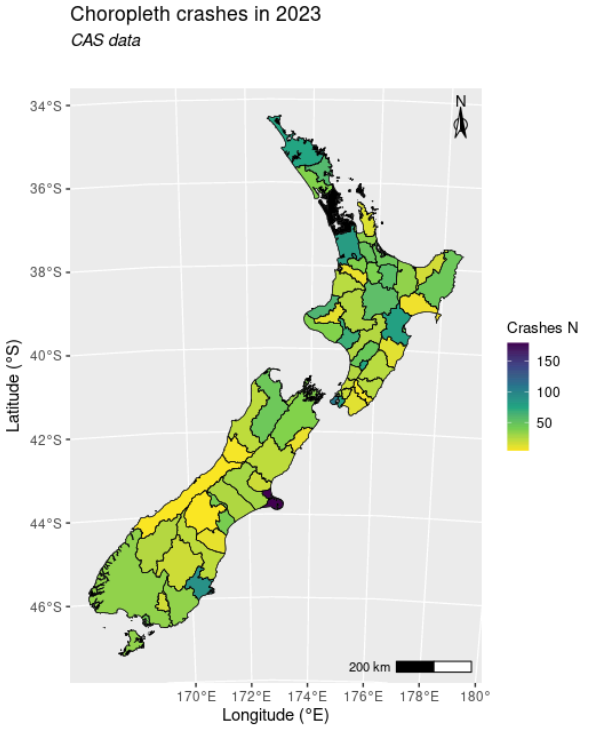
Lastly the generalised case of a poisson, a negative binomial bayesian regression was applied in the last five years to unveil the effect of road/network variables in the number of crashes.

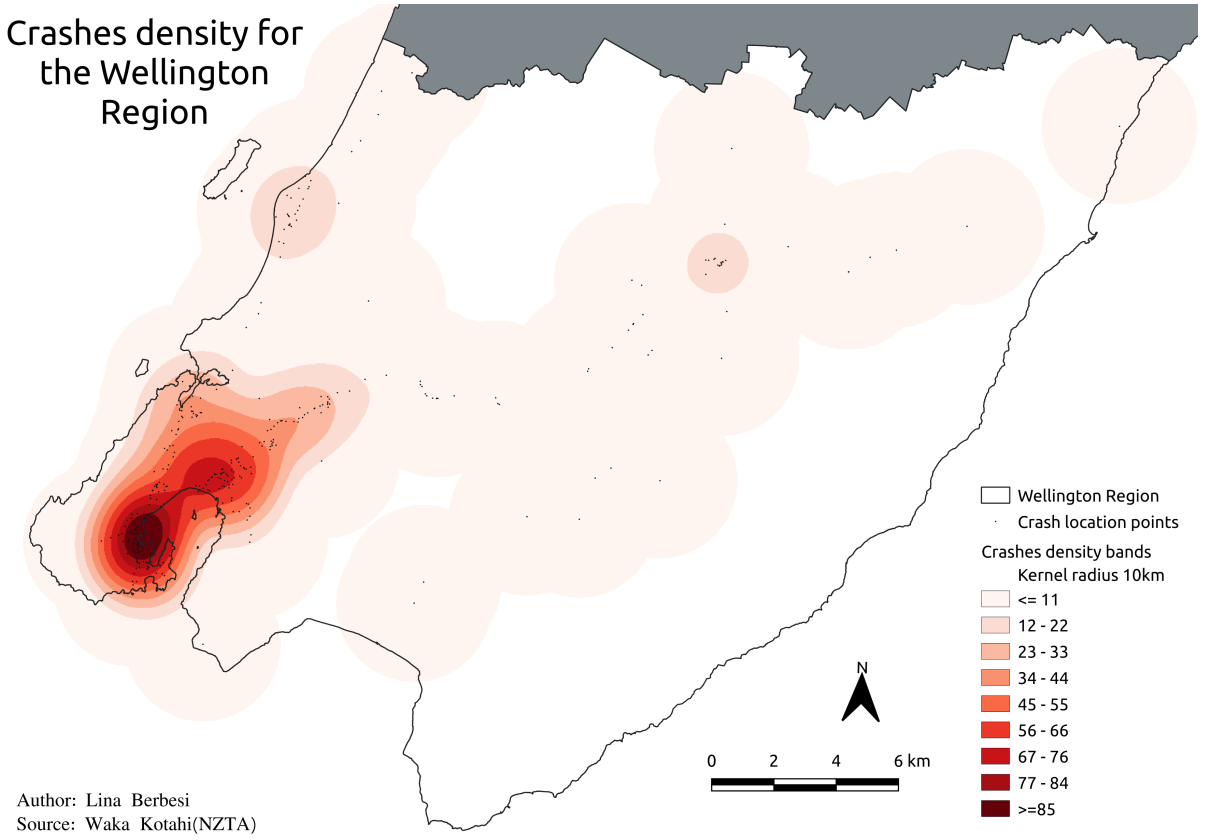
[1] NZTA https://opendata-nzta.opendata.arcgis.com/

Insights

Urban centres, Auckland and Wellington City, register the higher number of crashes in the North Island. Followed by rural districts like Far North and Hastings. While in the South Island crashes seem to be clustered in the Canterbury and Otago regions, more specifically around Christchurch and Dunedin.

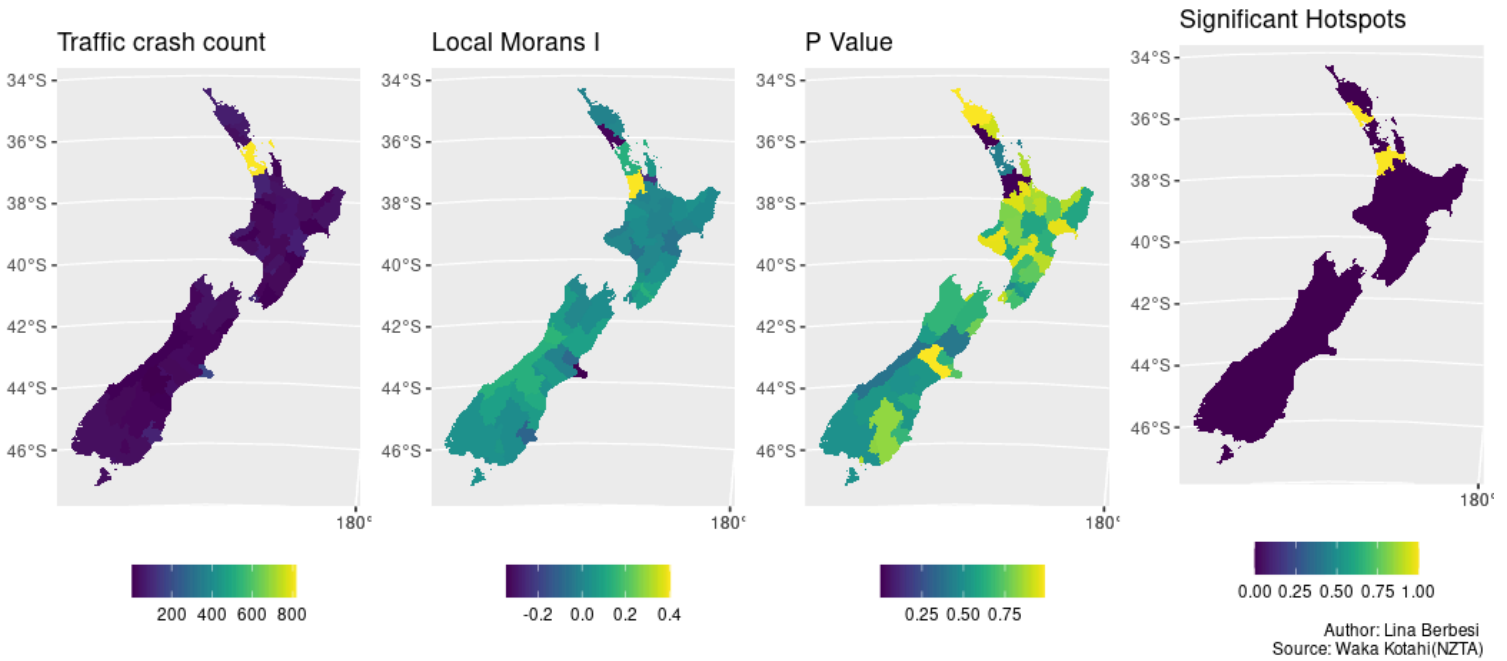






When using kernel density estimation (hot spot analysis) over the Wellington region , it can be seen that there are high clusters of traffic accidents in the Wellington City and Lower/Upper Hutt Districts. This is particularly true for the suburbs of Thordon, Pipitea, Wellington central, Te Aro, Oriental Bay and Mount Victoria where crashes are registered at rates of seventy-seven or over. Clusters with medium to low rates are registered in Masterton and Kāpiti.

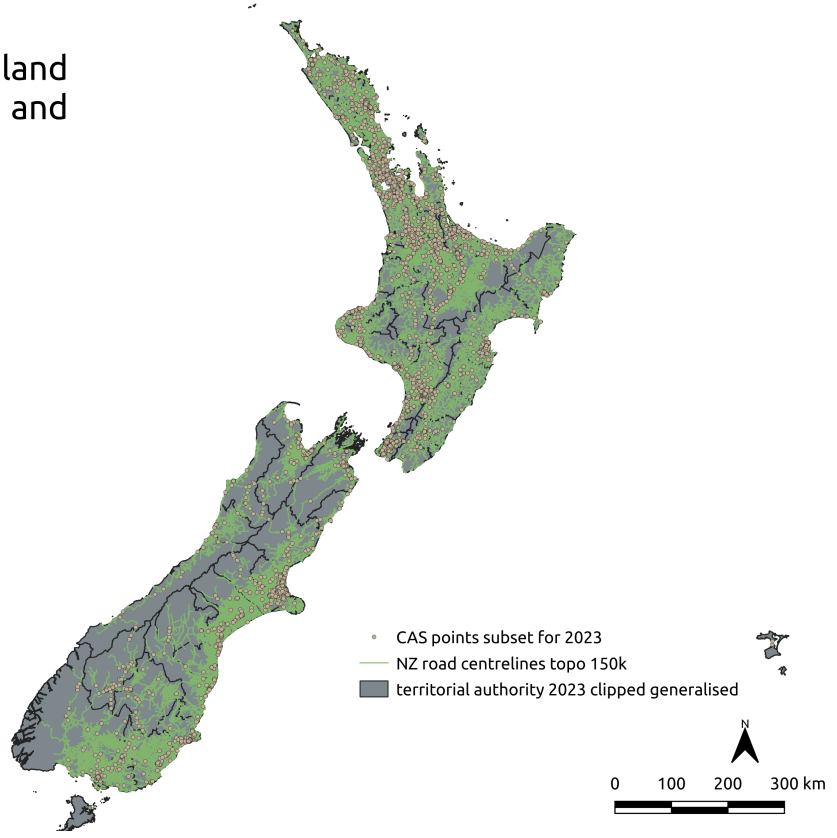
To measure the spatial autocorrelation Moran's I was calculated. In addition to Moran's I, p-values and significant hotspots where p-values were less than 0.05 to be considered statistically significant were calculated. Significant hotspots show areas where the null hypothesis from Moran's I, attributes distribute randomly, should be rejected.



Model

For the model additional network variables such as the count of roads by polygon were included. Lower geography aggregation, meshblocks, although desirable is hard to accomplish in terms of information publicly available.

Overlay of New Zealand Roads Centrelines and Crashes for 2023



Author: Lina Berbesi
Source: Waka Kotahi(NZTA)

For the model fitting instead of using a poisson linear regression where only a single regressor can be used to explain the response to traffic accidents. A Bayesian negative binomial approach was preferred since it allows to incorporate multiple information from a set of predictor variable vectors x_i through more than a single beta $beta_k$. The negative binomial considers the results of a series of trials that can be classified either as a success or a failure. Being the poisson a special case of the negative binomial distribution.

$$y_i \sim \text{NegBinomial}\left(\mu_i, r\right)$$

Where $log\left(\mu_i\right) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}$ for i $\in 1...$n

```
# Bayes Model
nb_bayes <- brms::brm(
    crashcnt ~ speedlimit + hghwycnt + I(sealed > 0),
    data = crash_paneldata_nzta_train,
    family = negbinomial(link = "log")
  )


# Frequentist Model
nb_freq <-
  MASS::glm.nb(crashcnt ~ speedlimit + hghwycnt + I(sealed > 0),
               data = crash_paneldata_nzta_train)
```

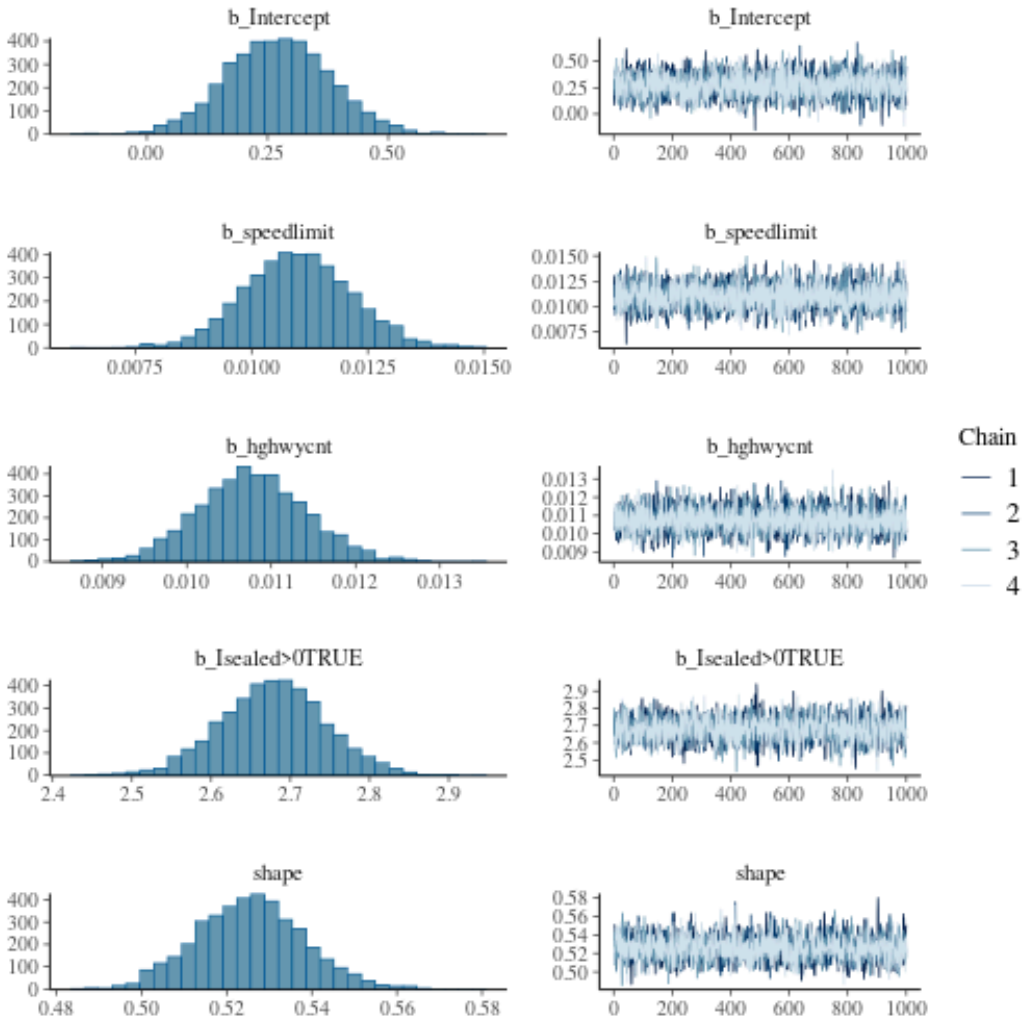
```
summary(nb_bayes)$fixed[,1:5]
```

```
##                  Estimate    Est.Error    l-95% CI   u-95% CI      Rhat
## Intercept     0.27214588 0.1067875497 0.063420036 0.48484964 1.0004056
## speedlimit    0.01100034 0.0011556058 0.008757319 0.01324745 0.9994353
## hghwycnt      0.01077597 0.0006415639 0.009553326 0.01205613 1.0013450
## Isealed>0TRUE 2.67861649 0.0667054768 2.547652542 2.81078313 1.0019642
```

```
summary(nb_freq)$coefficients
```

```
##                     Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept)       0.27358023 0.0951534011  2.875149 4.038366e-03
## speedlimit        0.01097579 0.0010270415 10.686803 1.173466e-26
## hghwycnt          0.01073592 0.0006652002 16.139386 1.348757e-58
## I(sealed > 0)TRUE 2.67844799 0.0636587409 42.075102 0.000000e+00
```

```
plot(nb_bayes)
```



Conclusions

- Categorical variables in the crash characterization indicate that multinomial logistic regression models could also be fitted to predict variables such as the crash severity.
- Additional numerical variables can be drawn for the roads network. In this case, the number of highways per polygon/geographical boundary was included into the model to complement the existing information. Demographic variables (not included due to time constraints) such as population although not strictly indicative of the network/roads quality could be a good indicative of other non-publicly available variables such as the roads flow(average of daily traffic) which could also have an impact in the number of crashes.
- Local regressions at TA or meshblock level could potentially be applied using a Geographically Weighted Regression model if more time was available.

Conclusions

- The proposed bayesian negative binomial model, though not perfect, constitutes a sufficient model with a good acceptance ratio, a robust log posterior to step size and tree depth, $\hat{r}$ values are all below 1.01, an effective sample size above 0.5 and a small MCMC error to posterior sd.
- Bayesian priors, although initially set, were dropped due to the nature of the response variable, discrete instead of continuous.
- Both Bayesian and Frequentist approaches converge to the same model coefficients.
- Fixed effects on the territorial authorities or the regions can be added but it will extend the convergence time.

Thank you!

@linkedin.com/in/lina-berbesi
lina.berbesi@gmail.com