POLITECNICO
MILANO 1863

School of Industrial and Information Engineering

Spatial point pattern analysis for lansdlide susceptibility in Hong Kong

Bayesian Statistics

Authors: **Niccolò Donadini, Aygalic Jara–Mikolajczak, Samulele Marchioni, Nathaniel Nethercott, Roberta Troilo, Christopher Volpi**

Student ID's: 10622527 (Donadini), 10797403 (Jara–Mikolajczak),10619375 (Marchioni), 10815538 (Nethercott), 10659317 (Troilo), 10615393 (Volpi)
Advisors: Alessandra Guglielmi, Matteo Gianella
Academic Year: 2022-23

Contents

1 Background

Landslides have always been one of the most common natural hazards in Hong Kong. Generally caused by seasonal rainstorms, landslides have claimed many lives and destroyed homes. Although the Government has strived to tackle slope safety problem, it is obvious that landslide risk cannot be totally eliminated. On average, about 300 landslides are reported to the Geotechnical Engineering Office (GEO) each year [1].

A first approach by the GEO to reduce the risk of landslides was developed in the 1970s. At the time, a focus on structural slope stabilisation was the most widely used approach to reduce landslide risk. Hong Kong instead set up the world's first regional landslide warning system: the landslip warning system. The warning was raised during heavy periods of rainfall to alert the public of potential landslide danger and broadcast across radio and TV news stations at regular intervals. In 2005 the GEO also introduced the landslide potential index, a statistical model used to warn when there is the highest potential for fatal and severe landslide incidents. Unlike the landslip warning, which is issued ahead of rainstorms, the landslide potential index estimates the risk of landslides directly after major rainstorms have ended.

In addition to having acted to reduce the risk of landslides happening in the first place, geotechnical engineers in Hong Kong have also compiled a comprehensive landslide catalogue with information about precise location (geographic coordinates), geometry, geology and formation history.

Having data similar to these at our disposal as obtained from government of Hong Kong's public access datasets, we would like to use point process models to propose a statistical tool capable of understanding the distribution of landslide locations based on spatial covariates such as geometry and geology of the terrain where the landslide happened. Furthermore, we can enhance our model by taking into consideration the spatial correlation between events.

This tool could be used, along side the two current systems, to identify areas of Hong Kong which are more hazardous based on the data pattern of landslides occurred in past eras. Understanding the relationship between geographic features and the propensity for landslides could further allow for more effective and informed government spending to implement relevant infrastructures to reinforce identified at-risk regions and consequently save lives in the process.

2 Formulation

Point pattern analysis of marked data processes can allow for practitioners to uncover and interpret key features driving the process realizations. In a geostatistical setting this is relevant as it allows one to explain spatially-continuous phenomena through discrete pattern realizations and the knowledge of the geographic covariates.

Following the notation of Banerjee et al.[2], we define the spatial region of interest by $D \subset \mathbb{R}^2$ and the random variables corresponding to realizations of our process by $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$ with each $\mathbf{s}_i \in D$. Note that the number of realizations, $n \in \mathbb{N}$, itself is also a random quantity. We define by $\mathbf{S} \subset D$ the shorthand form for the vector of our point process realization over D.

Though in its theoretical formulation D should be a connected set, the geography of Hong Kong violates this condition with Hong Kong Island and Lantau Island (as well as the general archipelago along the coast) representing disjoint, non-empty subsets of D. In practice, however, one would be hard-pressed to extend this form of geostatistical analysis on a country-wide scale given the inherent irregularities in nation geographies.

2.1 Idealized model

We assume the point pattern realization is driven by some multivariate location density f defined over D, and that this density has the property of symmetry in its arguments, since points comprising the point

pattern $\mathbf{S}$ are unordered. With respect to this "unorderdness" of the $\mathbf{s}_i$'s the likelihood for $\mathbf{S}$ takes the form

$$\mathcal{L}(\mathbf{S}) = n! f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) \cdot P(N(D) = n) \tag{1}$$

where $N(D) : \mathbb{R}^2 \to \mathbb{N}$ is a counting measure describing the total number of observations in D. As is done customarily [2], we represent $N(D)$ through a Poisson process driven by an underlying *intensity function* $\lambda(\mathbf{s}) : D \to \mathbb{R}$. We will later elaborate on λ as a GLM dependent on the spatial covariates of D. In this form we have:

$$N(D) \sim Po(\lambda(D)), \quad \lambda(D) = \int_D \lambda(s) ds$$

In general, for any $B \subseteq D$ we have that $N(B) \sim Po(\lambda(B))$, and hence that $\mathbb{E}[N(B)] = \lambda(B) = \int_B \lambda(\mathbf{s}) ds$. Thus, knowledge of the intensity surface - provided it has been fit to adequately represent $\mathbf{S}$ - allows for the straightforward estimation of phenomena occurrences in any arbitrary region of the domain in the mean sense through numerical integration of the intensity function. For tiled surfaces the computational complexity of this problem diminishes greatly.

The independence of the Poisson process on disjoint sets immediately implies that the location density can be written as:

$$f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) = \prod_i^n f(\mathbf{s}_i) = \prod_i^n \frac{\lambda(\mathbf{s}_i)}{\lambda(D)}$$

so that the intensity function totally defines the location density. It is immediate to see that in this form $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$ is a density on D.

Now, bearing in mind the expression of $\mathcal{L}(\mathbf{S})$ in 1, we obtain

$$\begin{aligned}
\mathcal{L}(\mathbf{S}) &= n! f(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n) \cdot P(N(D) = n) \\
&= n! \prod_i^n \frac{\lambda(\mathbf{s}_i)}{\lambda(D)} \cdot \frac{\lambda(D)^n \exp(-\lambda(D))}{n!} \\
&= \left(\prod_i^n \lambda(\mathbf{s}_i) \right) \exp(-\lambda(D))
\end{aligned}$$

Currently, we're still only considering one form of randomness in the modelling - that being the number of points we expect to see when the intensity function is fixed. When there is no additional source of randomness in the intensity function, we're dealing with either a homogeneous Poisson process (HPP), or a non-homogeneous Poisson process (NHPP); the distinction between the two being strict stationarity of the intensity function in the former [3].

To benefit from insight obtained by applying a Bayesian framework to the problem at hand - namely quantifying the effect of the geostatistical covariates on a region's propensity for landslides - a good first step is to move in the direction of the Cox model and to consider a GLM formulation using features defined continuously over D. The general form for the intensity function then becomes

$$\lambda(s) = g(X(s)^T \beta) \lambda_0(s)$$

where $X(s) \in \mathbb{R}^{k+1}$ represents the collection of spatial covariates with the constant 1 prepended to account for the model intercept, $\beta \in \mathbb{R}^{k+1}$ is the vector of regression coefficients, and $\lambda_0(s)$ is an error process with mean 1. As is done in practice, and also given it is the natural link function for the Poisson distribution, we consider the case where $g(\cdot) = \exp(\cdot)$ so that

$$\log \lambda(s) = X(s)^T \beta + \log \lambda_0(s) \tag{2}$$

Finally, letting $w(s) = \log \lambda_0(s)$ we get the familiar form

$$\log \lambda(s) = X(s)^T \beta + w(s) \tag{3}$$

add a comment about how mean of lambda = 1 implies mean of w needs to be ...

The form of 3 is quite attractive from a modelling perspective, in the case where appropriate choices of $w(s)$ can capture spatial correlation of observations. That being said, in this form we face a few challenges computationally. Indeed, in the computation of the likelihood of the point pattern, $\mathcal{L}(s)$, we need to estimate the integrated quantity $\lambda(D) = \int_D \exp(X(s)^T\beta + w(s))ds$.

Given the discrete nature and oftentimes low resolution of geostatistical measurements this integral would instead reduce to something resembling

$$\lambda(D) = \int_D \exp(X(s)^T\beta + w(s))ds = \sum_i^M \int_{B_i} \exp(X(s)^T\beta + w(s))ds \approx \sum_i^M \exp(X(\mathbf{s}^*)^T\beta) \int_{B_i} \exp(w(s))ds$$

where $B_1, B_2, ..., B_M$ is a partition of D sufficiently refined enough such that $X(s) = c_i$ for $\forall s \in B_i$ according to the available measurements. The final integral would then be approximated straightforwardly through basic Monte Carlo.

Returning back to the case of spatial correlation we can make tangible the notion of neighbouring influences by assuming the intensity surface is a Gaussian process at the log scale with expectation given by the regression component. This framework is referred to as the Log Gaussian Cox Process (LGPC), and with it our model becomes

$$\begin{aligned} s|\lambda &\sim \lambda(s)/\lambda(D) \\ \log\lambda(\mathbf{S})|\beta,\sigma^2 &\sim \mathcal{N}_{N(D)}(X(\mathbf{S})\beta, \sigma^2\rho(\cdot;\phi)) \\ N(D)|\lambda &\sim Po(\lambda(D)) \\ \beta|\nu^2 &\sim \mathcal{N}(\mathbf{0}, \nu^2 I_{k+1}) \\ (\sigma^2,\phi,\nu^2) &\sim \pi \text{ informative} \end{aligned}$$

2.2 Pragmatic Model

The model we've developed until this point captures nicely the description of point processes defined at a continuous resolution, however, as we've alluded to earlier, geostatistical data is oftentimes collected on much courser scales. Furthermore, given that hundreds of landslides are recorded in a given year, fitting the Stan model becomes computationally intractable due to both the need to sample at each iteration from a multivariate normal for the spatial noise as well as to numerically integrate the intensity surface realization in order to estimate $\lambda(D)$. We thus seek an analogous discretized version of our modelling goal.

In this setting we impose an overlaid grid into which landslide observations are aggregated on a cell-wise basis and consider the geographical covariates to be effectively constant in each of the cells. Choosing and justifying the resolution of the grid is of great importance since without proper provisioning the aggregated covariates may no longer be informative in the predictive problem. The trade-off in the modelling challenge becomes choosing cell dimensions such that they are not too coarse as to lump together geographically diverse regions, but also not too fine so that we're computationally burdened.

Letting s_i^* $i \in 1, 2, ..., m * n$ denote the centroid of the i-th grid cell of a $m \times n$ grid with $m, n \in \mathbb{N}_{>0}$,

we have in this new framework that

$$\begin{aligned} Y(s_i^*)|\lambda &\sim Po(\lambda(s_i^*)) \\ \log\lambda(s_i^*)|\beta, w, \epsilon &= X(s_i^*)\beta + w(s_i^*) + \epsilon(s_i^*) \\ \beta|\nu^2 &\sim \mathcal{N}(\mathbf{0}, \nu^2 I_{k+1}) \\ \epsilon(s_1^*), \epsilon(s_2^*), ...\epsilon(s_{m*n}^*)|\tau^2 &\overset{iid}{\sim} \mathcal{N}(0, \tau^2) \\ \mathbf{w}|\sigma^2, \phi &\sim \mathcal{N}_{m*n}(\mathbf{0}, \sigma^2\rho(\cdot;\phi)) \\ (\nu^2, \tau^2, \sigma^2, \phi) &\sim \pi \text{ informative} \end{aligned}$$

with $Y(s_i^*)$ denoting the number of observed landslides in cell i. Generally we take all the noise variances to be distributed inverse-gamma or half-Cauchy with appropriate specifications. From here on let $\boldsymbol{\theta} = (\nu^2, \tau^2, \sigma^2, \phi)$.

At this point we are again faced with the decision of selecting the type of covariance kernel for the spatial random effects, $\{w(s_i)^*\}_i$, in the model with the available choices falling between using an isotropic kernel for the weakly stationary spatial effects such as exponential or Matérn, or proceeding in the direction of a CAR formulation and using the proximity matrix to drive the spatial effects. Indeed, aggregating spatial data into areal observations provides a bit of flexibility since by retaining coordinate information we have both a notion of the relative distances between cells and on the other hand the concept of neighbouring observation units.

The main advantage in selecting a CAR model for our spatial random effects lies in the sparsity of the proximity matrix which allows for the fitting of models on higher resolution grids to be executed much faster than the dense covariance matrix resulting from the exponential or Matérn kernels - though perhaps at the cost of assessing how the influence of neighbours changes as a function of the distance between them.

Following once again the direction of [3] we define the proximity matrix, $A = [a_{i,j}]$, such that $a_{i,j} = 1$ if cells i and j share an edge, and take D_A as a diagonal matrix with diagonal entry at $i = \sum_j^{m*n} a_{i,j}$ the total number of neighbours for cell i. With this in place the CAR formulation for the spatial effects is:

$$\begin{aligned} \mathbf{w}|\sigma^2 &\sim \mathcal{N}_{m*n}(0, \sigma^2\Sigma) \\ w_i|w_{-i}, \sigma^2 &\sim \mathcal{N}\left(\rho\frac{\sum_j a_{i,j}w_j}{\sum_j a_{i,j}}, \frac{\sigma^2}{\sum_j a_{i,j}}\right) \end{aligned}$$

with $\rho \in (0,1)$ to ensure non-singularity of $\Sigma^{-1} = D_A - \rho A$.

On the other hand, greater insight into the spatial effects might be gleaned from instead opting to use a parametric covariance kernel. In this case both the variance of the spatial effects as well as the correlation length of the kernel would be fit in the model. The drawback, however, we are once again working with a dense covariance matrix and sampling the spatial noise in the Stan model will be comparatively more expensive than the CAR case. Formalizing this approach we have

$$\begin{aligned} \mathbf{w}|\sigma^2, \phi &\sim N_{m*n}(\mathbf{0}, \sigma^2\rho(\cdot;\phi)) \\ \rho(s_i^*, s_j^*) &= \exp(-||s_i^* - s_j^*||/\phi) \end{aligned}$$

Finally, integrating out the spatial random effects and re-writing the expression for the intensity surface we notice that conditional on the parameters, the estimation of $\log\lambda(\mathbf{S}^*)$ can be done through kriging.

$$\log\lambda(\mathbf{S}^*)|\beta, \boldsymbol{\theta} \sim \mathcal{N}_{m*n}(X(\mathbf{S}^*)\beta, \sigma^2\Sigma + \tau^2 I_{m*n}) \tag{4}$$

so that the posterior predictive density for estimating the number of landslides at a new location, s_{new}^*, becomes

$$P(Y(s_{new}^*) = k)|Y(\mathbf{S}^*) = \int P(Y(s_{new}^*) = k|\log\lambda(s_{new}^*))P(\log\lambda(s_{new}^*)|\beta, \boldsymbol{\theta}, Y(\mathbf{S}^*))P(\beta, \boldsymbol{\theta}|Y(\mathbf{S}^*))d\beta d\boldsymbol{\theta} \tag{5}$$

where Σ in 4 can either be generated from the covariance kernel or be as specified in the CAR model. In any case, the posterior predictive problem for estimating the expected number of observations in a new cell can be addressed using Monte Carlo estimation and log Gaussian kriging.

3 Implementation

3.1 Dataset and Preliminaries

Our dataset is composed of the recordings of 111 408 landslides occurred in Hong Kong over the last 100 years. We are provided with information about where the landslide occurred and the geological features related to the event. In particular, our features are:

Covariate	**Description**
SLIDE TYPE	categorical variable classifying the observations as relict landslide("R"), landslide with recent channelized debris flow ("C"), landslide with open hillslope ("O"), recent coastal landslide ("S")
M_WIDTH	width of the landslide main scarp (in metres)
S_LENGTH	length of the landslide source area (in metres)
SLOPE	slope: ground slope angle across the landslide head
COVER	cover: categorical variable classifying the landslides based on the vegetation cover. Values fall into the categories of totally bare of vegetation ("A"), partially bare of vegetation ("B"), completely covered by grass ("C"), covered in shrubs and/or trees ("D")
YEAR_1	year of the serial photograph on which the landslide was first observed
HEADELEV	elevation of the landslide's crown (in mPD, i.e. metres above principal datum)
TAILELEV	elevation of landslide toe (in mPD)
ELE_DIFF	elevation difference between landslide crown and toe in metre
GULLY	categorical variable classifying landslides belonging to a previously recorded area of gully erosion ("Y") and landslides belonging outside of this area ("N")
NORTHING & EASTING	northing and easting coordinates of the landslides

Table 1: Data dictionary for Hong Kong landslide data set

To begin our point pattern analysis, we reduced our training samples considering only the landslides recorded from 2000 onwards and we ended up with a pattern of landslides corresponding to the one in the following figure.

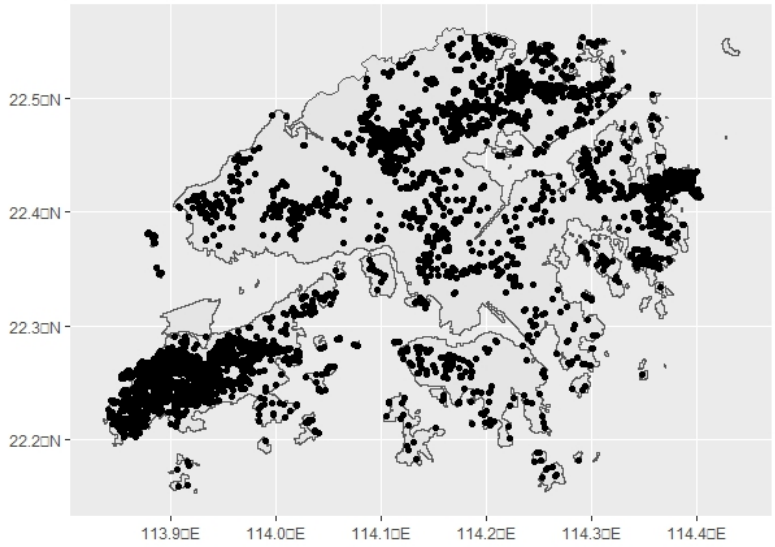


Figure 1: Landslides occurrences from 2000

In order to discretize the spatial covariates over the Hong Kong area we choose a certain cell resolution and we partitioned the area into m*n cells. We then obtained covariates to feed our model with by averaging the values of all the observations in each grid cell. This is why the choice of the number of quadrants of the grid was the one of main issue of our project: how much did we want to get close to the data in order to make a good predictor but without overfitting?

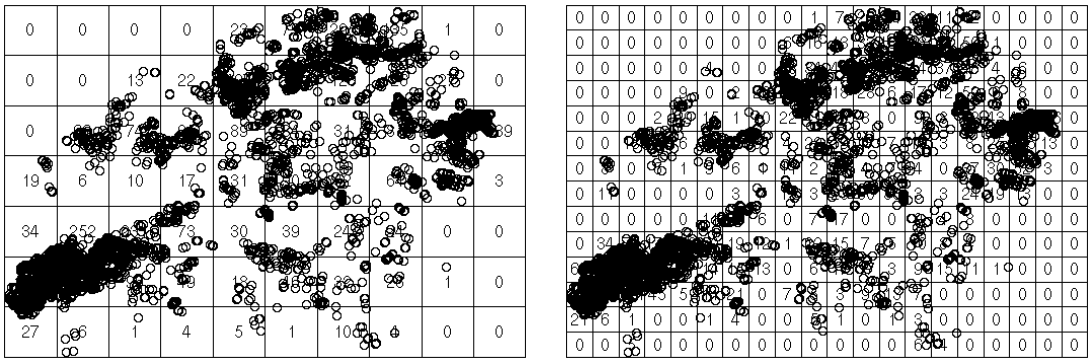
Furthermore, having a finer resolution meant much more time for the model implementation. In this sense, a good choice for us was the one to create a mask that could filter from the m*n cells only the one actually belonging to the Hong Kong area. In this way, our prediction could not include cells which were outside the nation's boundaries.

Figure 2: Visualization of the mask superimposed over the grid to extract only relevant cells for the model fitting

Grid resolution

In order to make an informed choice for the selected grid resolution we generated elbow plots for the within sum of squares (WSS) of the covariates as a function of a grid scaling hyper parameter (the number of cells in the x direction). Heuristically, as grid cell sizes decrease the underlying geography represented therein becomes more homogeneous thus the variance of the features within each cell should diminish.

Similar to hierarchical clustering techniques, we sought to identify an "elbow" in the plot which indicates a refinement level such that the intra-cell variance drops off significantly after which marginal gains are made from further cell divisions. These efforts help justify one of the central modelling assumptions that features are constant over cells, and allow us to choose a computationally feasible level of refinement for analysis. Using this approach we see that an elbow range falls somewhere between 20 and 50 units - corresponding to roughly 1-3 km in physical terms.



(a) 7x10 (b) 14x20

Figure 3: Resulting quadrant counts for different levels of grid refinements

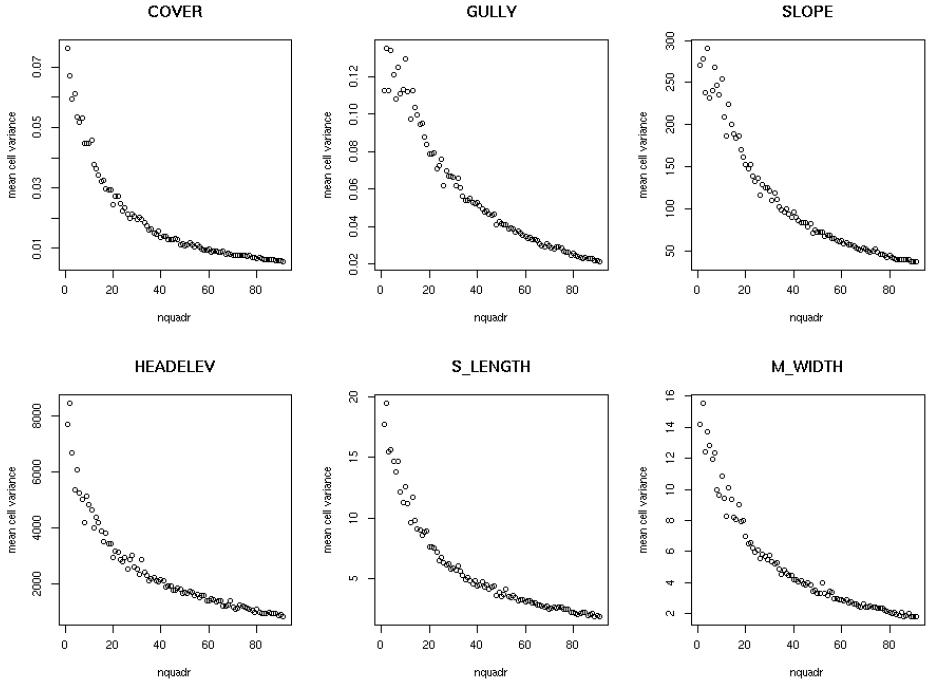


Figure 4: Generated elbow plots for the spatial covariates as a function of grid resolution size

Covariate selection

In order to identify the most significant features in determining a region's propensity for landslides, we considered a simplified model without spatially correlated noise which allowed for significantly reduced fitting times in Stan. We compared several Poisson process models fit using a grid resolution of 35x50 as informed by the previous section, adding at each new model an additional variable. This was done as

an inexpensive screening process to select influential covariates so that we could more efficiently manage computational strain in the actual model fitting later on. The formulation for the intensity used in this case is the following:

$$\begin{aligned} \log \lambda(s_i^*)|\beta, \epsilon = X(s_i^*)\beta + \epsilon(s_i^*) \\ \beta|\nu^2 \sim \mathcal{N}(\mathbf{0}, \nu^2 I_{k+1}) \\ \epsilon(s_1^*), \epsilon(s_2^*), ...\epsilon(s_n^*)|\tau^2 \overset{iid}{\sim} \mathcal{N}(0, \tau^2) \\ (\nu^2, \tau^2) \sim \pi \text{ informative} \end{aligned}$$

The numerical features which turned out to be the most significant were ELEDIFF (which represents the difference between HEADELEV and TAILELEV) and SLOPE. Nevertheless, the former refers to a feature that can be recorded only after the landslide has occurred, hence it does not make sense to consider its influence on the occurrence of landslides in the first place.

The other two features which turned out to be very significant were the categorical covariates COVER and GULLY - which we once again remind the reader refer respectively to the vegetation coverage of the area and whether or not erosion has occurred in the region previously. This can be seen in the plot of the posterior distribution of the coefficients. In order to integrate these covariates into the model we first transformed the COVER from a categorical variable to a binary one with values representing being "totally bare of vegetation" or "partially or completely covered by vegetation". When aggregating both binary features over cells we chose to assign a positive label to any cell containing at least one observation with a positive value for the feature in question.

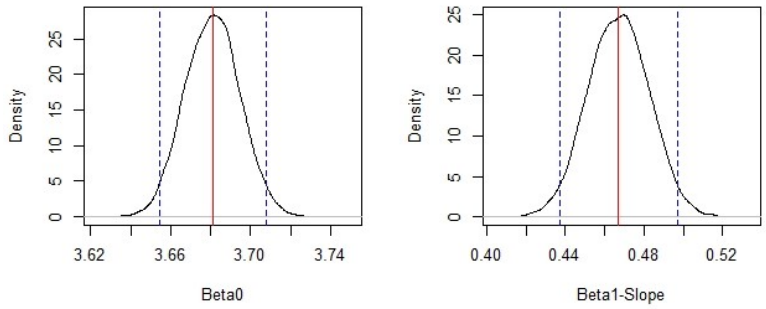


Figure 5: NHPP with slope as only spatial covariate

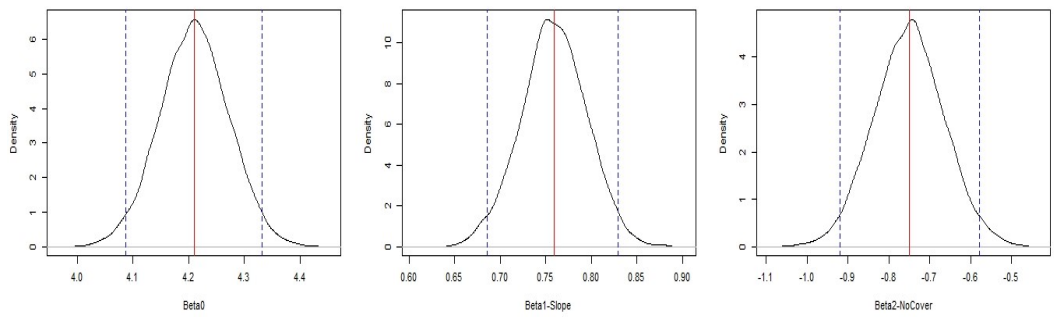


Figure 6: NHPP with slope and vegetation cover as covariates

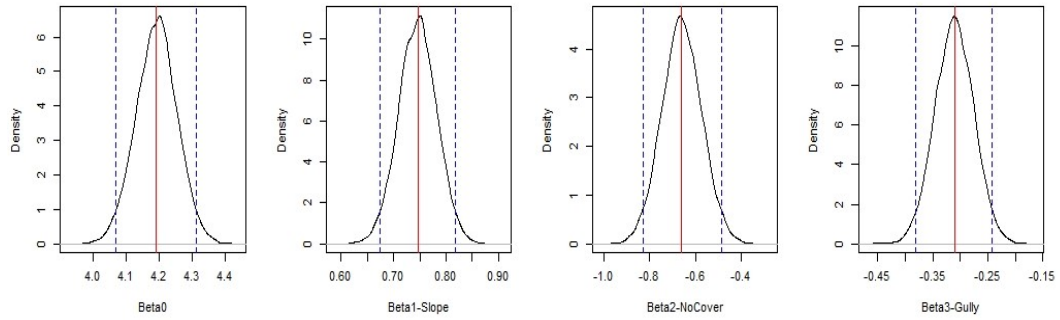


Figure 7: NHPP with slope, vegetation cover and gully erosion as covariates

As we can observe from the plots, the beta coefficients always have a distribution which does not include the value 0 on the x-axis and this made us believe that all these features would be significant. To make this intuition stronger, we used the "loo" R package to confront these models though some significant indices such as WAIC or LOO-CV. The derivation of the expected log predictive density was not as easy as expected, nor it gave us reliable results. Nevertheless, we still were able to compare the models and realize through the `loo_compare` command that the model with all the three covariates was better to explain the output.

```
        elpd_diff se_diff
model2     0.0       0.0
model1 -5376.6       0.0
```

Here is displayed the output of the `loo_compare`, where model1 is the one with only slope and model2 is the one with slope and vegetation cover.

```
        elpd_diff se_diff
model2     0.0       0.0
model1 -6141.6       0.0
```

Analogously, here is displayed the output of the `loo_compare`, where model1 is the one with slope and vegetation cover as covariates and model2 is the one with slope, vegetation cover and gully erosion as covariates. In addition, we performed a similar analysis using Bayes factor and bridge sampler of the fitted models, deriving the same conclusions.

3.2 Model Fitting

Rather than considering all the data collected from the year 2000 onward, we instead proceeded in the direction of using landslide observations from 2016, 2017, and 2018 to inform the model fitting, and reserved the samples from 2019 for prediction.

Indeed, considering 20 years worth of data is potentially problematic for a number of reasons. First, it is generally known that climate phenomena have been undergoing a drastic change over the past few decades hence considering all observations in this time would reflect process realizations under different time-dependent climate paradigms. Second, from a purely application-based point of view the model learns to recover the regression coefficients and spatial parameters to explain the observations it was fit with. Hence fitting the model on 20 years worth of data yields parameters which can be used to estimate landslide occurrences in regions for 20 years into the future. Of course assuming that each year contributes equally to the total number of landslides one can divide the predicted quantity by the time frame to estimate the number of landslides in the next year - but again it is unlikely given the variability of climate trends that this equal-contribution-of-landslides-on-an-annual-basis assumption holds on the scale of decades.

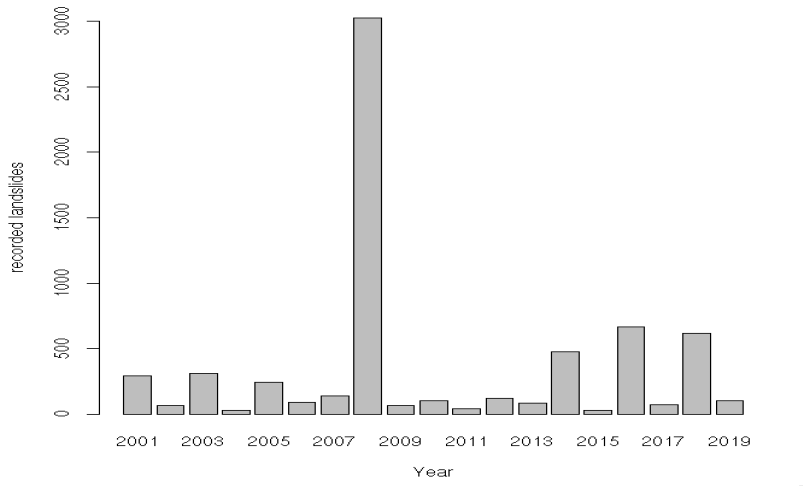


Figure 8: Distribution of recorded landslides in Hong Kong from 2000-2019. In 2008 a record breaking rainstorm struck Lantau Island resulting in over 3000 landslides [4]

Returning to the discretized model framework introduced earlier, we elaborate and formalize our model below. Additionally, we select a grid of resolution 21x30 which corresponds to grid cell dimensions of roughly 2km x 2km. Under this resolution we had 166 marked observations representing landslide occurrences in the training set, and 55 in the test set.

$$
\begin{aligned}
Y(s_i^*)|\lambda &\sim Po(\lambda(s_i^*)) \\
\log \lambda(s_i^*)|\beta, w, \epsilon &= X(s_i^*)\beta + w(s_i^*) + \epsilon(s_i^*) \\
\beta|\nu &\sim \mathcal{N}(\mathbf{0}, \nu^2 I_{k+1}) \\
\epsilon(s_1^*), \epsilon(s_2^*), ...\epsilon(s_n^*)|\tau^2 &\overset{iid}{\sim} \mathcal{N}(0, \tau^2) \\
\mathbf{w}|\sigma^2, \phi &\sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \rho(\cdot; \phi)) \\
\tau^2 &\sim inv - gamma(1, 1) \\
\nu^2 &\sim inv - gamma(1, 1) \\
\sigma^2 &\sim \mathcal{HC}(0, 1) \\
\phi &\sim inv - gamma(4, 1)
\end{aligned}
$$

with $\rho(s_i^*, s_j^*) = \exp(-||s_i^* - s_j^*||/\phi)$ the exponential covariance kernel and n the number of samples. Here we chose the initial hyper parameters for ϕ to reflect correlation at small scales as suggested by [2] and [5]. The associated Stan file for the implementation can be found in appendix A.

The covariates, $X(s^*)$, considered in the model were SLOPE, HEADELEV, COVER, and GULLY, with the latter of these being aggregated over grid cells in a consistent fashion to what was outlined before. Before model fitting the numerical features were normalized and the normalizing hyperparameters were saved for later use in the pre-processing of the test data.

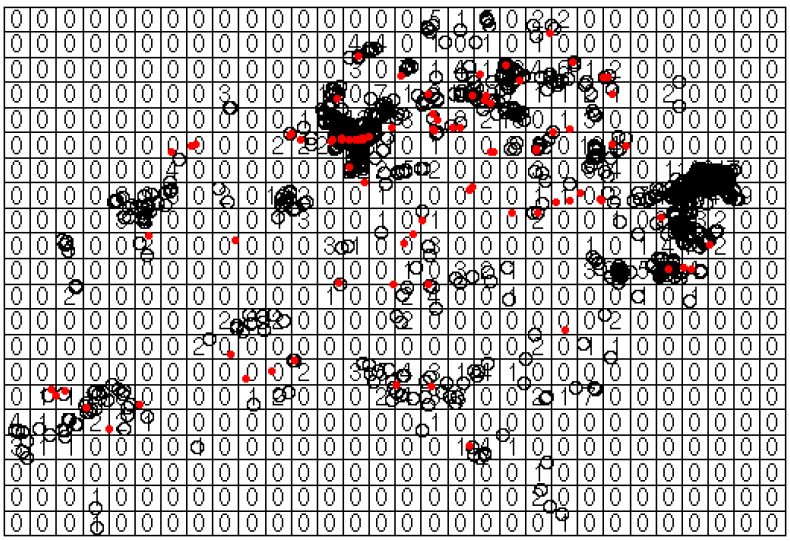


Figure 9: Landslide distribution from years 2016-2018 over the selected grid with observations from 2019 in red

We fit the model using the hyperparameter specifications for `rstan` outlined in the table below and obtained the corresponding chain histories for the parameters.

Parameter	**Value**
chains	3
warmup	1000
iter	6000
adapt_delta	0.999
max_treedepth	13

Table 2: `rstan` parameters for model fitting

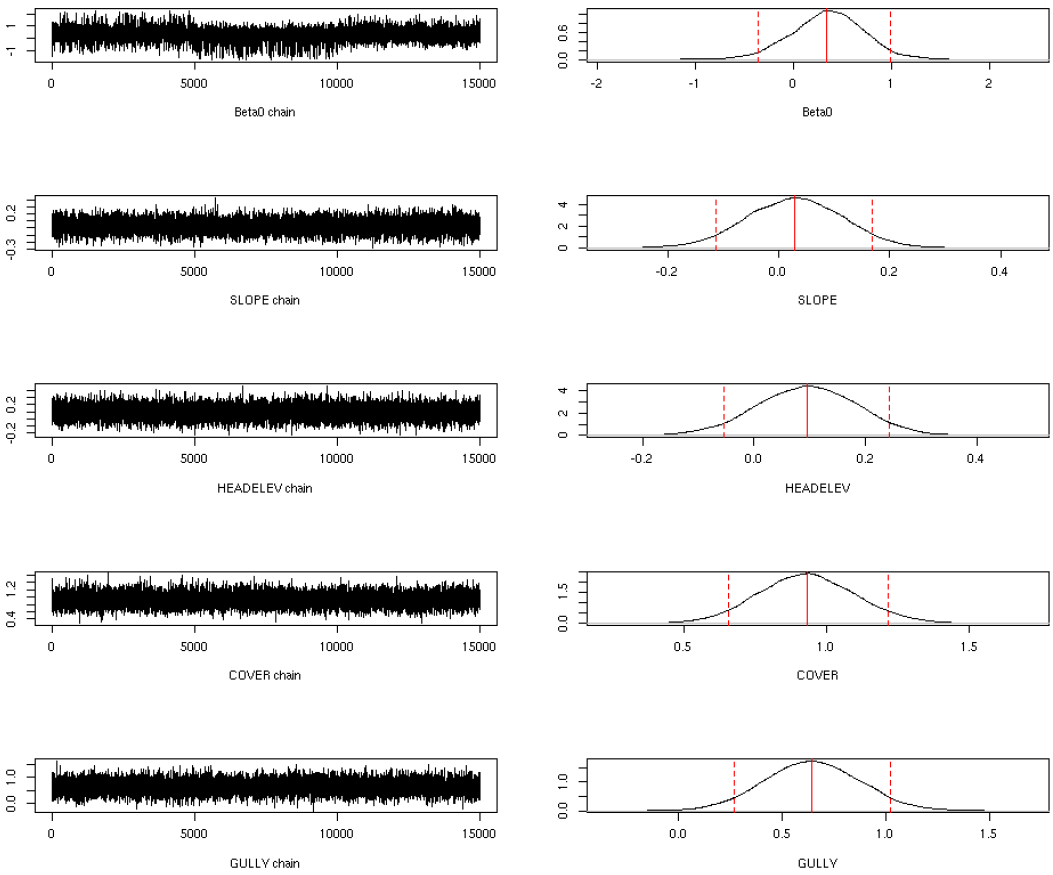


Figure 10: Posterior fit for the model regression coefficients

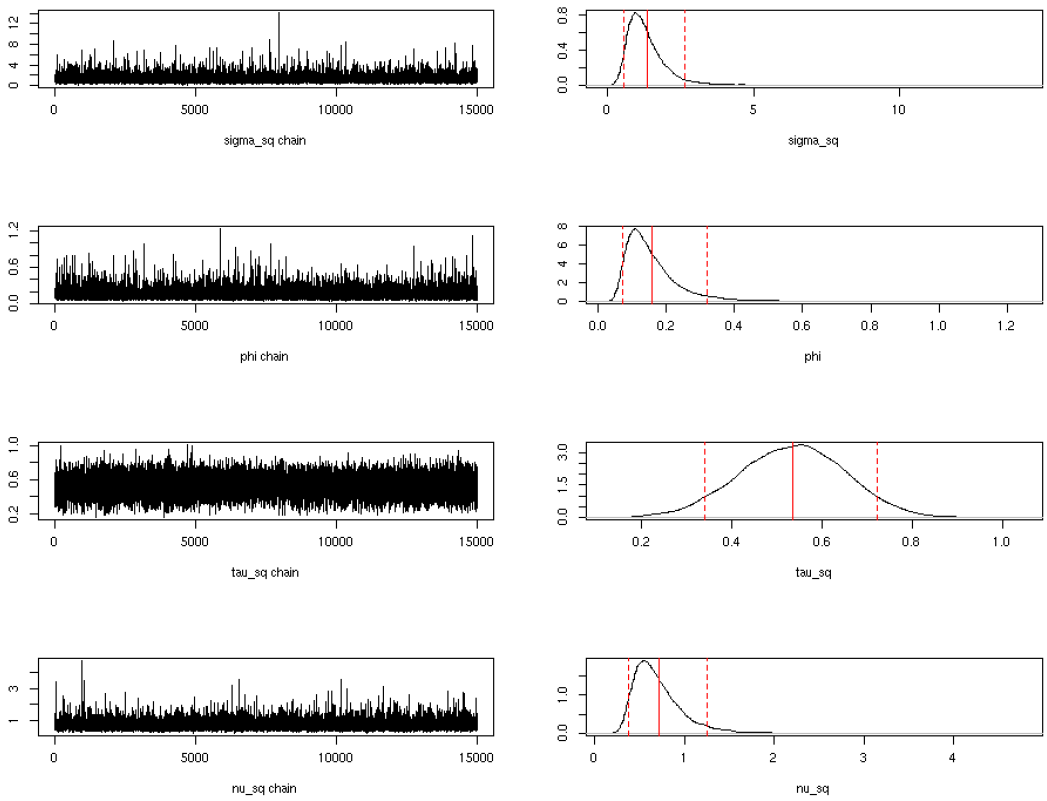


Figure 11: Posterior fit for the noise terms and scale parameter

With the posterior distributions in place from fitting on the training data, we were interested in assessing the predictive performance of the model on the test set.

As a consequence of the discretization approach grid cells corresponding to observations in the test set were not totally disjoint from those in the training set - indeed there was around a 70% overlap between the two. In the limiting case where the grid cell size shrinks to 0 we would have totally disjoint training and test sets, but again computational restrictions prevented such granularity from being implemented. To account for this we re-ran the pre-processing pipeline on the landslide data from 2019 independently from the training set so that the aggregated quantities in the cells represent the features associated with the new data. Numerical features were normalized using the mean and standard deviation extracted from the training set.

Since by construction the intensity surface evaluated over the cells is normally distributed at the log scale as is given in 4 conditionally on β and $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \nu^2)$ we benefit from being able to use gaussian regression/kriging in the estimation of the intensity surface for our new samples. Letting n represent the number of samples in the training set and m the number of samples in the test set we have

$$\begin{bmatrix} \log \lambda(\mathbf{S}^*{}_{train}) \\ \log \lambda(\mathbf{S}^*{}_{test}) \end{bmatrix} | \beta, \boldsymbol{\theta} \sim \mathcal{N}_{n+m} \left(\begin{bmatrix} X(\mathbf{S}^*{}_{train})\beta \\ X(\mathbf{S}^*{}_{test})\beta \end{bmatrix}, \begin{bmatrix} \Sigma_{train,train} & \Sigma_{train,test} \\ \Sigma_{test,train} & \Sigma_{test,test} \end{bmatrix} \right)$$

where Σ's are generated from $\sigma^2 \rho(\cdot; \phi) + \tau^2 I_{n+m}$. Now provided the knowledge that $\log \lambda(\mathbf{S}^*{}_{train}) = Z$, we get that

$$\log \lambda(\mathbf{S}^*{}_{test}) | \beta, \boldsymbol{\theta}, \log \lambda(\mathbf{S}^*{}_{train}) = Z \sim \mathcal{N}_m \left(\bar{\mu}, \bar{\Sigma} \right)$$

with

$$\bar{\mu} = X(\mathbf{S}^*_{test})\beta + \Sigma_{test,train} \cdot \Sigma^{-1}_{train,train}\,(Z - X(\mathbf{S}^*_{train})\beta)$$
$$\bar{\Sigma} = \Sigma_{test,test} - \Sigma_{test,train}\Sigma^{-1}_{train,train}\Sigma_{train,test}$$

Using the chain history extracted from the model fitting for the parameters we then estimate the expected number of landslides in each of the locations in our training set through Monte Carlo approximation consistent with 5. The posterior predictions at each location are plotted below with the associated 90% confidence interval denoted by the dashed red lines. Making the corresponding Bonferonni correction with our desired confidence of $\alpha = 0.1$ and regenerating the adjusted quantiles we end up predicting with that anywhere from 0 to 1603 landslides will occur in the year 2019 with probability $1 - \alpha$. Though historically-speaking this upper estimate is possible given the 2008 data, this predictive range reflects significant uncertainty in the model.

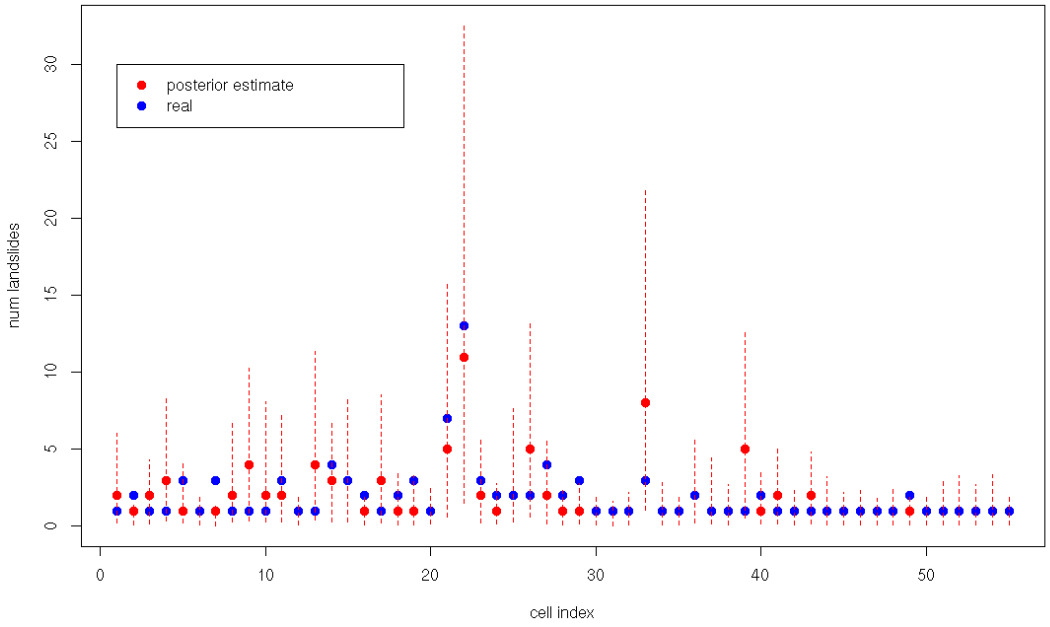


Figure 12: Gaussian regression predictions for the reserved 2019 landslides data

4 Limitations

Though the modelling here-presented was of sufficient complexity to uncover a relationship between geographic covariates as well as the spatial correlation between landslide sites, it failed to take into account the influence of precipitation which is a well-known landslide catalyst [4]. Indeed, since rainfall data were not included as a feature in the provided data set, we attempted to integrate external data sets to recover this information, but were unable to successfully do so since the government of Hong Kong did not have this data readily available. Though it was in theory possible to combine the over 50 csv's included on the Hong Kong public database repository concerning rainfall data at independent observation sites, this direction was unattractive given the lack of uniform coverage of these observatories over the country as well as the fact that they only provided data from the past year. That being said, including rainfall and other meteorological and geological features would certainly better inform the model.

Additionally, comments should be made concerning the approach taken to carry out the predictions for the 2019 landslide data. As was previously mentioned, we only considered predictions over the spatial regions where we knew beforehand landslides would occur. Since the model parameters were fit using strictly samples from observed landslide sites, the model was able to explain why some regions experienced more landslides

than others, **but** could not faithfully predict over the entirety of Hong Kong. The choice to proceed in this direction was based on the lack of descriptive data concerning the landslide phenomena and nation-wide meteorological data as a whole. A more informed data set and finer grid cell resolution would be needed in order to produce a predictive model capable of explaining in a detailed way the complexity of the landslide point patterns.

References

[1] *Seismicity of hong kong - cedd.gov.hk*, Nov. 2022. [Online]. Available: https://www.cedd.gov.hk/filemanager/eng/content_454/IN_2022_21E.pdf.

[2] G. Banerjee Carlin, *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2014.

[3] A. Guglielmi, *An introduction to bayesian models for spatial data*, 2022.

[4] *Slope safety in hong kong*, Sep. 2021. [Online]. Available: https://hkss.cedd.gov.hk/hkss/en/slope-safety-in-hong-kong/slope-safety-system/challenge-brought-about-by-climate-change/index.html#:~:text=As%20for%20Hong%20Kong%2C%20in,and%20havoc%20in%20western%20Lantau..

[5] B. Cretois, *Fitting point process models in stan - part 2*, Dec. 2020. [Online]. Available: https://benjamincretois.netlify.app/post/fitting-point-process-models-in-stan-part-2/.

A Stan file

```
functions{
    matrix GP(matrix D, real sigma_sq, real scale, real delta) {
        int N = dims(D)[1];
        matrix[N, N] K;
        for (i in 1:(N-1)) {
          K[i, i] = sigma_sq + delta;
          for (j in (i + 1):N) {
            K[i, j] = sigma_sq * exp(- D[i,j] / scale );
            K[j, i] = K[i, j];
          }
        }
        K[N, N] = sigma_sq + delta;
        return K;
    }
}
data{
  int<lower = 1> N;
  int<lower = 1> p;
  matrix[N,p] X; //covariates
  matrix[N,N] DMat;
  int<lower = 0> y[N];
}
parameters{
  //regression coefficients
  vector[p] beta;
  real<lower=0> nu_sq;

  //independent noise
  vector[N] eps;
  real<lower=0> tau_sq;

  //spatial noise
  vector[N] w;
  real<lower=0> sigma_sq;
  real<lower=0> scale;
}

transformed parameters {
    vector[N] mu;
    for(i in 1:N) {
      mu[i] = exp(row(X, i) * beta + w[i] + eps[i]); //link
    }
}

model{
  //spatial noise
  matrix[N,N] SIGMA;
  SIGMA = GP(DMat, sigma_sq, scale, 0.01);
  w ~ multi_normal(rep_vector(0,N), SIGMA);

  for (s in 1:N) {
        y[s] ~ poisson(mu[s]);
```

```
  }

  //independent noise
  eps ~ normal(0, tau_sq);

  //priors for the coefficients
  beta ~ normal(0, nu_sq);


  target += cauchy_lpdf(sigma_sq | 0, 1); //tutorial - HC noise
  target += inv_gamma_lpdf(scale | 4, 1); //tutorial
  target += inv_gamma_lpdf(tau_sq | 1, 1);
  target += inv_gamma_lpdf(nu_sq | 1, 1);
}
generated quantities {
  }
```