

# 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.<sup>1</sup>

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/10.1177/03611981211037882>

# Action/ Results

Accidents data was obtained from CAS<sup>1</sup>

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.

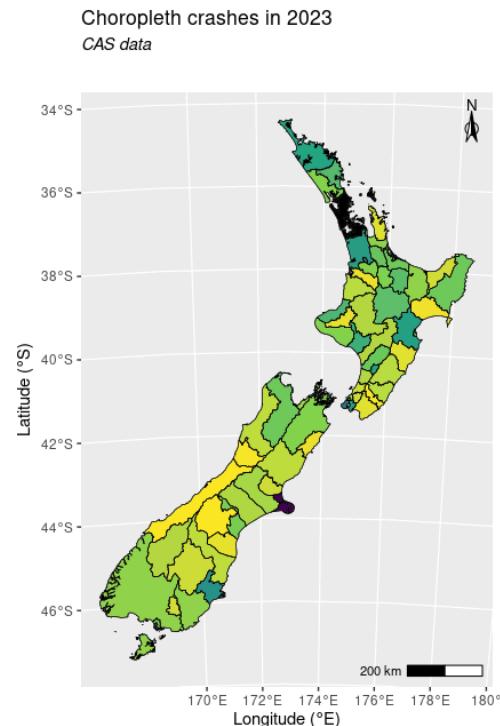
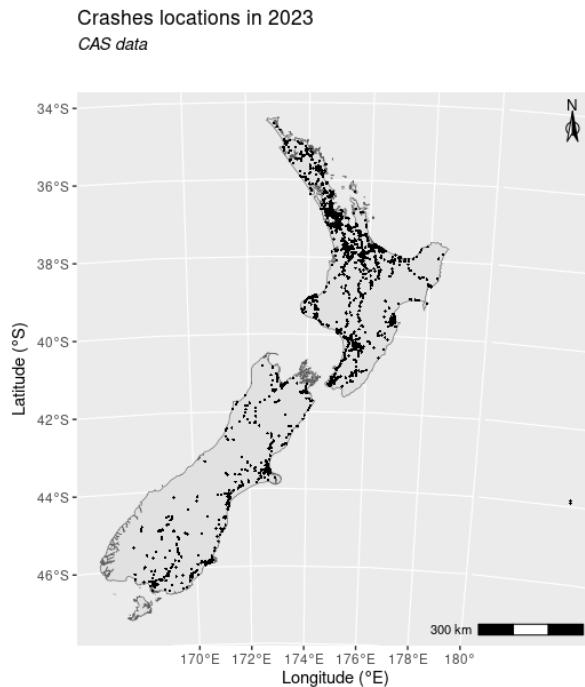
Point pattern analysis and density analysis for accidents that occurred in 2023 were carried out to identify how do they behave related to the degree of spatial aggregation.

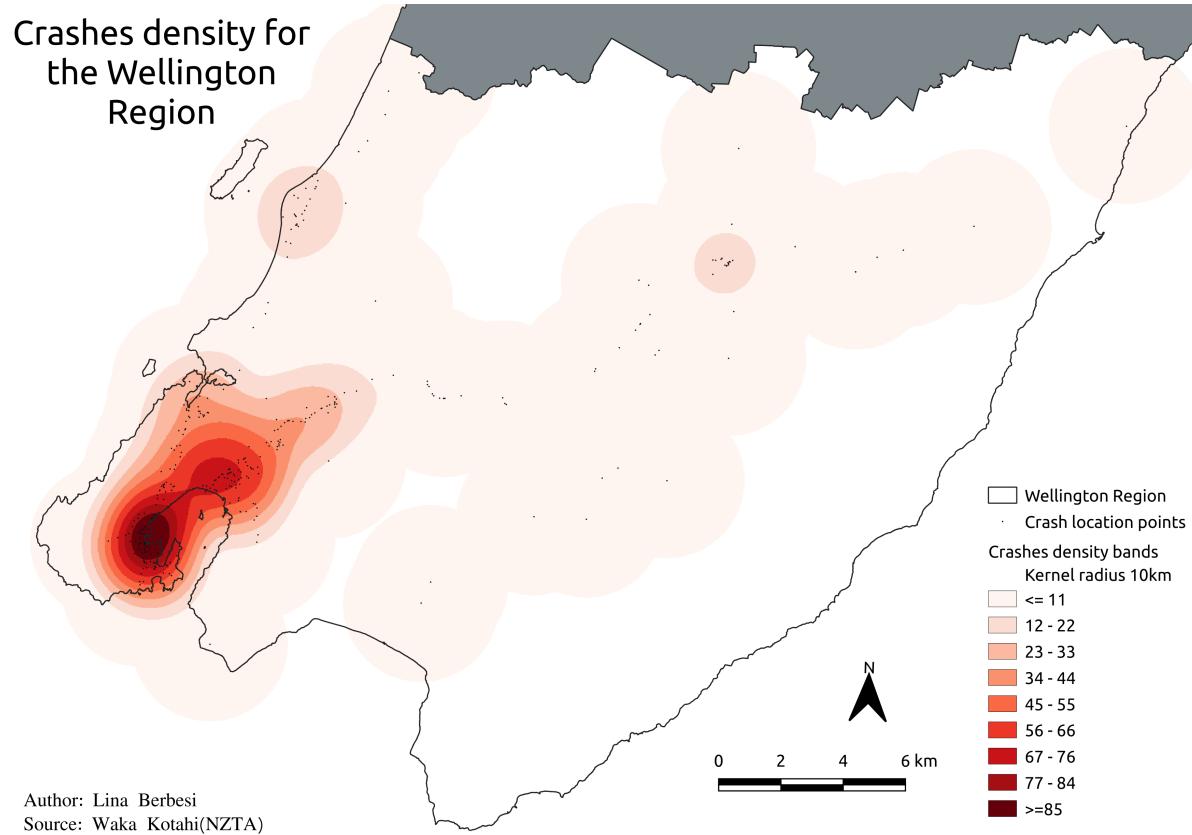
Lastly a poisson bayesian regression was applied on the last five years to unveil the effect of road/network variables in the number of crashes.

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

# Insights

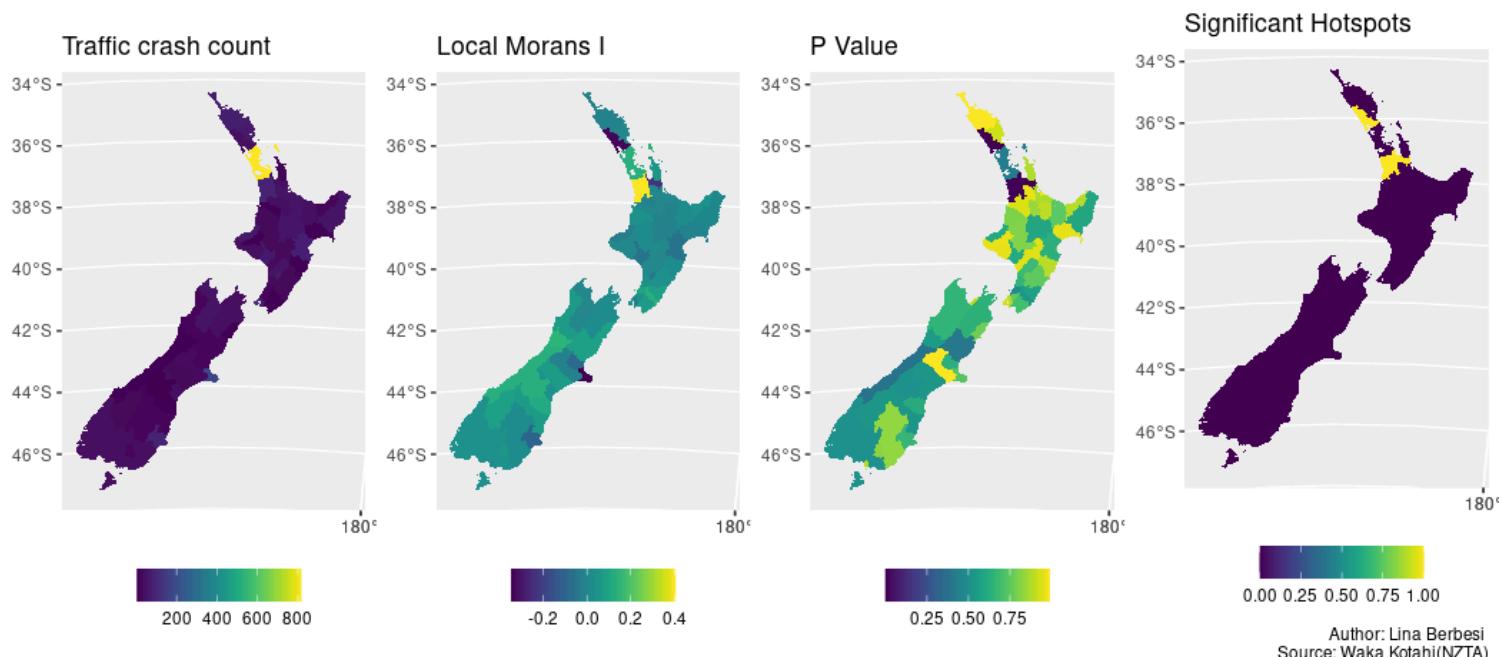
Urban centers, Auckland and Wellington City, register the higher number of crashes in the North Island. Followed by rural districts like Far North and Hastings. While in the South Island crashes seem to be clustered in the Canterbury and Otago regions, more specifically around Christchurch and Dunedin.





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

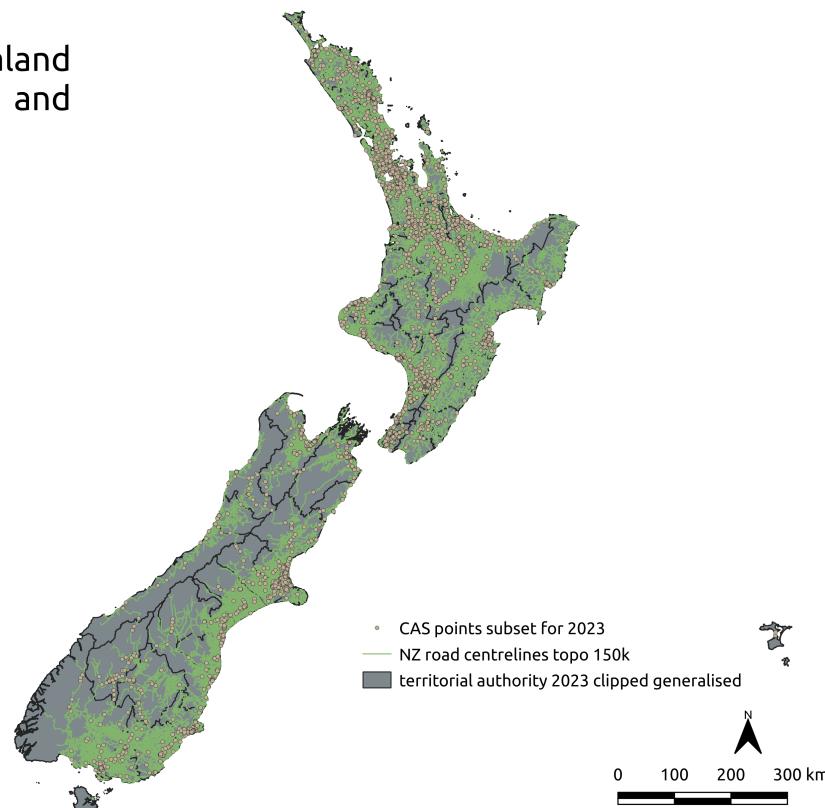
To measure the spatial autocorrelation Moran's I was calculated. In addition to Moran's I, p-values and significant hotspots where p-values were less than 0.05 to be considered statistically significant were calculated. Significant hotspots show areas where the null hypothesis from Moran's I, attributes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



For the model fitting instead of using a poisson linear regression where only single values are obtained for the model parameters of the traffic accidents. Bayesian linear poisson regression was preferred since it allows to incorporate external information of what it is being known about the parameters and also it gives back distributions instead of single values.

*with(datamodel, list(Y = crashescount, X = sealed, Z = speedlimit, W = hgh*

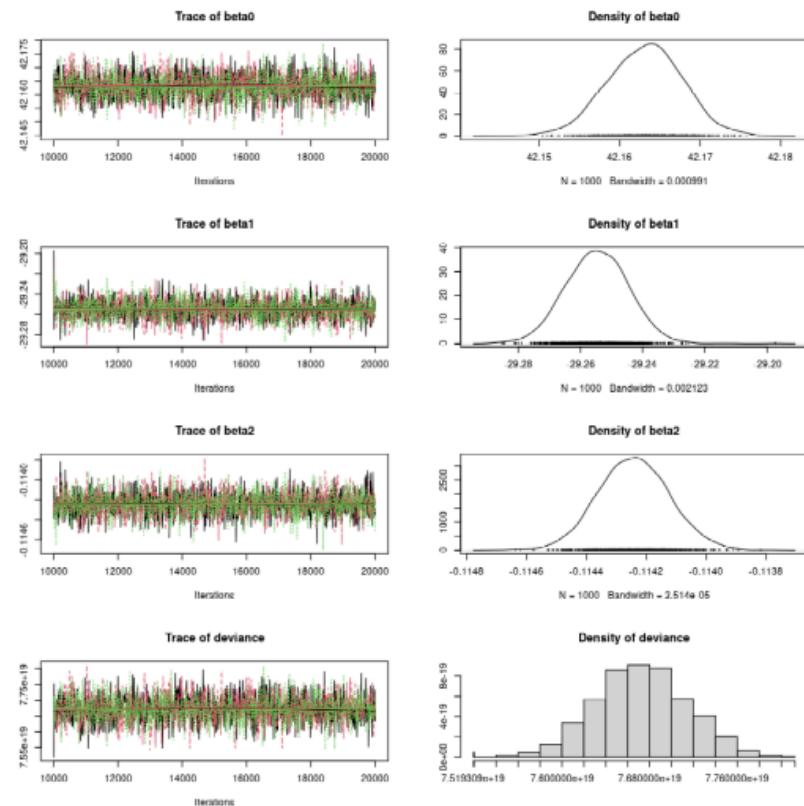
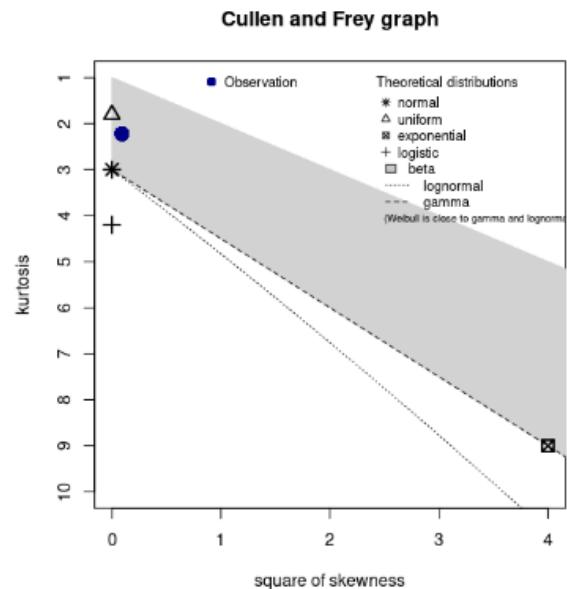
```
# Likelihood model for Y[i]
poisson_bayesian_model<-for(i in 1:length(Y)) {
  Y[i] ~ dpois(l[i])
  log(l[i]) <- beta0 + beta1*X[i] + beta2*Z[i] + beta3*W[i]
}
# Prior models for beta0,beta1,beta2 and beta3
beta0 ~ dnorm(43.73782 , 211.2847 ^2)
beta1 ~ dnorm(0, 1)
beta2 ~ dnorm(0, 26.62074 ^2)
beta3 ~ dnorm(0,41.33789 ^2)
}
```

```
> poisson_bayes_model
Inference for Bugs model at "poisson_model.txt", fit using jags,
 3 chains, each with 20000 iterations (first 10000 discarded), n.thin = 10
n.sims = 3000 iterations saved
      mu.vect    sd.vect      2.5%      25%      50%      75%    97.5% Rhat n.eff
beta0  4.21630e+01 5.00000e-03 4.21540e+01 4.21600e+01 4.21630e+01 4.21670e+01 4.217300e+01 1.001 2600
beta1 -2.925500e+01 1.00000e-02 -2.927400e+01 -2.926200e+01 -2.92550e+01 -2.92490e+01 -2.923600e+01 1.001 3000
beta2 -1.140000e-01 0.000000e+00 -1.140000e-01 -1.140000e-01 -1.140000e-01 -1.140000e-01 -1.140000e-01 1.001 3000
beta3  6.000000e-03 0.000000e+00 6.000000e-03 6.000000e-03 6.000000e-03 6.000000e-03 6.000000e-03 1.001 3000
deviance 7.673857e+19 4.161513e+17 7.592354e+19 7.645106e+19 7.67411e+19 7.70239e+19 7.755811e+19 1.001 3000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 8.661411e+34 and DIC = 8.661411e+34
DIC is an estimate of expected predictive error (lower deviance is better).
```

The parameter setting was evaluated individually and was approximated in most cases to a normal distribution for simplicity, setting individual standard deviation parameters.



# Conclusions

- Categorical variables in the crashes characterization indicate not only a Poisson but a multinomial logistic regression model could 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 poisson model with response distribution  $Y_i \sim P(\lambda)$ , link function  $\log(\lambda_i) \sim \eta_i$  and linear predictor  $\eta_i \sim \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 W_i$ . Though not perfect, constitutes a sufficient model with a good acceptance ratio, a robust lot posterior to step size and tree depth, rhat values are all below 1.01, an effective sample size above 0.5 and a small MCMC error to posterior sd.
- Bayesian prior  $\beta_0 \sim N(44, 211^2)$ ,  $\beta_1 \sim N(0, 1)$ ,  $\beta_2 \sim N(0, 27^2)$  and  $\beta_3 \sim N(0, 41^2)$  although drawn from the variables distributions could be further refined with beta distributions.
- By looking at the results it can be seen that the number of crashes holds an inverse relation with the categorical variable that indicates whether the road surface is sealed or not. The same occurs with the speed limit. While for the number of highways in a polygon, there is direct relationship that seems to indicate crashes are often linked to primary roads instead of secondary or tertiary roads.

# Thank you!

 [@linkedin.com/in/lina-berbesi](https://www.linkedin.com/in/lina-berbesi)  
 [lina.berbesi@gmail.com](mailto:lina.berbesi@gmail.com)