

# Flexible hierarchical risk modeling for large insurance data via NumPyro

Christopher Krapu and Mark Borsuk  
 Department of Civil & Environmental Engineering  
 Duke University

November 2023

## Abstract

Data analysis and individual policy-level modeling for insurance involves handling large data sets with strong spatiotemporal correlations, non-Gaussian distributions, and complex hierarchical structures. In this research, we demonstrate that by utilizing gradient-based Markov chain Monte Carlo (MCMC) techniques accelerated by graphics processing units, the trade-off between complex model structures and scalability for inference is overcome at the million-record size. By implementing our model in NumPyro, we leverage its built-in MCMC capabilities to fit a model with multiple sophisticated components such as latent conditional autoregression and spline-based exposure adjustment, achieving an 8.8x speedup compared to CPU-based implementations. We apply this model to a case study of 2.6 million individual policy-level claim count records for automobile insurance from Brazil in 2011. We illustrate how this modeling approach significantly advances current risk assessment processes for numerous, closely related outcomes. The code and data are available at <https://github.com/ckrapu/bayes-at-scale>.

## 1 Introduction

Modeling for risk assessment and prediction in insurance is challenging due to the tension between (1) the desire for accurate forecasts of expected claim counts and/or severity, and (2) the requirement for reliable uncertainty quantification. As total policy value typically exceeds available liquid assets for most types of modern insurance, a key objective in this setting is to forecast the probability of ruinous losses under a chosen ratemaking and loss adjustment scheme. Unfortunately, this probability does not allow a straightforward factorization into independent distribution functions because of strong correlations across policies stemming from similarities in geography, demographics, property type, and other factors. Consequently, an effective modeling workflow must account for as many of these correlation sources as possible. A vitally important output of such statistical modeling is the predictive covariance matrix which, for new data points, quantifies their cross-correlation structure. Although a machine learning-based approach can utilize large data sets for accurate record-level predictions, which may be highly useful operationally [Ding et al., 2020], it is uncertain whether any current research for calibrated uncertainty quantification from machine learning [Kuleshov et al., 2018, Angelopoulos and Bates, 2022] applies to the multivariate context induced by multiple policies or outcomes from the same policy. Moreover, it is often desirable for models to be useful in both small- and large-data situations such as in [Zhang, 2017], and to be structured in a way that allows for some interpretability. These requirements are not yet met by the machine learning models frequently used in insurance modeling [Lupton, 2022].

## Parameter estimation and uncertainty quantification

With regard to the statistical approach, in scenarios where the data generating process, synonymous with *forward model*, can be written down easily, probabilistic inference might not be straightforward or computationally feasible for a maximum likelihood or restricted maximum likelihood approach (see [Wahl et al., 2022] for a recent example). Generally, these methods will require a bespoke algorithm for model fitting and producing parameter estimates, restricting their usability to experts in statistical subfields. Packages designed for a broader range of probabilistic models [Lunn et al., 2009, Carpenter et al., 2017, Salvatier et al., 2016] using sampling-based Monte Carlo inference methods like Markov chain Monte Carlo (MCMC) [Hastings, 1970, Brooks, 1998, Geyer, 1992] have found widespread acceptance. A major challenge of MCMC methods,

including Gibbs sampling and the Metropolis-Hastings algorithm, is the substantial time complexity of  $\mathcal{O}(d^2)$  where  $d$  denotes the dimension of the target density, generally equivalent to the combined number of latent variables and/or parameters in a model. This quadratic scaling arises from a general  $\mathcal{O}(d)$  time to evaluate the model log posterior, scaling with the number of parameters, and a theoretical  $\mathcal{O}(d)$  mixing time [Roberts and Rosenthal, 1998, Beskos and Stuart, 2009] for a broad class of distributions. Taken together, these lead to the  $\mathcal{O}(d^2)$  scaling. Alternatives to MCMC for Bayesian model fitting include stochastic variational inference [Hoffman et al., 2013] and the integrated nested Laplace approximation reviewed [Rue et al., 2016] and benchmarked against MALA in [Taylor and Diggle, 2014]. A more comprehensive review of approximate Bayes methods is given in [Martin et al., 2023].

However, we now have MCMC methods that incorporate proposal generation using gradients of the log-posterior with regard to latent variables and parameters [Duane et al., 1987, Roberts and Tweedie, 1996, Neal, 2011]. For certain distributions, these methods can have  $\mathcal{O}(d^{4/3})$  or even  $\mathcal{O}(d^{5/4})$  time complexity for the Metropolis-adjusted Langevin algorithm (MALA) and Hamiltonian Monte Carlo, respectively. These algorithms are a substantial field of research and are explained thoroughly in [Betancourt, 2017]. An obstacle to the general use of these MCMC algorithms is the need for gradients of the log-posterior density. However, automatic differentiation software like Torch [Paszke et al., 2019], TensorFlow [Abadi et al., 2015], Jax [Frostig et al., 2018], and Theano [Bergstra et al., 2010] enable this and integrate with probabilistic programming frameworks like Stan [Carpenter et al., 2017], PyMC [Salvatier et al., 2016, Abril-Pla et al., 2023], Pyro [Bingham et al., 2019], and NumPyro [Phan et al., 2019] for statistical modeling. Techniques for scaling differentiation to large models in machine learning also apply to MCMC, benefiting from GPU-based parallel computing for rapid model fitting. The No-U-Turn sampler [Hoffman and Gelman, 2014] has become a general-purpose method due to few tuning parameters and diverse applicability. Initial efforts [Chin and McNulty, 2023] have exposed the insurance modeling community to a probabilistic programming approach, yet there is an unmet need in the literature to communicate the workflow’s pros and cons against the unique requirements of risk assessment and insurance modeling. In this work, we use NumPyro, recognizing several promising features that could benefit the modeling community.

To summarize, by writing models in probabilistic programming frameworks, we can achieve both flexibility and scalability. A detailed discussion of probabilistic programming, particularly in environmental statistics, is provided in [Krapu and Borsuk, 2019]. By employing general, model-agnostic inference methods, we can incorporate various model components; if we code the forward model in such a framework, we then automatically leverage GPU with minimal effort. We illustrate this by developing a log-additive spatial count regression model with complex submodel components. In the Methods section, we describe our statistical model and its intended data. The Results section presents summaries and analyses of the model’s inferential outputs, while the Discussion covers limitations and further improvement opportunities.

## 2 Methods

To provide context to the material from the previous section, we designed a case study to understand the capabilities and constraints of the proposed approaches. This study involves an analysis of 2.6 million auto insurance policies from *SUSEP*, the Brazilian federal insurance supervisory agency. The records, originally compiled for the supplemental material of the book *Computational Actuarial Science with R* [Charpentier, 2015, Dutang and Charpentier, 2022], detail individual auto policies from 2011, which may include claims. Our focus, however, is solely on the collision data. Each record is linked to a municipality name and contains the automobile brand (e.g., Volkswagen, Kia, Iveco) and the model (e.g., Taurus, Camry, F-150). We preserved the original brand categories but simplified the model classification into 15 vehicle types depicted in Figure 7. Each record also specifies the vehicle year and an exposure value.

### Model

The general form of the model considered in this work is a log-additive regression for count data, which records the number of collision claims per policy. We use a Poisson likelihood, expressing the  $i$ -th policy’s Poisson rate parameter  $\lambda_i$  as

$$y_i \sim \text{Poisson}(\lambda_i) \quad (1)$$

$$\log \lambda_i = \underbrace{\log \alpha_i}_{\text{Base exposure}} + \underbrace{g(\alpha_i)}_{\text{Exposure adjustment}} + \sum_{k=1}^P v_{x_{ik}} + u_{j[i]} + S_{t_i} \quad (2)$$

where  $\alpha_i$  denotes the base exposure,  $g(\alpha_i)$  denotes an exposure adjustment using splines,  $\sum_{k=1}^P v_{x_{ik}}$  captures contributions from policy-level categorical variables related to the vehicle's brand and category,  $u_{j[i]}$  is a city-level random effect associated with the city  $j$  of policy  $i$ , and  $S_{t_i}$  represents a Gaussian random walk over time with  $t_i$  indicating the vehicle's manufacturing year for policy  $i$ . We will provide further details on these terms below. There are 7,756 free parameters and latent variables in this model, with most arising from the latent city-level spatial effect. Figure 1 presents a visual representation of this model in plate notation.

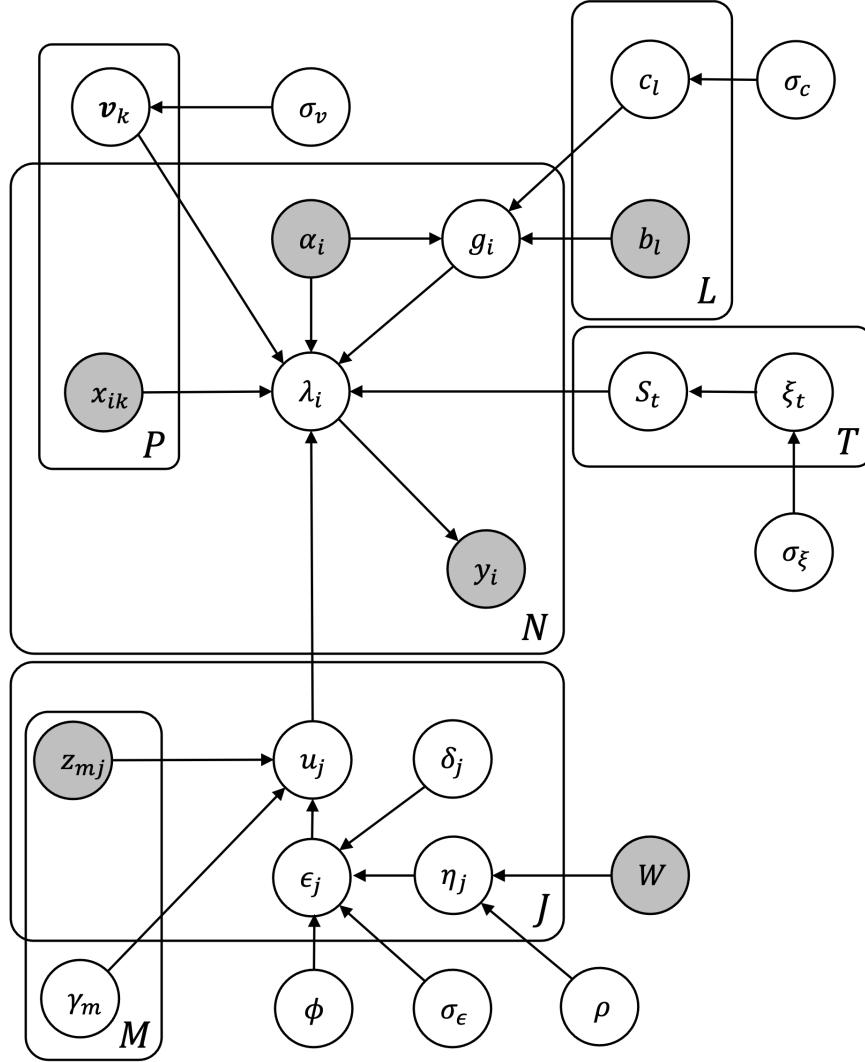


Figure 1: Graphical representation of the statistical model. Shaded circles denote observed variables, while unshaded circles denote latent variables and parameters. Rectangular plates indicate replication of the variables within the plate.

### Exposure adjustment

While the exposure score is highly correlated with observed claim counts and provides substantial predictive power, early analysis indicated a monotonic relationship between log exposure and log claim count that strongly deviated from linearity, particularly at the higher percentiles of the exposure distribution. Considering the data includes various claim types beyond collision, the exposure score likely represents a broader risk measure, rather than being finely tuned for collision claims alone. To adjust this exposure score, we modeled an additive effect on log exposure  $\log \alpha_i$  using basis splines to construct a nonlinear function of

$\log \alpha_i$ . The nonlinear function  $g(\alpha_i)$  is defined as:

$$\sigma_g \sim \text{HalfNormal}(1) \quad (3)$$

$$c_l \stackrel{\text{iid}}{\sim} N(0, \sigma_g) \quad (4)$$

$$g(\alpha_i) = \sum_{l=1}^L b_l(\log \alpha_i) \cdot c_l \quad (5)$$

where  $b_l(\cdot)$  denotes the basis spline functions,  $L$  is the number of spline coefficients, and  $c_l$  are the spline coefficients estimated by the model. We positioned the spline knots at evenly spaced percentiles of the log-transformed exposure values to ensure a uniform distribution throughout the range of the data. This technique captures subtle variations in the exposure-claim count relationship, especially in the distribution's tails. We employed the `interpolate` module from the software package `SciPy` [Virtanen et al., 2020] to determine  $b_l$ . We acknowledge that [Basile et al., 2014] combines nonlinear modeling with spatially-correlated latent variables, similarly to our approach.

### Categorical predictors

Each policy record is associated with a vehicle brand and category, represented as discrete, unordered variables. Consequently, in equation (2),  $P = 2$ , but we could include additional covariates that are continuous or discrete. Letting  $n_k$  denote the number of classes in the  $k$ -th categorical variable, we have

$$\sigma_v^{(k)} \sim \text{HalfNormal}(1) \quad (6)$$

$$v_1^{(a)}, \dots, v_{n_k}^{(k)} \stackrel{\text{iid}}{\sim} N(0, \sigma_v^{(k)}) \quad (7)$$

The interpretation of the  $v_1^{(k)}, \dots, v_{n_k}^{(k)}$  variables is that they adjust the risk according to the vehicle's brand and category, such as a motorcycle or pickup. The priors used here could be used with policy-level continuous covariates, although none are present in the current data set. For an example of a model that incorporates per-category effects for auto insurance modeling, see [Frees and Valdez, 2008]

### City effect

Most insured assets are tied to specific locations or areas on Earth, so it is common to address unmodeled variation using spatial statistical techniques. Examples include kriging to account for spatial correlation in drought severity [Paulson et al., 2010], employing spatially autocorrelated indicator variables for feature selection [Paulson et al., 2010], and using a Gaussian line process to represent hail damage Miralles et al. [2023]. For our study, policy records are categorized by cities; therefore, we treat these as *areal* data suitable for modeling with conditional autoregressions. Modeling spatial correlation at the level of individual policies is impractical because we do not have spatial coordinates more precise than the policyholder's city. As a result, we opt to model at the city level, utilizing spatial priors to account for correlations across the  $J$  components of the city effect vector  $\mathbf{u}$ , which we model as a latent conditional autoregression (CAR) [Cressie and Chan, 1989]. There are  $J = 3785$  cities and municipalities, and we organize the autoregressive prior for  $u_j$ , the effect from city  $j$ , around a modified Besag-York-Mollie (BYM) prior [Besag et al., 1991], reparameterized to enhance identifiability [Simpson et al., 2015], with city-level covariates

$$\begin{aligned} \sigma_u, \sigma_\epsilon &\sim \text{HalfNormal}(1) \\ \delta_j &\stackrel{\text{iid}}{\sim} N(0, 1) \\ u_j &= \sum_{m=1}^M \gamma_m z_{mj} + \epsilon_j \\ \phi &\sim \text{Beta}(1, 1) \\ \rho &\sim \text{Beta}(2, 2) \\ \boldsymbol{\epsilon} &= \sigma_\epsilon \left( \sqrt{1 - \phi} \boldsymbol{\delta} + \sqrt{\phi} \boldsymbol{\eta} \right) \\ \boldsymbol{\eta} &\sim N(\mathbf{0}, \mathbf{D}(\mathbf{I}_J - \rho \mathbf{D}^{-1} \mathbf{W})^{-1}) \end{aligned}$$

Here, the city effect is represented by city-level covariates  $z_{mj}$ , independent Gaussian noise  $\delta_j$ , and spatially correlated terms  $\eta_j$ .  $W$  is the spatial adjacency matrix where  $W_{ab} = 1$  if  $a$  is a neighbor of  $b$ . This matrix is created by finding the five nearest neighboring cities for each city and setting the corresponding

entries in  $D$  to 1. We then make  $W$  symmetric by ensuring that  $W_{ab} = 1$  if  $W_{ba} = 1$  for all distinct  $a$  and  $b$ . Non-neighboring entries in  $D$  remain zero. Therefore,  $D$  is a  $J$ -dimensional diagonal matrix where  $D_{jj}$  is the count of neighbors for city  $j$ , which must be at least five. For illustrations of latent Gaussian Markov random field applications, see [Jin et al., 2005; Gelfand and Vounatsou, 2003]. The parameter  $\rho$  represents the spatial autocorrelation level; the Beta(2, 2) prior promotes edge-avoiding behavior away from 0 or 1 fosters more effective MCMC convergence. The  $\phi$  parameter dictates the reliance of the city effect on uncorrelated ( $\delta_j$ ) versus correlated terms ( $\eta_j$ ) when assessing the city's influence on policyholder risk. We apply the sparse CAR likelihood in NumPyro, leading to  $\mathcal{O}(J)$  computation of  $\log p(\boldsymbol{\eta})$ , as detailed in [Jin et al., 2005].

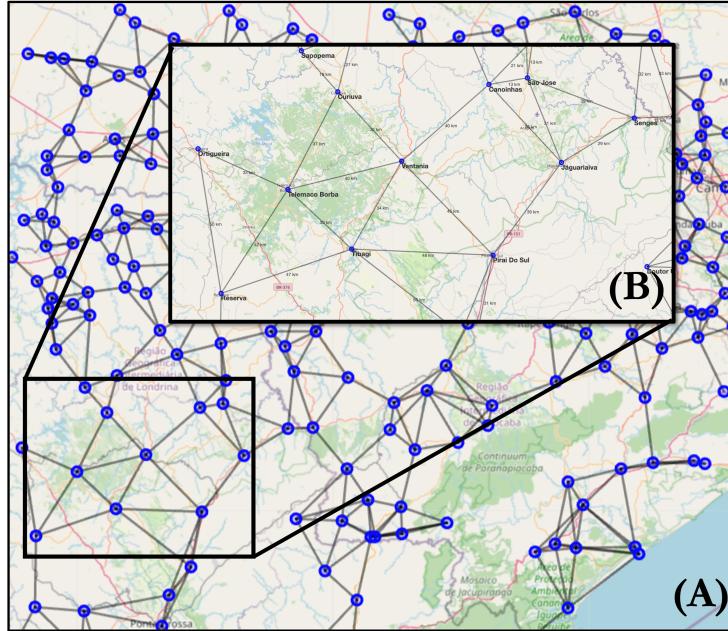


Figure 2: Adjacency network for conditional autoregressive city-level effect. Blue circles denote cities as identified by the Google Maps Geocoding API. Gray lines depict the city adjacency graph  $W$  edges established for our model's conditional autoregression. Edges are symmetric, and the number of neighboring cities can vary.

### Temporal effect

To include vehicle age in the model structure in a way that is flexible and possibly nonlinear, we created a latent time series. This series is a Gaussian random walk spanning from years 1971 to 2011, with yearly increments. The purpose of this approach is to construct a time-varying random effect [Boucher et al., 2008] that ensures modest year-to-year variations in the effect but allows for significant overall changes over the 40 years without prescribing a specific parametric form.

$$\sigma_\xi \sim \text{HalfNormal}(1)$$

$$\xi_t \stackrel{\text{iid}}{\sim} N(0, \sigma_\xi)$$

$$S_{t'} = \sum_{t < t'} \xi_t$$

The variables  $\xi_t$  represent increments in the latent time series, and  $S_{t'}$  is the corresponding cumulative sum. This model is not suited for continuous temporal coordinates; for such cases, a latent Gaussian process [Jia et al., 2023] or changepoint model [Chib, 1998] would be preferable.

### Scale priors

To fully specify a Bayesian model, we must assign prior distributions to scale or variance parameters such as  $\sigma_x i$  and  $\sigma_g$ , ensuring these priors are defined only for nonnegative values. We prefer the half-normal or folded-normal prior for these parameters as they impart minimal information on the log scale; the HalfNormal(1) prior places considerable probability mass on values ranging from 0 to 3. In terms of categorical effects, a log scale value of 3 indicates that expected claim counts are  $\exp(3) \approx 20$  times greater on the original scale of the data. This concept is elaborated in [Gelman and Carpenter, 2020]. Other priors such as the Half-Cauchy [Gelman, 2006] or horseshoe [Carvalho et al., 2009] are alternatives for stronger regularization.

All modeling components were executed in NumPyro and Jax, utilizing Jax’s alternative module for NumPy, `jax.numpy`. The following section delves into data preparation and summary statistics.

### 3 Data

To fit the model from the previous section, we used 2,658,372 records of auto insurance policies from 2011 in Brazil, as collected by SUSEP, the Brazilian insurance oversight agency, and collated as the `brvehins2` data set maintained by [Dutang and Charpentier, 2022] as supporting material for the textbook *Computational Actuarial Science in R* [Charpentier, 2015].

We removed 944 rows with vehicle manufacture dates prior to 1971 to avoid introducing extra parameters due to the time-varying effect corresponding to a small sample size. We also discarded an additional 8,346 rows that had exposure values set exactly to zero. 141,004 records lacked information on vehicle type; these were also removed from the data set. From the vehicle model information, we extracted the vehicle brand ( $n_{brand} = 104$ ) and the vehicle category ( $n_{category} = 15$ ); both are listed in Figure 7. This classification data from SUSEP has proven to be a strong predictor of the risk of theft in previous studies [de Azevêdo et al., 2023]. [Peres et al., 2019] provides a comprehensive analysis of the Brazilian auto insurance market.

Nonzero claim counts are relatively rare in this data set, with 92% of policies reporting no collision claims during 2011. The marginal distributions of claim counts and exposure values are shown in Figure 3. We observe a distinct peak in the histogram of exposure values at the median value, which may indicate flawed reporting, the application of a standard exposure scoring formula to a large number of similar inputs, or another cause. We chose not to intervene or further investigate this anomaly.

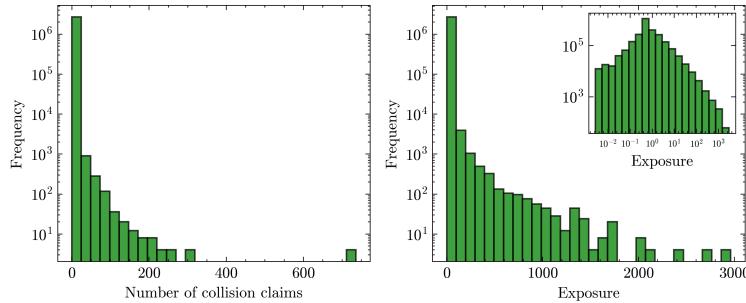


Figure 3: Histograms for the observed claim counts and exposure values for all policies considered.

Originally, this data set did not include coordinates linking each city or municipality with its geographic location. To address this, we employed the Google Maps Geocoding API to map each of 3,785 cities to a city/state/country string. For instance, ‘Imbuia, Santa Catarina, Brazil’ refers to the city of Imbuia in the state of Santa Catarina. This method successfully linked all unique cities to latitude and longitude. 1,404 rows listed the state as ‘Santa Catarina’, but had no city or city code and were listed as `Area` value of ‘Oeste’. These were renamed to ‘Oeste Catarinense’. The states with the most policies were Sao Paulo (860,584), Minas Gerais (346,948), and Rio Grande do Sul (339,644). The states with the fewest policies were Piaui (5,112), Acre (2,632), and Rondonia (1,928).

Having identified coordinates for each city, we performed a spatial join with a data layer for the municipal boundaries. Although the linked municipalities nearly cover the more populated areas of Brazil, many are geographically isolated with no adjacent municipalities, preventing us from creating an adjacency matrix based on the municipal boundaries.

To understand the spatial structure of collision risk, we included city-level covariates. For this, we calculated zonal averages of several globally-available raster data layers using the municipal boundaries for each city through Google Earth Engine [Gorelick et al., 2017]. The layers used for this covariate extraction are listed below. We calculated the mean for each pixel touching a municipal polygon in each case.

**Water occurrence:** For each pixel, the Global Surface Water project [Pekel et al., 2016] monitors the proportion of time water was observed from 1984 - 2022 using Landsat data at a 30 m spatial resolution.

**Precipitation:** The WorldClim project [Fick and Hijmans, 2017] provides long-term monthly climate data interpolated from weather stations from 1970-2000 at a spatial resolution of 1 km. We selected average rainfall for December, typically the wettest month.

**Population density:** To reflect the increased need for collision insurance with higher population density [Sherden, 1984], we utilized the WorldPop project’s [Tatem, 2017] estimate of population density for 2011 at a 100 m resolution.

**Enhanced vegetation index:** The MODIS satellite data set provides an average EVI for January 2011, used as an indicator of urban development and agricultural land use [Huete et al., 2002].

**Elevation:** The elevation data from the Shuttle Radar Topography Mission, with a 90 m resolution, serve as a simple proxy for distance from the coast and are correlated with development patterns.

**Topographic diversity:** The topographic diversity index from [Theobald et al., 2015], measured at a 270 m resolution, helps identify areas with considerable variations in elevation or landforms.

**Forest cover:** Including the global forest cover data at a 270 m resolution from [Shimada et al., 2014] provides additional land use information.

**Travel friction index:** The Global Friction Surface data sets provide indices for the difficulty of motorized and nonmotorized travel at a 1 km resolution, relevant for average trip velocity and collision risk [Weiss et al., 2018].

The specific image IDs and bands applied are detailed in Table 1. All covariates were standardized to have a mean of zero and unit variance before modeling. Scatter plots illustrating cross-variable correlations and marginal distributions are presented in Figure 3.

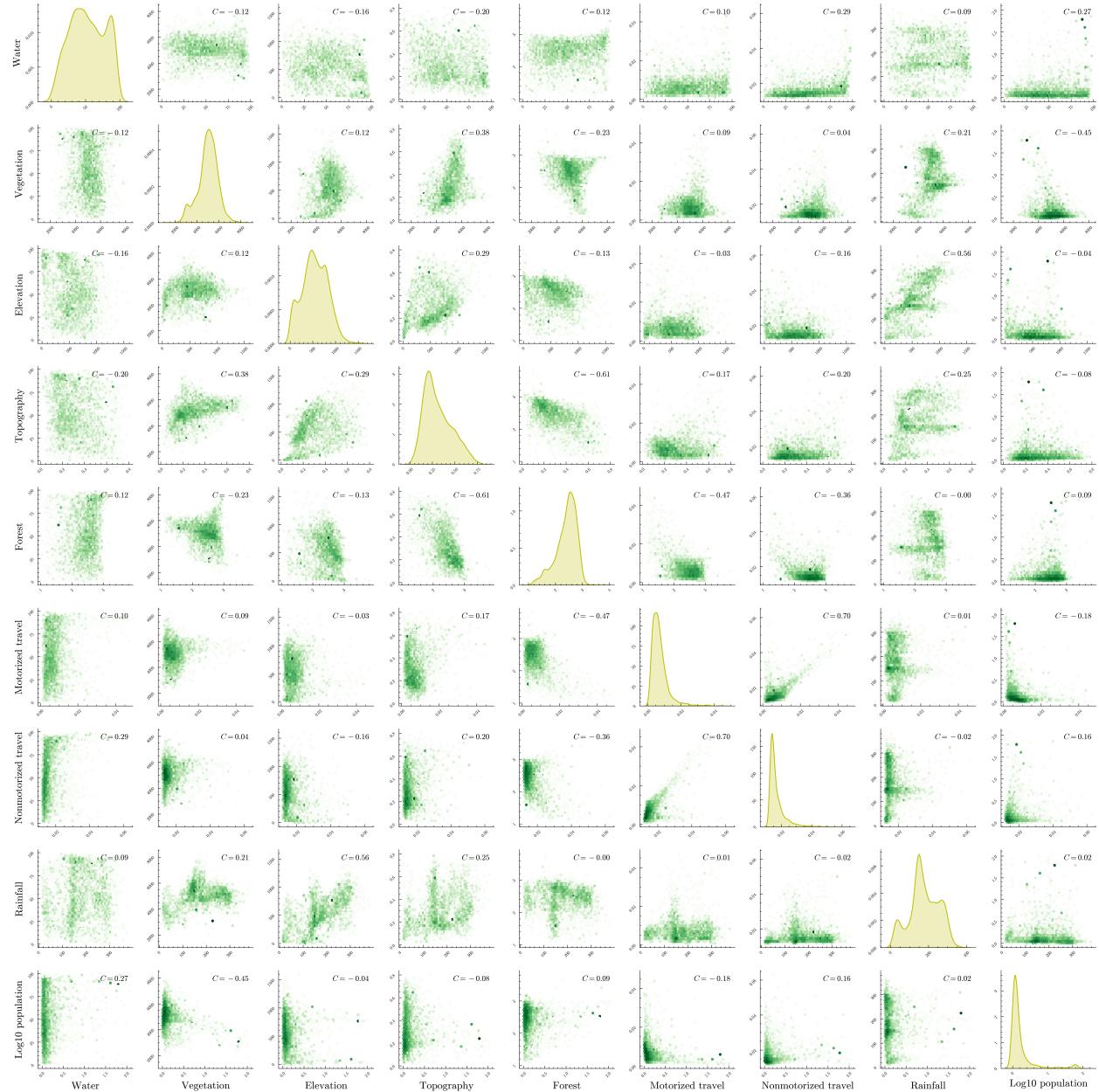


Figure 4: City-level covariates obtained through zonal averaging. Correlations for each pair of variables are displayed in the upper right of the subplot.

## Model Fitting and Benchmarking

### Benchmarking GPU-based acceleration

We compared the time required to evaluate the model's log-posterior density using the full data set mentioned previously. This comparison is more suitable than measures like minimal effective sample size per second

Covariate	Image ID	Band name	Spatial resolution
Water	JRC/GSW1_4/GlobalSurfaceWater	occurrence	30 meters
Population	WorldPop/GP/100m/pop/BRA_2011	population	100 meters
Vegetation	MODIS/061/MOD13A3/2011.01_01	EVI	1000 meters
Elevation	CGIAR/SRTM90_V4	elevation	90 meters
Topography	CSP/ERGo/1_0/Global/ALOS_topoDiversity	constant	270 meters
Forest	JAXA/ALOS/PALSAR/YEARLY/FNF4/2017	fnf	270 meters
Motorized travel	Oxford/MAP/friction_surface_2019	friction	1000 meters
Nonmotorized travel	Oxford/MAP/friction_surface_2019	friction_walking_only	1000 meters
Rainfall	WORLDCLIM/V1/MONTHLY/12	prec	1000 meters

Table 1: Earth Engine images used for city-level covariates. All images were processed using the zonal mean to produce a city-level covariate matrix with dimensions of 3785 by 9

because it avoids bias from inefficiencies in specific implementations of Hamiltonian Monte Carlo or the No-U-Turn sampler, and directly targets the fundamental computational challenges in performing gradient-based MCMC for our purpose. Calculating the gradient 100 times on both GPU and CPU, we recorded a median evaluation time of 5 milliseconds on the GPU and 44 milliseconds on the CPU. This indicates approximately an 8.8 times speed increase when moving from CPU to GPU. All experiments were conducted on a Lambda Labs virtual machine outfitted with an Nvidia A10 GPU with Nvidia driver version 525.85.12 and CUDA version 12.0. The Jax and NumPyro versions installed were 0.4.13 and 0.12.1, respectively. The outcomes from this benchmarking are consistent with the runtime for the entire model fitting process detailed in the next paragraph.

### Model fitting

To obtain parameter estimates under the specified model, we ran the No-U-Turn sampler in NumPyro for 2000 warmup iterations and 2000 sampling iterations. The warmup iterations usually take place before the Markov chain has converged to the stationary distribution and are discarded. We retained every 20th sample due to GPU memory constraints. For finding a starting point for sampling, we used the built-in stochastic variational inference (SVI) [Hoffman et al., 2013] optimization routine for 200,000 iterations and the Adam optimizer with a learning rate of  $10^{-3}$ . The chains were run sequentially; the total runtime of the entire model fitting procedure was 6.7 hours, with chain 1 requiring 194 minutes, chain 2 requiring 189 minutes, and the SVI initialization requiring 19 minutes. To assess MCMC convergence, we calculated  $\hat{R}$  [Gelman and Rubin, 1992] and effective sample size, noting all  $\hat{R}$  values were under 1.10 and all effective sample sizes were over 35. The variables with the largest  $\hat{R}$  values and smallest effective sample sizes were elements of  $\eta$ , the spatially-correlated contributions to the city-level random effect. Based on these statistics, we found no evidence of non-convergence of the Markov chains.

## 4 Results

### Model fit

Figure 5 displays several visual indicators of in-sample model fit. Since a Poisson likelihood was adopted, the relative frequency of predicted claim counts falling outside the predictive  $2\sigma$  credible intervals implied by the Poisson model ( $\lambda_i \pm 2\sqrt{\lambda_i}$ ) is a critical measure of overdispersion. Although Figure 5A illustrates this graphically, an overdispersion test was also conducted by calculating the ratio of the sum of squared Pearson residuals to the degrees of freedom, yielding a test statistic of 1.107 which indicates minimal evidence of overdispersion. Regarding predictive means, the observed and predicted quantiles match well, as shown in Figure 5D. Figure 5B illustrates a strong correlation between predicted and true claim counts which increases with larger  $\lambda_i$ , consistent with the  $\sqrt{\lambda_i}$  scaling of expected deviation for a Poisson distribution. The posterior mean values of  $\lambda_i$  were used to determine the per-record log-likelihood and absolute error. On average, a mean log-likelihood of  $-0.27$  and mean absolute error of 0.16 were produced. The coefficient of determination specific to Poisson-distributed data  $R_{P,P}^2$  was computed using Pearson residuals [Cameron and Windmeijer, 1996], resulting in  $R_{P,P}^2 = 0.94$ . It should be noted that this model incorporates the exposure score as a predictive element, which could be the result of a prior modeling effort.

### Nonlinear exposure adjustment

The posterior estimate of  $g(\alpha_i)$  displayed in Figure 6 indicates a relatively stable relation between exposure and  $g$  for exposures in the range  $10^{-2}$  to 1.0. From 1.0 upwards, there is an increasing linear trend. Since this model lacks a global intercept and regularizes the parameters towards zero through scale priors with

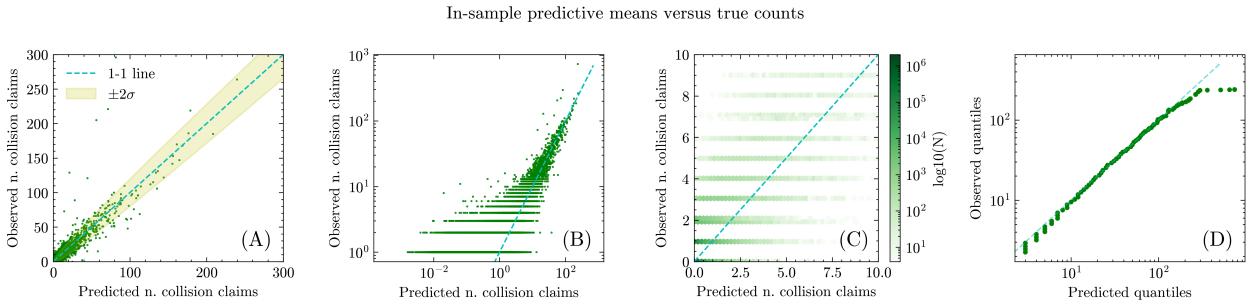


Figure 5: Relationship between true and in-sample posterior predictive means.

modes at zero, the posterior mean of  $g(\alpha_i)$  is not centered at zero. In summary, these estimates indicate that the optimal adjustment to the exposure score remains fairly constant for exposure values near the modeled range from  $10^{-1}$  to 1, with a more pronounced adjustment for policies with high exposure values.

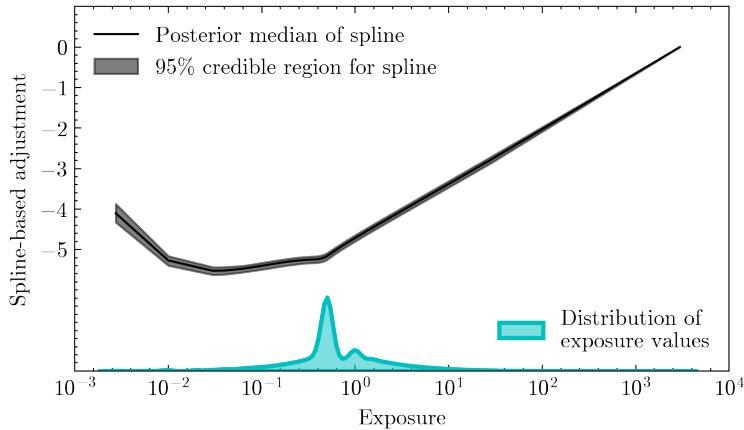


Figure 6: Posterior estimates of nonlinear exposure adjustment

## Categorical predictors

As demonstrated in Figure 7, the posterior estimates for the categorical predictors vary significantly, with some categories showing notable increases in expected claim counts and others decreases. Brands with large impacts include *Engresa*, *Subaru*, and *Ducati*. Approximately one quarter of the brands showed posterior coefficient distributions markedly different from zero, as determined by 95% credible intervals. In terms of vehicle type, sports cars tend to be associated with higher claim counts, whereas trucks, buses, compact cars, and motorcycles typically have lower claim counts. Nonetheless, only compact cars and motorcycles display significantly negative coefficients, according to 95% credible intervals.

## Spatial effect

Despite the inclusion of several city-level covariates, we find that none of these covariates' coefficients have values significantly different from zero, assessed via the 95% posterior credible intervals (Figure 8). The posterior mean values for these coefficients suggest that population density and vegetation may be positively associated with claim counts, while rainfall and forest cover appear to be negatively associated with claim counts. There is also evidence of spatial autocorrelation in  $\eta$  values, demonstrating the utility of this model component in identifying patterns of risk which are not fully explained by covariates, as depicted in Figure 9.

## Temporal effect

The impact of car manufacture date on expected collision claim counts is shown in Figure 10. The relationship between car age and claim count is relatively level for vehicles aged 20 to 30 years. We also noted a nearly linear increase in this effect for cars younger than 20 years, with the trend intensifying as vehicle age decreases. In summary, the risk assessed is stable for vehicles ranging from 20 to 30 years old and decreases with the age for cars under 20 years old.

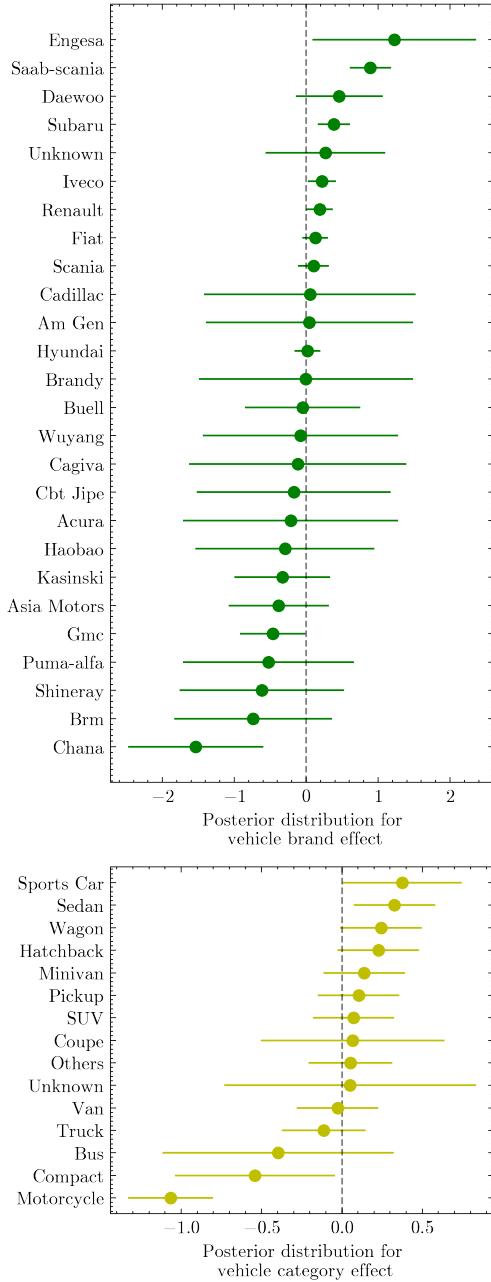


Figure 7: Posterior summaries for parameters linked to vehicle brand and category. Due to space constraints, we show only every fourth brand. All categories are displayed. Lines indicate  $2\sigma$  credible interval bounds and dots represent posterior mean estimates of model parameters.

## 5 Discussion

The case study presented in this work has integrated several modeling subcomponents to address different aspects of risk modeling for insurance. We note while elements such as the spline-based exposure adjustment, vehicle category, and time-varying effect demonstrated strong signals manifesting as coefficients or parameters with posterior credible intervals excluding zero, the city-level covariates did not. We also observed a relatively large number of model parameters and coefficients with wide posterior credible intervals crossing zero, suggesting that the data is not strong enough to constrain these values to useful ranges. A major limitation of this work is the absence of shrinkage priors, which are designed specifically for sparsification [Carvalho et al., 2009], preventing us from applying substantial regularization to push parameter estimates towards zero when needed. Employing such priors would also allow for more model components, like feature-brand interactions or varying effects for the categorical predictors. A more comprehensive analysis could include policy-level covariates like driver age, driving record, and other risk or exposure indicators. This was intended as a demonstration for insurance-relevant models using large ( $> 1000$ ) parameter and data sets in a general-purpose statistical framework like NumPyro. Nonetheless, there is no barrier to creating models with significantly more parameters and data, given enough computational resources. We anticipate future studies using multi-GPU setups, enabling models with  $10^5$  or more parameters and  $> 10^8$  observations.

## Future work

We anticipate further interest in statistical modeling for insurance with large datasets and substantial models, and we view several avenues as particularly relevant directions for future work. There exists considerable

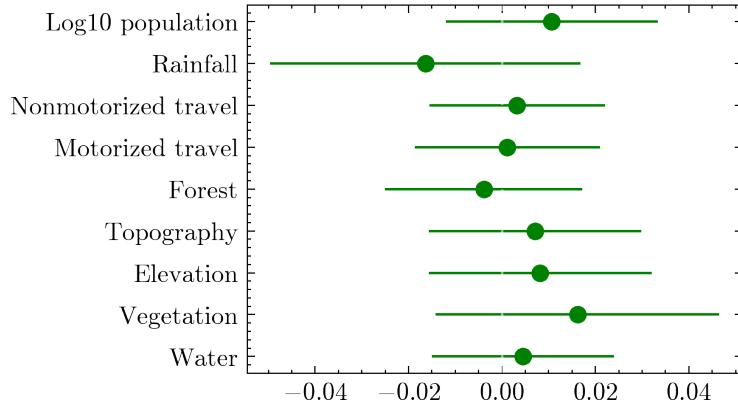


Figure 8: Posterior summaries for parameters relating city-level covariates to predicted counts. Lines indicate  $2\sigma$  credible interval bounds and dots indicate posterior means.

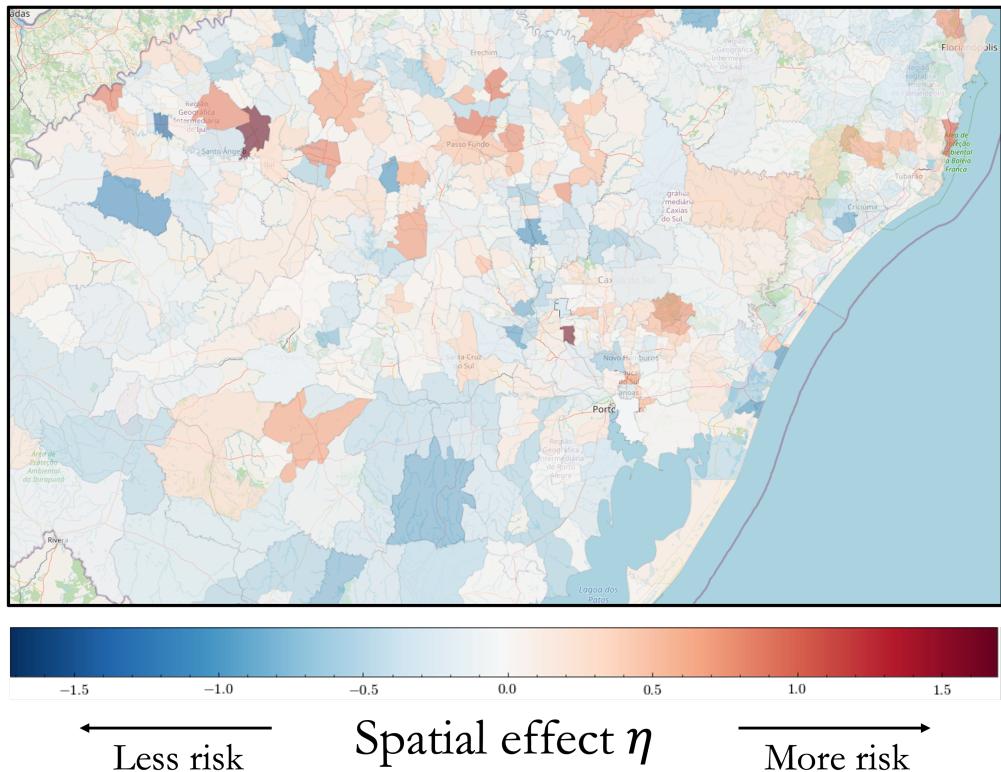


Figure 9: Posterior mean estimates of spatially correlated values of  $\eta$  in the state of Rio Grande do Sul.

potential for extending the model to include multiple types of counts with interesting correlation structures [Bermúdez and Karlis, 2011] to help share information from different observation types [Anastasiadis and Chukova, 2012]. For example, we can imagine a scenario where the risk due to robbery and the risk due to fire are highly correlated because of underlying factors related to emergency services and population density. Individual models of each type may be fit on data without a strong enough signal to constrain parameter estimates; together, however, a multivariate model may share information and consequently enable more insights into joint risk factors. To accommodate data with varying degrees of dispersion reflecting different amounts of underlying heterogeneity, using likelihoods suitable for multiple dispersion types Guikema and Coffelt [2008] may be productive. For a more practical and useful modeling exercise of direct relevance to insurance operations, we may want to model claim frequency and severity together [Frees and Valdez, 2008, Chin and McNulty, 2023] since both types of outcomes are essential for pricing and reserving.

In early modeling iterations, we identified prominent disparities in risk between municipalities close to major urban centers and those farther away. While the model version ultimately presented explains a significant amount of this variation using covariates and the spatial random effect, we are interested in using a latent mixture model to cluster together cities with similar excess risk patterns. Generally, this type of model is hard to accommodate in a gradient-based MCMC framework as the latent class label is usually discrete and thus does not admit a log-posterior which is differentiable in all latent variables. However, both Stan and NumPyro have the ability to either enumerate or automatically marginalize out latent variables of modest cardinality, thus avoiding this issue and allowing for the use of gradient-based MCMC. We would also be interested in using computationally less expensive approximations to the Poisson likelihood [Cameron and Johansson, 1997] to speed up model fitting.

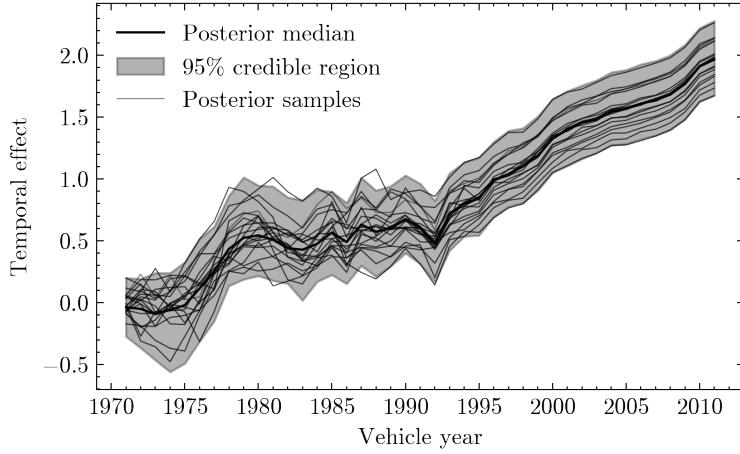


Figure 10: Posterior summary for time-varying effect.

## 6 Conclusion

This study demonstrates the value of state-of-the-art statistical modeling frameworks like NumPyro that work seamlessly on GPUs to speed up model fitting without sacrificing any flexibility in the selection of model components. We developed a log-additive model for collision claim frequencies of auto insurance policies in Brazil. This model includes a nonlinear exposure adjustment, categorical predictors, a city-level random effect with spatial correlation, and a dynamic temporal effect. The model succeeded in explaining much of the data variance, though the city-level covariates showed weak effects. Moreover, we discovered that NumPyro’s GPU-based implementation made the model fitting process nine times faster than using a CPU, as shown in the log posterior gradient calculations, enabling completion in a practical timeframe. This study should interest both practitioners and researchers in the insurance sector, as well as those who employ Bayesian methods for analyzing extensive spatial datasets.

## 7 Data Availability

All code and data are available for replication at <https://github.com/ckrapu/bayes-at-scale>.

## References

- Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. page 19, 2015.
- Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. PyMC: A modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, September 2023. ISSN 2376-5992. doi: 10.7717/peerj-cs.1516.
- Simon Anastasiadis and Stefanka Chukova. Multivariate insurance models: An overview. *Insurance: Mathematics and Economics*, 51(1):222–227, July 2012. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2011.01.013.
- Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022.
- Roberto Basile, María Durbán, Román Mínguez, Jose María Montero, and Jesús Mur. Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245, November 2014. ISSN 0165-1889. doi: 10.1016/j.jedc.2014.06.011.
- James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: A CPU and GPU Math Compiler in Python. page 7, 2010.
- Lluís Bermúdez and Dimitris Karlis. Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, 48(2):226–236, March 2011. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2010.11.001.
- Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, March 1991. ISSN 1572-9052. doi: 10.1007/BF00116466.
- Alexandros Beskos and Andrew Stuart. Computational Complexity of Metropolis-Hastings Methods in High Dimensions. In Pierre L’ Ecuyer and Art B. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 61–71, Berlin, Heidelberg, 2009. Springer. ISBN 978-3-642-04107-5. doi: 10.1007/978-3-642-04107-5\_4.
- Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint:1701.02434*, page 60, 2017.
- Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019. ISSN 1533-7928.
- Jean-Philippe Boucher, Michel Denuit, and Montserrat Guillén. Models of Insurance Claim Counts with Time Dependence Based on Generalization of Poisson and Negative Binomial Distributions. *Variance*, 2 (1), 2008.
- Stephen P. Brooks. Markov Chain Monte Carlo Method and Its Application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):69–100, 1998. ISSN 0039-0526.
- A. Colin Cameron and Per Johansson. Count Data Regression Using Series Expansions: With Applications. *Journal of Applied Econometrics*, 12(3):203–223, 1997. ISSN 0883-7252.
- A. Colin Cameron and Frank A. G. Windmeijer. R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization. *Journal of Business & Economic Statistics*, 14(2):209–220, 1996. ISSN 0735-0015. doi: 10.2307/1392433.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling Sparsity via the Horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 73–80. PMLR, April 2009.

Arthur Charpentier. *Computational Actuarial Science with R*. 1st edition, 2015. ISBN 978-1-138-03378-8.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, October 1998. ISSN 0304-4076. doi: 10.1016/S0304-4076(97)00115-2.

Stephanie Chin and Greg McNulty. A Bayesian Model for Estimating Long Tailed Excess of Loss Reinsurance Loss Costs. *CAS E-Forum*, Summer, August 2023.

Noel Cressie and Ngai H. Chan. Spatial Modeling of Regional Variables. *Journal of the American Statistical Association*, 84(406):393–401, 1989. ISSN 0162-1459. doi: 10.2307/2289922.

Beatriz Margarida Zanotto de Azevêdo, Luiz Augusto Finger França Maluf, Joelson Oliveira Sampaio, and Vinicius Augusto Brunassi Silva. Estimating vehicle types effects on auto insurance premiums in São Paulo City: A GAMLS Approach. *CONTRIBUCIONES A LAS CIENCIAS SOCIALES*, 16(10):21008–21044, October 2023. ISSN 1988-7833. doi: 10.55905/revconv.16n.10-140.

Kexing Ding, Baruch Lev, Xuan Peng, Ting Sun, and Miklos A. Vasarhelyi. Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies*, 25(3):1098–1134, September 2020. ISSN 1573-7136. doi: 10.1007/s11142-020-09546-9.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2), September 1987.

Christophe Dutang and Arthur Charpentier. CASdatasets, 2022.

Stephen E. Fick and Robert J. Hijmans. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, October 2017. ISSN 08998418. doi: 10.1002/joc.5086.

Edward W. Frees and Emiliano A. Valdez. Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*, 103(484):1457–1469, 2008. ISSN 0162-1459.

Roy Frostig, Matthew James Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. In *SysML*, Stanford, CA, February 2018.

Alan E. Gelfand and Penelope Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics (Oxford, England)*, 4(1):11–25, January 2003. ISSN 1465-4644. doi: 10.1093/biostatistics/4.1.11.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, September 2006. ISSN 1936-0975, 1931-6690. doi: 10.1214/06-BA117A.

Andrew Gelman and Bob Carpenter. Bayesian Analysis of Tests with Unknown Specificity and Sensitivity. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 69(5):1269–1283, August 2020. ISSN 0035-9254. doi: 10.1111/rssc.12435.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4), 1992.

Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992. ISSN 0883-4237.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202: 18–27, December 2017. ISSN 0034-4257. doi: 10.1016/j.rse.2017.06.031.

Seth D. Guikema and Jeremy P. Coffelt. A flexible count data regression model for risk analysis. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 28(1):213–223, February 2008. ISSN 1539-6924. doi: 10.1111/j.1539-6924.2008.01014.x.

W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970. ISSN 0006-3444. doi: 10.2307/2334940.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, page 31, 2014.

Matthew D Hoffman, David Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 2013.

A Huete, K Didan, T Miura, E. P Rodriguez, X Gao, and L. G Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1):195–213, November 2002. ISSN 0034-4257. doi: 10.1016/S0034-4257(02)00096-2.

Yisu Jia, Stefanos Kechagias, James Livsey, Robert Lund, and Vladas Pipiras. Latent Gaussian Count Time Series. *Journal of the American Statistical Association*, 118(541):596–606, January 2023. ISSN 0162-1459. doi: 10.1080/01621459.2021.1944874.

Xiaoping Jin, Bradley P. Carlin, and Sudipto Banerjee. Generalized Hierarchical Multivariate CAR Models for Areal Data. *Biometrics*, 61(4):950–961, December 2005. ISSN 0006341X. doi: 10.1111/j.1541-0420.2005.00359.x.

Christopher Krapu and Mark Borsuk. Probabilistic programming: A review for environmental modellers. *Environmental Modelling & Software*, 114:40–48, April 2019. ISSN 1364-8152. doi: 10.1016/j.envsoft.2019.01.014.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2796–2804. PMLR, July 2018.

David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009. ISSN 1097-0258. doi: 10.1002/sim.3680.

Daniel Lupton. Machine Learning in Insurance. 2022.

Gael M. Martin, David T. Frazier, and Christian P. Robert. Approximating Bayes in the 21st Century. *Statistical Science*, -1(-1), January 2023. ISSN 0883-4237. doi: 10.1214/22-STS875.

Ophélia Miralles, Anthony C. Davison, and Timo Schmid. Bayesian modeling of insurance claims for hail damage, August 2023.

Radford Neal. MCMC Using Hamiltonian Dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, volume 20116022. Chapman and Hall/CRC, May 2011. ISBN 978-1-4200-7941-8 978-1-4200-7942-5. doi: 10.1201/b10905-6.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. <https://arxiv.org/abs/1912.01703v1>, December 2019.

Nicholas D. Paulson, Chad E. Hart, and Dermot J. Hayes. A spatial Bayesian approach to weather derivatives. *Agricultural Finance Review*, 70(1):79–96, May 2010. ISSN 0002-1466. doi: 10.1108/00021461011042657.

Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, December 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature20584.

Viviane Maria Peres, Wilfredo L. Maldonado, and Osvaldo Candido. Automobile insurance in Brazil: Market concentration and demand. *Revista Contabilidade & Finanças*, 30:396–408, May 2019. ISSN 1519-7077, 1808-057X. doi: 10.1590/1808-057x201808300.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv:1912.11554 [cs, stat]*, December 2019.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, February 1998. ISSN 1369-7412, 1467-9868. doi: 10.1111/1467-9868.00123.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, December 1996. ISSN 1350-7265.

Håvard Rue, Andrea Riebler, Sigrunn H. Sørbye, Janine B. Illian, Daniel P. Simpson, and Finn K. Lindgren. Bayesian Computing with INLA: A Review, September 2016.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, April 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.55.

William A. Sherden. An Analysis of the Determinants of the Demand for Automobile Insurance. *The Journal of Risk and Insurance*, 51(1):49–62, 1984. ISSN 0022-4367. doi: 10.2307/252800.

Masanobu Shimada, Takuya Itoh, Takeshi Motooka, Manabu Watanabe, Tomohiro Shiraishi, Rajesh Thapa, and Richard Lucas. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sensing of Environment*, 155:13–31, December 2014. ISSN 0034-4257. doi: 10.1016/j.rse.2014.04.014.

Daniel P. Simpson, Håvard Rue, Thiago G. Martins, Andrea Riebler, and Sigrunn H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors, August 2015.

Andrew J. Tatem. WorldPop, open data for spatial demography. *Scientific Data*, 4(1):170004, January 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.4.

Benjamin M. Taylor and Peter J. Diggle. INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284, October 2014. ISSN 0094-9655. doi: 10.1080/00949655.2013.788653.

David M. Theobald, Dylan Harrison-Atlas, William B. Monahan, and Christine M. Albano. Ecologically-Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning. *PLOS ONE*, 10(12):e0143619, December 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0143619.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2.

Jens Christian Wahl, Fredrik Lohne Aanes, Kjersti Aas, Sindre Froyn, and Daniel Piacek. Spatial modelling of risk premiums for water damage insurance. *Scandinavian Actuarial Journal*, 2022(3):216–233, March 2022. ISSN 0346-1238. doi: 10.1080/03461238.2021.1951346.

D. J. Weiss, A. Nelson, H. S. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, E. Poyart, S. Belchior, N. Fullman, B. Mappin, U. Dalrymple, J. Rozier, T. C. D. Lucas, R. E. Howes, L. S. Tusting, S. Y. Kang, E. Cameron, D. Bisanzio, K. E. Battle, S. Bhatt, and P. W. Gething. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688):333–336, January 2018. ISSN 1476-4687. doi: 10.1038/nature25181.

Yanwei Zhang. Bayesian Analysis of Big Data in Insurance Predictive Modeling Using Distributed Computing, July 2017.