unibz
Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
Faculty of Economics and Management

BEMPS –
Bozen Economics & Management
Paper Series

NO 98/2022

Modelling spatial correlation between earthquake insured losses in New Zealand: a mixed-effects analysis

F. Marta L. Di Lascio, Ilan Noy,
Selene Perazzini

Modelling spatial correlation between earthquake insured losses in New Zealand: a mixed-effects analysis

F. Marta L. Di Lascio* Ilan Noy† Selene Perazzini‡

December 22, 2022

Abstract

Earthquake insurance is a critical risk management strategy that contributes to improving recovery and thus greater resilience of individuals. Insurance companies construct premiums without taking into account spatial correlations between insured assets. This leads to potentially underestimating the risk, and therefore the exceedance probability curve. We here propose a mixed-effects model to estimate losses per ward that is able to account for heteroscedasticity and spatial correlation between insured losses. Given the significant impact of earthquakes in New Zealand due to its particular geographical and demographic characteristics, the government has established a public insurance company that collects information about the insured buildings and any claims lodged. We thus develop a two-level variance component model that is based on earthquake losses observed in New Zealand between 2000 and 2021. The proposed model aims at capturing the variability at both the ward and territorial authority levels and includes independent variables, such as seismic hazard indicators, the number of usual residents, and the average dwelling value in the ward. Our model is able to detect spatial correlation in the losses at the ward

*Faculty of Economics and Management, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy. E-mail: *marta.dilascio@unibz.it*

†School of Economics and Finance, Victoria University of Wellington, New Zealand. E-mail: *ilan.noy@vuw.ac.nz*

‡DMS StatLab, Department of Economics and Management, University of Brescia, Brescia, Italy. E-mail: *selene.perazzini@unibs.it*

level thus increasing its predictive power and making it possible to assess the effect of spatially correlated claims that may be considerable on the tail of loss distribution.
Classification-JEL: C10, C21.
Keywords: Earthquake losses; Insurance; Mixed-effects model; Spatial correlation; Variance component model.

1 Introduction

Disasters caused by natural hazards have obvious adverse consequences for people and the economies they affect (IPCC, 2022, UNDRR, 2022). Catastrophe insurance - for residential and commercial properties, for infrastructure, and for business interruption - is often seen as an integral part of the disaster risk management toolkit. However, insurance is not a panacea as it transfers the financial risk to the insurer (and sometimes through re-insurance to global financial markets). Evidence does suggest, though, that insurance and similar financial risk transfer instruments (such as catastrophic bonds) can enable improved recovery and thus increase resilience ([Owen et al., 2021], [Nguyen et al., 2020]).[1].

However, the literature has long been documenting that insuring catastrophic risks is complex and not easily achieved [Kunreuther and Michel-Kerjan, 2014]. Indeed, coverage rates for catastrophic risks from private insurers are notoriously low in many areas where there is significant risk ([Nguyen and Noy, 2020] and [Barbieri et al., 2022]), and even public insurance systems struggle to provide widespread cover, except for instances where insurance is mandatory [Schwarze and Croonenbroeck, 2017]. For documentation of the extent of the catastrophic insurance cover gap globally, see [Lloyd's, 2017].

There are both demand and supply obstacles that appear to reduce actual insurance penetration and that help explain the large global insurance gap, even though many different types of disaster insurance products are available or could easily be offered. Generally, private insurers are reluctant to cover the risk of hazards like earthquakes or hurricanes, as this risk is heavily correlated across their portfolio, and the risk is also often difficult to quantify. Therefore, it is often governments that end up providing this insurance, and surprisingly rarely re-insure the extreme tail of this risk

[1]For a review of this literature, see [Kousky, 2019]

internationally through the global reinsurance market [Ito and McCauley, 2022]. Examples of these public insurance schemes include flood insurance programs in the United States and the UK (National Flood Insurance Program and FloodRE, respectively), micro-insurance for crop losses in India, seismic and geothermal hazard insurance in New Zealand, and earthquake insurance in Turkey.

Earthquakes, in particular, are a very significant hazard in many countries, especially around the Pacific Ocean Rim, in mountainous Central and South Asia, and the Central and Eastern Mediterranean. Other regions may not experience such strong earthquakes, but very high vulnerability makes them equally risky (e.g., some parts of the Caribbean). Mortality from earthquakes can be very high, and indeed the highest mortality risk, globally, is very clearly from earthquakes and the tsunamis they generate. More than half a million people died in the four most lethal events since the turn of the century (2004 in Indonesia, 2008 in China, 2010 in Haiti, and 2011 in Japan), by far more than in any other type of disaster. Earthquakes also destroy very high-valued physical assets and infrastructure - the Japan 2011 earthquake-tsunami disaster was the costliest disaster event ever. Yet, in most of these cases, the levels of insurance penetration were quite low, and not much has changed since these events occurred.

Given these observations, it is not surprising that risk transfer tools, and especially insurance, constitute a significant focus for policy efforts in all the countries facing high seismic risks. Of specific relevance for us is the supply of earthquake insurance to residential homes. This insurance is almost always supplied by governments directly (e.g., New Zealand), backed by the government explicitly (e.g., Japan), backed by the government implicitly (e.g., California), or provided in an ad-hoc manner (e.g., Italy).

The need to provide some kind of safety net, especially for residential homeowners, imposes a difficult decision on public policymakers who control these public insurance schemes. As one alternative, they could institute full risk-based pricing of these insurance schemes, thereby aligning incentives for re-settlement or strengthening for homeowners. They could also charge politically unpalatable rates from different homeowners (and often more from low-income ones). On the other extreme, such schemes could use a flat premium fee and charge everyone exactly the same, irrespective of the risks they face. This choice also has potentially important redistributive

ramifications, as low-risk homeowners will subsidize high-risk ones. Overall, governments are recognizing that aligning incentives, and preventing moral hazard, is potentially important. Using the insurance premium as a risk signaling device is increasingly recognized as potentially useful, especially since evidence suggests that other types of risk signaling (e.g., warning on property titles, or published hazard maps online) are not as effective as one could reasonably expect them to be (e.g., [Filippova et al., 2020]).

Recognizing that risk-based premiums are one of the most plausible ways to generate de-risking behaviour by homeowners, our purpose here is to propose a novel way to calculate risk-based premiums based on detailed spatial modelling of the hazard using the existing record of insurance claims in Aotearoa New Zealand. New Zealand is an obvious choice for such a case study given its high frequency of earthquakes, its comprehensive insurance penetration (>95 percent), and the presence of a single public insurer that is willing to share its comprehensive unit-record claims data with researchers.

We estimate expected earthquake losses accounting for their spatial correlation. This in turn allows the insurer to differentiate premiums based on local characteristics. Given the limited data available on the housing stock itself, we can calculate this risk differentiation at the ward level, i.e. electoral district. Our main contribution is the model's ability to capture and quantify the effect of spatially correlated claims, as even low spatial correlation may considerably inflate the tail of loss distribution. We were able to detect spatial correlation between wards, which are fairly small areas. Though spatial correlation between the wards is low, it has a bigger effect on aggregate losses. These findings are also relevant for the assessment of the amount of liquidity insurers need, and the amount of re-insurance they should purchase, given their portfolio of exposures.

The rest of the paper is organised as follows. The next section describes the innovative data we use. Section 3 describes the mixed-effects model we estimate. Next, the empirical analysis is presented in section 4, and we end with a discussion and conclusion section (sections 5 and 6, respectively).

2 Data

We here describe the innovative dataset analysed in this paper. Our database represents an almost unique structured source of information on natural dis-

asters, such as earthquakes. Indeed, it captures approximately all buildings in New Zealand and their earthquake losses over a rather long observation period, i.e. about 20 years.

Data on losses have been provided by the New Zealand Earthquake Commission (EQC hereafter) and refer to the earthquake insurance coverage EQCover. The database collects information about the insured buildings and any claims lodged against the EQC between 2000 and 2021. Given the extraordinarily high insurance penetration rate in New Zealand, the insured dwellings database contains information about the 95% of the housing stock in the country. Insured properties are localized by longitude and latitude (with minor adjustment to preserve anonymity), and have thus been assigned to their respective ward and territorial authority (TA hereafter), which is the second tier of local government in New Zealand, through reverse geocoding. For this, we refer to the New Zealand 2019 local boundary maps released by Land Information New Zealand (LINZ hereafter)[2]. In case of missing coordinates, records have been referenced by means of postcodes. Overall, our database concerns 236 wards and 66 TAs in New Zealand.

As far as claims are concerned, we limited our analysis to open or accepted claims only (i.e., we remove rejected claims). To overcome issues generated by the time gap between the moment at which the damage occurred and the opening of the claims as well as the effect of earthquake sequences, insured losses are commonly aggregated over each catastrophic event. Moreover, to capture spatial correlations we consider the sum of the insurer's losses due to claims reported between 2000 and 2021 aggregated at the ward level. Indeed, a ward is the ideal geographical area to consider because it is sufficiently large to detect low spatial correlations but small enough to differentiate local risk.

Table 1 summarizes all the variables used for the proposed analysis. We use variables concerning mainly the seismic risk, the characteristics of insured buildings, and inhabitants. Three variables in the EQC database have been included in the analysis: total loss per ward divided by the number of dwellings in the ward (Y), the median Cresta zone[3] of the ward (X_1), and the mean value of dwellings in the ward (X_2). Additional information has been sourced from the national statistics institute called Statistics NZ

[2] `https://datafinder.stats.govt.nz/layer/98742-ward-2019-generalised/`

[3] `https://about.cresta.org/`

(STATS NZ hereafter). In particular, the number of individuals usually resident in a ward (X_3) and the rate of housing with reported problems (heating, mold, etc..) (X_6) and the average weekly income (X_7), which are available at the larger region level and have been disaggregated at the ward level. In addition, the Z seismic risk score (X_4) [4] used to determine building standards has been taken for the wards from "NZ standards"[5]. Finally, the rate of earthquakes (Z) in the territorial authority has also been included in the model and it has been computed considering all the earthquakes of magnitude bigger than 3.5 that have occurred in New Zealand from January 1900 to May 2020 and reported in the GeoNet earthquake catalogue[6].

[4] https://www.building.govt.nz/managing-buildings/managing-earthquake-prone-buildings/how-the-system-works/z-values-seismic-risk/

[5] https://www.standards.govt.nz/shop/nzs-1170-52004/

[6] https://www.geonet.org.nz/data/types/eq_catalogue

Table 1: Description of the innovative database analysed.

Variable	Name	Description	Range	Mean (Standard deviation)	Source
i	Ward	Ward ID	-	-	LINZ
j	TA	TA ID	-	-	LINZ
Y	Loss	Total loss 2000-2021 of the ward divided by total number of dwelling in the ward	[0.03,71711]	3390.00 (11425.97)	EQC
X_1	Cresta zone	Median value of Cresta zone in the ward	[1,16]	8.55 (4.67)	EQC
X_2	Dwelling value	Average dwelling value in the ward	[138087,346545]	233473 (32841.11)	EQC
X_3	Usual residents	Number of individuals usually resident in the ward	$[729, 176459]$	20310.3 (30630.19)	STATS NZ
X_4	Z seismic risk score	Seismic risk index per ward	$[0.1, 0.55]$	0.26 (0.10)	NZ Standard
X_5	Density	Number of dwellings per km^2 per ward	$[0, 7.93]$	0.51 (0.97)	LINZ, EQC
X_6	Housing with problems	Rate of housing with problems (e.g. mold, humidity, etc...) in the ward	$[0.23, 0.41]$	0.33 (0.04)	STATS NZ
X_7	Average weekly income	2019 average weekly income in NZ $ in the ward	$[332, 482]$	390.3 (43.19)	STATS NZ
Z	Earthquake rate	Rate of eartquakes of magnitude $\geq$3.5 in the TA	$[0, 0.13]$	0.015 (0.03)	GeoNet
	Long, Lat	Coordinates of ward's centroid	-	-	EQC

3 A mixed-effects model with spatial within-group correlation

Here we develop a mixed-effects model able to account for heteroscedasticity and spatial correlation between insured earthquake losses. The model formulation is inspired by the work of Laird and Ware [1982]. Here the ward's losses per building (Y) are represented by a two-level variance component model with m level 2 units, i.e. the territorial authorities, and n level 1 units, i.e. the wards

$$\log(y_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 \log(X_{3ij}) + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \\ + \beta_6 X_{6ij} + \beta_7 X_{7ij} + u_{1j} + u_{2j} Z_j + \varepsilon_{ij} \quad (1)$$

where $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$, coefficients β_k with $k = 0, \ldots, 7$ are the fixed effects of the model, while u_{1j} and u_{2j} are the random effects, and ε_{ij} is the within-group error. Since the distribution of losses per ward is highly positively skewed, the logarithmic transformation has been chosen to normalize the dependent variable (log-Loss hereafter). The model includes independent variables aimed at capturing the variability at both the ward and TA level, denoted respectively by X_k, with $k = 1, \ldots, 7$, and Z and described in section 2. The log-transformation has been applied to the number of usual residents, X_3, as the range of values is considerably high. Finally, the variable Z_j associated with the random effect u_{2j} is the earthquake rate of the TA. It is worth noticing that in the defined model, we have two levels of random variation but one nested level of random effects. Hence, to avoid misunderstandings, we refer to the model defined as a two-level model of variance components and avoid using terminology from the literature on multilevel modelling.

In order to introduce a source of variability associated with the TAs and a within-TA variability, for the model in Eq. (1) we assume the following

Assumption 1

$$\boldsymbol{u}_1 \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{u1}\boldsymbol{I}), \quad \boldsymbol{u}_2 \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{u2}\boldsymbol{I}), \quad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}) \quad (2)$$

where $\boldsymbol{u}_1 = (u_{11}, \ldots, u_{1j}, \ldots, u_{1m})$, $\boldsymbol{u}_2 = (u_{21}, \ldots, u_{2j}, \ldots, u_{2m})$, $\boldsymbol{u} =$

$(\boldsymbol{u}_1, \boldsymbol{u}_2)$, $\boldsymbol{I}$ is a $(m \times m)$-dimensional identity matrix, and

$$\Psi = \begin{pmatrix} \sigma^2_{u_1}\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{u_2}\boldsymbol{I} \end{pmatrix} \tag{3}$$

is the $(2m \times 2m)$-dimensional variance-covariance matrix of the random effects.

Assumption 2

$$\begin{aligned} &\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_{ij}}), \quad \sigma^2_{\varepsilon_{ij}} = \sigma^2_{\varepsilon} g^2(\mu_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta}) \\ &\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0, \quad \forall\, j \neq j', \forall\, i, i' \end{aligned} \tag{4}$$

where σ^2_{ε} is the error variance component constant over i and j, $\mu_{ij} = E[y_{ij}|u_{1j}, u_{2j}]$, $\boldsymbol{v}_{ij}$ is a vector of variance covariates, $\boldsymbol{\delta}$ is a vector of variance parameters and $g(\cdot)$ is the variance function, assumed continuous in $\boldsymbol{\delta}$.

Note that if $g(\mu_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta}) = 1$ for all i and j, $\sigma^2_{\varepsilon_{ij}} = \sigma^2_{\varepsilon}$ and the model has homoscedastic variance.

Assumption 3

$$\begin{aligned} &\mathrm{cov}(u_{1j}, \varepsilon_{ij}) = 0 \quad , \qquad \mathrm{cov}(u_{2j}, \varepsilon_{ij}) = 0 \\ &\mathrm{cov}(u_{1j}, \varepsilon_{ij'}) = 0 \quad , \qquad \mathrm{cov}(u_{2j}, \varepsilon_{ij'}) = 0 \end{aligned} \tag{5}$$

$\forall$ i, j, and j'.

The variance of y_{ij} thus results as follows

$$\begin{aligned} \mathrm{var}(y_{ij}|\beta_0, \ldots, \beta_7, X_{1ij}, \ldots, X_{7ij}, Z_j) &= \mathrm{var}(u_{1j} + u_{2j} + \varepsilon_{ij}) \\ &= \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_{\varepsilon_{ij}} \end{aligned} \tag{6}$$

The assumptions in Eq.s (2)-(5) imply a block diagonal covariance matrix

that exhibits serial correlation between the wards in the same TA, that is

$$\begin{aligned}
\operatorname{cov}(y_{ij}, y_{i'j'}) &= \operatorname{cov}(u_{1j} + u_{2j} + \varepsilon_{ij}, u_{1j'} + u_{2j'} + \varepsilon_{i'j'}) \\
&= \operatorname{cov}(u_{1j} + u_{2j}, u_{1j'} + u_{2j'}) + \operatorname{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) \\
&= \begin{cases} \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_{\varepsilon_{ij}}, & \text{if } i = i', j = j' \\ \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{ii'}, & \text{if } i \neq i', j = j' \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{7}$$

where $\sigma^{(j)}_{ii'}$ is the within-group covariance. Thus, each block Σ_{jj} in the diagonal of the variance-covariance matrix of y_{ij} represents the variance-covariance matrix for the j-th TA with n_j wards. The generic Σ_{jj} is a symmetric square matrix

$$\Sigma_{jj} = \begin{pmatrix} \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_{\varepsilon_{ij}} & \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{12} & \cdots & \sigma^2_{u1} + \sigma^2_{u2} + \sigma^{(j)}_{1n_j} \\ \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{12} & \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_{\varepsilon_{ij}} & \cdots & \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{2n_j} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{n_j1} & \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{n_j2} & \cdots & \sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_{\varepsilon_{ij}} \end{pmatrix}. \tag{8}$$

The within-group correlation, which in our model is the correlation between wards within a TA, is thus

$$\rho_{ij,i'j} = \frac{\left(\sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^{(j)}_{ii'}\right)}{(\sigma^2_{u_1} + \sigma^2_{u_2} + \sigma^2_{\varepsilon_{ij}})}. \tag{9}$$

We assume $\rho_{ij,i'j}$ equal to a function $h(\cdot)$ of the Euclidean distance $d_{ii'}$ between the centroids of two wards i and i', where the centroid of a ward is the average of geographical coordinates (longitude and latitude) of all the points located in it, and a parameter r given by the distance where the variogram first flattens out and reaches the sill [Cressie, 1993, Cressie and Wikle, 2011]

$$\rho_{ij,i'j} = h(d_{ii'}, r). \tag{10}$$

Since the spatial correlation between two wards is stronger the closer they are and becomes equal to 0 after a certain distance - coherently with the assumption of our model -, we assume that wards in the same TA are correlated, while TAs are uncorrelated. The between-class correlation is, therefore, $\rho_{ij,i'j'} = 0$, but we take into account the potential effect of neighbouring

regions on the response variable in modelling empirical heteroscedasticity (see Sect. 4 for details).

3.1 Estimation method

Equations (1) and (7) require the estimation of eight fixed coefficients ($\boldsymbol{\beta} = \{\beta_0, \ldots, \beta_7\}$), $\sum_{j=1}^{m} \frac{n_j(n_j-1)}{2}$ variance components ($\sigma_{ii'}^{(j)}$ $\forall\, j = 1, \ldots, m$ and i and i' belong to $\{1, \ldots, n_i\}$), two random effects variances ($\sigma_{u_1}^2$ and $\sigma_{u_2}^2$), and the within-group error variance ($\sigma_{\varepsilon_{ij}}^2$ by varying i and j). In light of the Eq. (10), it is not necessary to estimate the covariances $\sigma_{ii'}^{(j)}$ directly, but it is sufficient to calculate the distances $d_{ii'}$ between each pair of wards i and i'. This strongly lightens the computational burden of the model estimation. For simplicity, we indicate with $\boldsymbol{\theta}$ the variance components vector of the entire model (including both the variance components of the random effects and the set of parameters for the variance function $g(\cdot)$ in Eq. (4)).

In order to avoid underestimation of variance components, we use the restricted maximum likelihood (REML hereafter) estimation method [Patterson and Thompson, 1971, Harville, 1977, Laird and Ware, 1982] by exploiting the variance-covariance parameterizations described in Pinheiro and Bates. [1996]. The restricted likelihood corresponding to the model in Eqs. (1) based on assumptions in Eqs. (2)-(5) and within-group correlation as in Eq. (10) is defined by integrating out the fixed effects from the likelihood as follows

$$L_R(\boldsymbol{\theta}, \sigma_\varepsilon^2 | \log(\boldsymbol{y})) \quad = \quad \int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2 | \log(\boldsymbol{y})) d\boldsymbol{\beta} \tag{11}$$

where $\log(\boldsymbol{y})$ is the vector of two-levels observations. The log-restricted likelihood $l_R(\boldsymbol{\theta}, \sigma_\varepsilon^2 | \log(\boldsymbol{y})) = \log L_R(\boldsymbol{\theta}, \sigma_\varepsilon^2 | \log(\boldsymbol{y}))$ produces the conditional estimate for $\hat{\sigma}_\varepsilon^2(\boldsymbol{\theta})$ from which we obtain the profiled log-restricted-likelihood $l_R(\boldsymbol{\theta}, \hat{\sigma}_\varepsilon^2(\boldsymbol{\theta}) | \log(\boldsymbol{y}))$. This is optimized with respect to $\boldsymbol{\theta}$ only, and using the resulting REML estimate of $\boldsymbol{\theta}$ to obtain the REML estimate of σ_ε^2. Similarly, it has been done to obtain the REML estimates of variance components $\boldsymbol{\theta}$. The optimization method used is the quasi-Newton method due to Broyden [1970], Fletcher [1987].

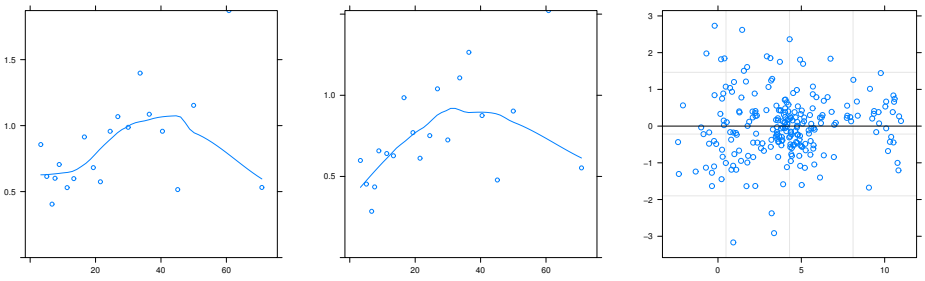

Figure 1: Model in Eq. (1) estimated assuming within-group correlation $\rho_{ij,i'j} = 0$ and homoscedasticity. Left: semivariogram of within-group residuals; the x-axis denotes the distance, the y-axis represents the semivariogram. Center: semivariogram of the normalized residuals; the x-axis denotes the distance, the y-axis represents the semivariogram. Both the semivariograms have been presented by setting the maximum distance to 80 km. Right: fitted (x-axis) versus standardized (y-axis) residuals.

4 Empirical analysis

The model in Eq. (1) with $\rho_{ij,i'j} = 0$, is estimated to investigate the spatial correlation hypothesis. The semivariogram of residuals obtained from this base model versus the distance between the wards within a TA is reported in Figure 1, left, and shows a clear pattern. Since various isotropic variogram models might capture the observed spatial correlation structure, several functional hypotheses have been estimated for $h(d_{ii'}, r)$ (see Eq. (10)). According to the behaviour of the semivariogram, the best fit is obtained when assuming a Gaussian correlation structure

$$h(d_{ii'}, r) = (1 - nugg)\mathrm{e}^{-\left(\frac{\mathrm{d_{ii'}}}{\mathrm{r}}\right)^2} \tag{12}$$

where a nugget effect $nugg$ [Cressie, 1993, Pinheiro and Bates, 2000] is introduced to account for abrupt changes at very small distances.

Moreover, the homoscedastic model residuals (see Fig. 1, right) show evidence of heteroscedasticity. Possible drivers of heteroscedasticity should be sought in the geospatial characteristics of the phenomena. We thus model heteroscedasticity with the following function

$$g(s_{ij}, q_{ij}, \boldsymbol{\delta}) = \delta_{s_{ij}} \delta_{q_{ij}} \tag{13}$$

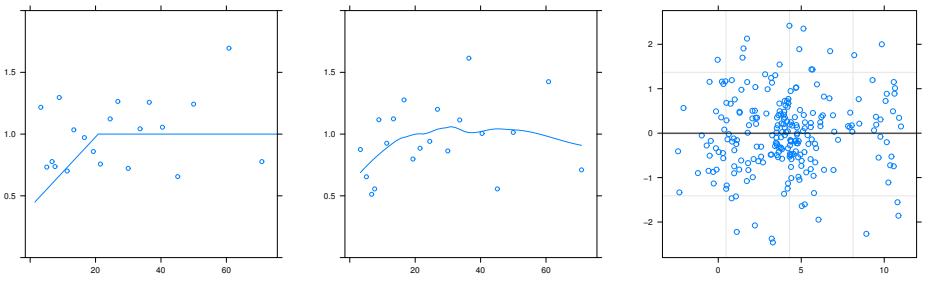

Figure 2: Model in Eq. (1) estimated assuming Gaussian within-group correlation as in Eq. (12) and heteroscedasticity as in Eq. (13). Left: semivariogram of within-group residuals; x-axis: distance; y-axis: semivariogram. Center: semivariogram of the normalized residuals; x-axis: distance; y-axis: semivariogram. Both the semivariograms have been presented by setting the maximum distance to 80 km. Right: fitted (x-axis) versus standardized (y-axis) residuals.

where s_{ij} is a dummy variable that takes value 1 if the ward is in the North Island (and 0 if it is in the South Island), q_{ij} is the number of regions (among the 16 into which New Zealand is divided) bordering the region to which the ward belongs, and $\boldsymbol{\delta}_{(\cdot)}$ is the subvector of variance parameters referring to the variable $(\cdot)$. The introduction of s_{ij} is suggested by looking at the wards with the highest (in absolute value) residuals, which are located on the North Island. In addition, since the correlation between TAs is assumed to be 0, we introduce the number of neighbouring regions q_{ij} to mitigate its possible role in heteroscedasticity.

Preliminary analyses of the regressors do not show problems related to multicollinearity. Results of the selected model, which is the model in Eq. (1) with both spatially correlated losses and heteroscedastic residuals (see Eq.s (12) and (13), respectively), are reported in Table 2. As one can notice, only two fixed effects appear not significant and the model has therefore been modified accordingly obtaining a final model with the both Akaike and Bayesian information criteria (AIC and BIC, respectively, hereafter) smaller. In particular, the intercept, the Cresta zone, the Z seismic risk score, the logarithm of the number of usual residents, the average weekly income, and the dwelling value appear good predictors of the $\log(Y)$. Not surprisingly, the average value of dwellings in the ward and the Z seismic risk score appear the main determinants of the value of losses. In contrast, the average weekly

Table 2: **Estimation results.** Model in Eq. (1) with within-group correlation as in Eq. (12) and heteroscedasticity as in Eq. (13). Column (A): all regressors as in Table 1; column (B): significant regressors only.

	Dependent variable:	
	log-Loss	
	(A)	(B)
Constant	-25.698^{***}	-27.835^{***}
	(5.299)	(5.287)
Cresta zone	0.383^{***}	0.391^{***}
	(0.044)	(0.044)
Z seismic risk score	5.603^{***}	5.432^{***}
	(1.712)	(1.667)
Density	0.0003	
	(0.0002)	
Housing with problems	-5.189	
	(4.367)	
log(Usual residents)	0.245^{***}	0.280^{***}
	(0.086)	(0.083)
Average weekly income	-0.011^{**}	-0.010^{**}
	(0.005)	(0.005)
Dwelling value	2.299^{***}	2.312^{***}
	(0.394)	(0.401)
	Random Effects	
$\hat{\sigma}_{u1}$	1.475	1.479
$\hat{\sigma}_{u2}$	0.0047	0.095
$\hat{\sigma}_{\varepsilon}$	0.588	0.544
	Within-group correlation	
$\hat{\mathrm{r}}$	20.538	1.984
$\widehat{nugg}$	0.656	0.071
	Heteroscedasticity ($\delta_{s_{ij}=0} = 1$, $\delta_{q_{ij}=4} = 1$)	
$\hat{\delta}_{s_{ij}=1}$	1.191	1.148
$\hat{\delta}_{q_{ij}=1}$	1.194	1.284
$\hat{\delta}_{q_{ij}=2}$	1.007	1.058
$\hat{\delta}_{q_{ij}=3}$	1.368	1.498
$\hat{\delta}_{q_{ij}=5}$	1.813	2.001
Log-restricted-likelihood	-355.319	-352.830
AIC	746.638	737.660
BIC	808.366	792.669
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: **Accuracy measures.** Comparison between the model in Eq. (1) with $\rho_{ij,i'j} = 0$ and homoscedasticity (columns 2 and 4) and the model in Eq. (1) with Gaussian spatial correlation as in Eq. (12) and heteroscedasticity as in Eq. (13) (columns 3 and 5). Columns 2 and 3 refer to the estimated losses at the ward level; the last two columns refer to the estimated losses at the TA level.

	log-Loss per ward		log-Loss per TA	
	$\rho_{ij,i'j} = 0$	$\rho_{ij,i'j}$ in Eq. (12)	$\rho_{ij,i'j} = 0$	$\rho_{ij,i'j}$ in Eq. (12)
RMSE	0.702	0.700	0.340	0.323
MAE	0.523	0.516	0.249	0.227
MAPE	43.496	44.599	7.716	7.462

income plays a much minor role. Using the standard UNDRR terminology[7], it is apparent that both measures of the *hazard* (Cresta zone and the Z seismic risk score), and measures of *exposure* matter for the determination of ward-level losses. $Vulnerability$, in as much as it is proxied by per capita income, however, does not seem to matter that much. This may be because the vulnerability of residential buildings to earthquake damage might not be that different across different wards with different income levels in the New Zealand context. In addition, figure 2 reports the semivariograms and the plot of residuals of the estimated model that clearly show the effectiveness in modelling spatial correlation and heteroscedasticity.

As for the model accuracy, we computed the root mean squared error (RMSE hereafter), the mean absolute error (MAE hereafter), and the mean absolute percentage error (MAPE hereafter) that compare the accuracy of the loss estimated using the proposed heteroscedastic spatially correlated model with those obtained with the basic homoscedastic uncorrelated model (see Table 3). As one can notice, the two models perform very similarly in terms of losses per ward and different accuracy measures support one model or the other. By contrast, all the accuracy measures suggest that the heteroscedastic spatially correlated model outperforms the basic one in estimating the losses at the TA level. However, one of the most relevant implications of our model is the impact of spatially correlated claims on the tail of loss distribution that we discuss in the next section.

[7] https://www.undrr.org/terminology.

5 Discussion

As noted above, we find that insured damage from earthquakes is, not surprisingly, spatially correlated across wards within a TA. In line with the existing literature [Kousky and Cooke, 2012, Cooke et al., 2010], we found that the correlation between wards in the insurance portfolio is on average extremely low. As reported in Table 4, the average correlation between wards in a TA is 0.0069.

Table 4: **Estimated spatial correlation between wards.** Estimates refer to the models presented in Table 2, column (B).

Spatial Correlation			
Mean (all wards)	Mean (wards in a TA)	3rd Qu. (wards in a TA)	Max.
0.0001	0.0069	2.088788e-09	0.5504

However, the low spatial correlation across wards is not immaterial to the identification of the tail of the distribution. As shown in Figure 3 and Table 5, the loss prediction obtained including spatial correlation quite satisfactorily approaches the observed log-losses both at the ward and at the TA level. In particular, our model is able to capture the right tail of the cumulative ward losses. This is a very satisfactory result since losses are strongly affected by the 2011 Canterbury earthquake sequence, which has been an extreme event.

Capturing the effect of spatial correlation improves the fitting of the exceedance probability curve $EP = P(AnnualLoss > x)$, which is used by insurers to determine the insolvency probability. Figure 4 shows the exceedance probabilities predicted by the homoscedastic non-spatially correlated model and by the heteroscedastic spatially correlated one. As one can easily notice, the uncorrelated model underestimates the aggregate expected losses, while the proposed spatial-correlation model better approaches the observed values. Spatial correlation particularly affects the prediction of losses with time to return greater than 20 years. These spatially correlated losses we identified in the right tail of the distribution are extremely important for determining the liquidity and solvency risks of an insurer, particularly when a large event occurs. They are also important in determining the amount of re-insurance that insurers should purchase, and have impli-

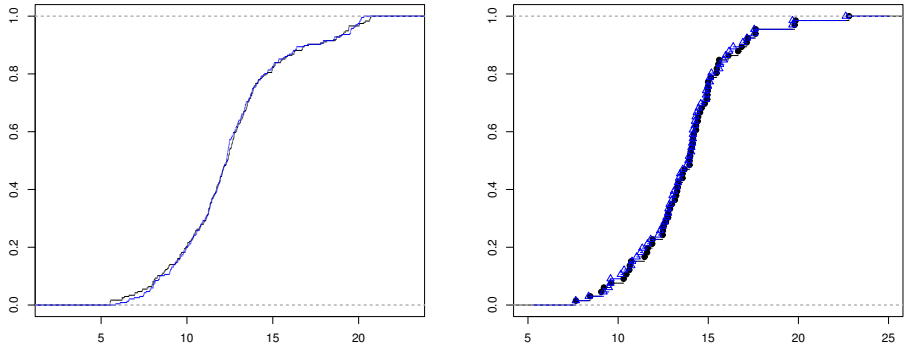

Figure 3: Comparison of empirical cumulative distribution functions of predicted (blue line) and observed (black line) $\log(y_{ij})$. Plots refer to predictions from the model in Eq. (1) with Gaussian within-group correlation as in Eq. (12) and heteroscedasticity as in Eq. (13). The x-axis denotes the values of $\log(y_{ij})$, the y-axis denotes the cumulative probability. Left: log-Loss at ward level. Right: log-Loss at TA level.

Table 5: **Loss right-tail quantiles.** Predicted values have been estimated by the model in Eq. (1) with Gaussian within-group correlation as in Eq. (12) and heteroscedasticity as in Eq. (13). Quantiles of the distribution of ward log-losses (columns 2 and 3) and quantiles of the distribution of TA log-losses (columns 4 and 5).

	log-Loss per ward		log-Loss per TA	
	Observed	Predicted	Observed	Predicted
90%	16.96	16.92	16.99	16.65
91%	17.94	17.91	17.14	16.96
92%	18.66	18.55	17.18	17.10
93%	18.80	18.79	17.39	17.33
94%	19.10	19.44	17.62	17.55
95%	19.29	19.54	17.63	17.55
96%	19.43	19.65	18.49	18.40
97%	20.00	19.84	19.79	19.67
98%	20.32	20.04	19.83	19.67
99%	20.65	20.13	20.88	20.71
100%	20.72	20.31	22.81	22.62

cations for any prudential regulatory practice of the insurance industry[8].

[8]In New Zealand, macro-prudential regulation of the insurance sector is the responsibility of the central bank (the Reserve Bank of New Zealand).

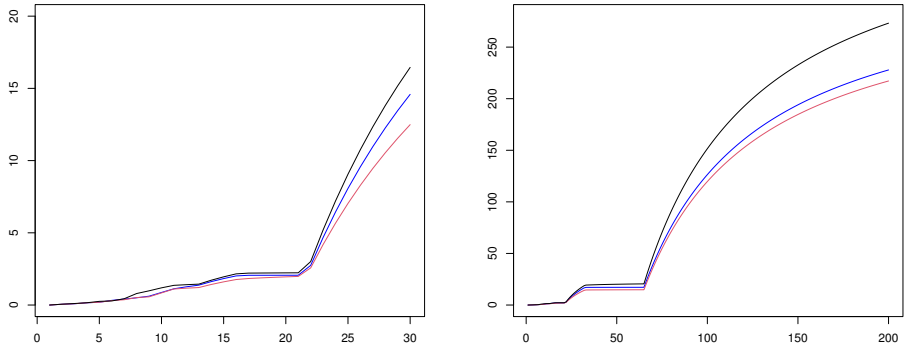

Figure 4: Observed exceedance probability curves (black lines) versus exceedance probability curves predicted using the model in Eq. (1) with $\rho_{ij,i'j} = 0$ and homoscedasticity (red lines), and exceedance probability curve predicted using the model in Eq. (1) with within-group correlation as in Eq. (12) and heteroscedasticity as in Eq. (13) (blue lines). The x-axis denotes the time to return $RT = \frac{1}{EP}$ and the y-axis denotes the total annual millions of losses in the Country. Left: RT up to 30 years; right: RT up to 200 years. Please note that the y-axis range varies among the plots to better show the curves.

By ignoring this very small spatial correlation, insurers and their regulator might be underestimating the risk of insolvency. This may have further repercussions for government budgets (who are often the implicit insurer-of-last-resort) and for long-term recovery trajectories of disaster-affected regions, in as much as recovery is dependent on the available funding from insurance claims.

6 Conclusion

The purpose of our modelling was to examine the feasibility of quantifying the likely insurance liability (or the level of risk-based insurance premiums) that can be estimated from existing claims data. We find that accounting for the spatial correlation in earthquake damages increases the predictive power of our mixed-effects model. Without accounting for these spatial correlations and heteroscedasticity, estimates based on historical data will under-estimate the risk, and therefore the exceedance probability curve. The model we developed is thus able to fit the data much better than the

equivalent and more standard homoscedastic non-spatial model. It is clear that, should the public insurer in New Zealand decide to start charging risk-based premiums, they should use a model that accounts for these features of the data. This type of model will also be useful to assess liability and risk more broadly, and thus price any reinsurance contracts that the insurer purchases.

More generally, the modelling approach we propose can be used in other instances in which an insurer is considering moving to risk-based pricing, or in other situations where risk-based pricing is more appropriate for supplying catastrophic natural hazard insurance cover.

Acknowledgements

The first author acknowledges the financial support from the Italian Ministry of University and Research (MIUR) under the Research Project of National Interest (PRIN) grant 2017TA7TYC. The second and the third author thank the New Zealand Resilience National Science Challenge for funding various stages of this work, and the EQC, and specifically Geoff Spurr, for supplying the data and helping us decode it.

References

N. Barbieri, M. Mazzanti, A. Montini, and A. Rampa. Risk Attitudes to Catastrophic Events: VSL and WTP for Insurance Against Earthquakes. *Economics of Disasters and Climate Change*, 6(2):317–337, 2022.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90, 1970.

R. Cooke, C. Kousky, and H. Joe. Micro correlations and tail dependence. In Kurowicka and Joe, editors, *Dependence Modeling: Handbook on Vine Copulae*. World Scientific, Singapore, 2010.

N. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

N. Cressie and C. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2011.

O. Filippova, C. Nguyen, I. Noy, and M. Rehm. Who Cares? Future sea-level-rise and house prices. *Land Economics*, 96(2):207–224, 2020.

R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, NY, USA, second edition, 1987.

D. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–340, 1977.

H. Ito and R. N. McCauley. A disaster under-(re)insurance puzzle: Home bias in disaster risk-bearing. *IMF Economic Review*, 2022.

C. Kousky. The role of natural disaster insurance in recovery and risk reduction. *Annual Review of Resource Economics*, 11(1):399–418, 2019.

C. Kousky and R. Cooke. Explaining the failure to insure catastrophic risks. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 37(2): 206–227, Apr 2012. ISSN 1468-0440.

H. Kunreuther and E. Michel-Kerjan. Chapter 11 - economics of natural catastrophe risk insurance. In M. Machina and K. Viscusi, editors, *Handbook of the Economics of Risk and Uncertainty*, volume 1 of *Handbook of*

the Economics of Risk and Uncertainty, pages 651–699. North-Holland, 2014.

N. Laird and J. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

Lloyd's. *A World At Risk: Closing the Insurance Gap*. Lloyds, London, 2017.

C. Nguyen, I. Noy, D. E. Sommervoll, and F. Yao. Redrawing of a housing market: Insurance payouts and housing market recovery in the wake of the christchurch earthquake of 2011. CESifo Working Paper Series 8560, CESifo, 2020.

C. N. Nguyen and I. Noy. Comparing earthquake insurance programmes: how would japan and california have fared after the 2010-11 earthquakes in new zealand? *Disasters*, 44(2):367–389, 2020.

S. Owen, I. Noy, J. Pastor-Paz, and D. Fleming. Measuring the Impact of Insurance on Recovery after Extreme Weather Events Using Nightlights. *Asia-Pacific Journal of Risk and Insurance*, 15(2):169–199, 2021.

H. D. Patterson and R. Thompson. Recovery of interblock information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

J. Pinheiro and D. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:298–296, 1996.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

R. Schwarze and C. Croonenbroeck. Economies of integrated risk management? an empirical analysis of the swiss public insurance approach to natural hazard prevention. *Economics of Disasters and Climate Change*, 1(2):167–178, 2017.