Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 1 Issue 2 December, 2024, pp. 65-70
https://idjams.com/index.php/jms/index

Modelling Claim Frequency with Spatial Effects for Accurate Insurance Premium Cost Calculations

Didymus Ibrahim Nyamweya [1,*], Jared Mong'are [2]

[1] *Institut Teknologi Bandung, Jl. Ganesha No. 10, Bandung, 40132, Indonesia*

[2] *University of Eldoret, Eldoret, PO Box 1125-30100, Kenya*

[*] *Corresponding author: didyibra@gmail.com*

Copyright: ©2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License-(CC-BY-SA) (https://creativecommons.org/licenses/by-sa/4.0/)

Article Info

Article history:
Received: July 2024
Revised: October 2024
Accepted: October 2024
Available online: December 2024

Abstract

This paper challenges the traditional assumption of independence between claim counts and amounts in non-life insurance. It explores the effectiveness of Generalized Linear Models (GLMs) in analyzing claim frequency data, a key component of accurate premium pricing. The proposed approach utilizes GLMs for both the marginal frequency and conditional severity of claims. Dependence between these factors is introduced by incorporating the number of claims as a covariate in the severity model. This strategy offers ease of implementation and interpretability, particularly when combined with Poisson counts and a log-link function. The resulting pure premium calculation considers the marginal mean frequency, a modified severity, and a dependence correction term. The paper further establishes the importance of spatial factors in claim frequency modelling for insurance businesses. It proposes a novel GLMM with a CAR (Conditional Autoregressive) model to account for these spatial effects. The impact of spatial factors on pure premium calculations is evaluated using simulated claim data.

Keywords: Claim Frequency; Generalized Linear Models; Conditional Autoregressive; Spatial Effects; Pure Premium

I. Introduction

The paper aims to investigate the impact of coronavirus disease (COVID-19) on claim severity and pure premium cost calculations. In this preliminary research the treatment cost is incorporated into an existing dataset. Treatment costs for individual COVID-19 cases will be included in the severity calculations, which are crucial for determining accurate pure premiums. The cost reference is based on a study by Jati et al. [1] which estimated treatment costs for COVID-19 patients in Indonesia [2]. Their research suggests a range of 130.4 - 133.2 million rupiah per patient, assuming a 14-day ICU stay followed by 14 days in a non-ICU ward, with remaining hospitalization costs following INACB guidelines [3]. The dataset used in this research encompasses 263,072 cumulative infected cases documented from February 2020 to April 2021. Generalized linear mixed-effect models (GLMM) offer a solution to this issue by meeting normality assumptions without requiring data transformation. This approach enables accurate assessment of differences between datasets, using the most suitable metric for the theoretical framework employed in this paper.

II. Methods

We utilize a dataset encompassing 263,072 cumulative infected cases from February 2020 to April 2021, specifically focusing on patients residing within the West Java Province of Indonesia. This province is further divided into 28 sub-regions (Kapubaten).

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 1 Issue 2 December, 2024, pp. 65-70
https://idjams.com/index.php/jms/index

A generalized linear model (GLM) refers to a class of linear regression models in which the response vector Y comes from an exponential family of distributions, with mean μ and its relationship to the covariates becomes linear through a link function which need not be the identity map. Generalized linear mixed models are used in cases where observations are non-normal and correlated.

The dataset provides valuable insights into various aspects of the COVID-19 pandemic within the region. It includes details such as the geographical coordinates of patients' residences, the dates of reported cases, daily death tolls, recoveries, and cumulative case counts. Due to potential privacy concerns, the dataset lacks information on individual demographics like age, gender, social class, or education level. We leverage the available descriptive variables, namely longitude and latitude, to explore spatial trends in claim severity and their implications for insurance pricing.

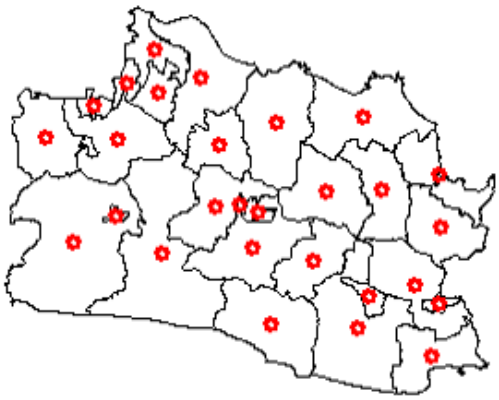

Figure 1. Map of West Java showing the centre points of the various districts

Graph data as adopted from Point Pattern method on data source [4].

Once the adjacency matrix has been formed from the graph data, implementation of the simulation follows in the R software. The matrix is constructed taking into consideration the total number of neighbours a sub-region. The number of neighbours is represented by w_i and ψ is the hyper parameter which shows the degree of dependence in the model, the spatial effects are modelled by the Conditional Autoregressive (CAR) model [5];

$$\boldsymbol{\gamma} \sim \text{MVN}(0, \sigma^2 Q^{-1}) \tag{1}$$

where Q specified as

$$Q_{ij} = \begin{cases} 1 + |\psi| * w_i, & i = j \\ -\psi, & i \neq j, i \sim j \\ 0, & otherwise \end{cases} \tag{2}$$

Spatial gamma($\boldsymbol{\gamma}$) random effects generated from R software [4].

Having estimated the gamma spatial effects, implement them in the Poisson GLMM model which has a mean given as

$$\mu_i = \exp(X_i'\beta + \gamma_{j(i)}) \tag{3}$$

where the variable are as discussed. The model above based on the dataset will have the following components:

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 1 Issue 2 December, 2024, pp. 65-70
https://idjams.com/index.php/jms/index

1. Response vector N is a 28-dimensional vector, which contains the covid count for each of the sub region in West Java (Tufvesson, et. 2019).
2. The predictors are in the form of a 28*2 design matrix X which include for the 28 sub regions.
   a. The latitude position of each region X_1
   b. The longitude position of each region X_2
3. The random component of the model is assumed to be a vector if gamma spatial which have already been calculated in the previous subtopic.

Estimation of the unknown beta parameters can be carried out in R software by use of the various packages found in R. In this simulation we mainly applied two of them and compared the results, the estimates are then used to compute the μ_i the 35 average claim necessary in premium determination.

III. Results and Discussion

Conditional Autoregressive Model (CAR) and the Poisson GLMM have been applied to a real dataset [8]. The results from the estimates of claim frequency and claim severity are then used to assess the importance of spatial factors in claim frequency modelling and severity calculations [6].
This paper incorporates treatment costs into the analysis. Treatment cost estimates for Indonesian COVID-19 patients, based on research by Jati et al, range from 130.4 to 133.2 million rupiah per case [7]. These estimates assume a 14-day ICU stay followed by 14 days in a non-ICU ward, with remaining hospitalization costs following INACB guidelines. The dataset employed in this research encompasses 263,072 confirmed COVID-19 cases documented in West Java Province, Indonesia, between February 2020 and April 2021. As of the final observation date, there were 39,667 active cases and a cumulative death toll of 2,188 within the province.

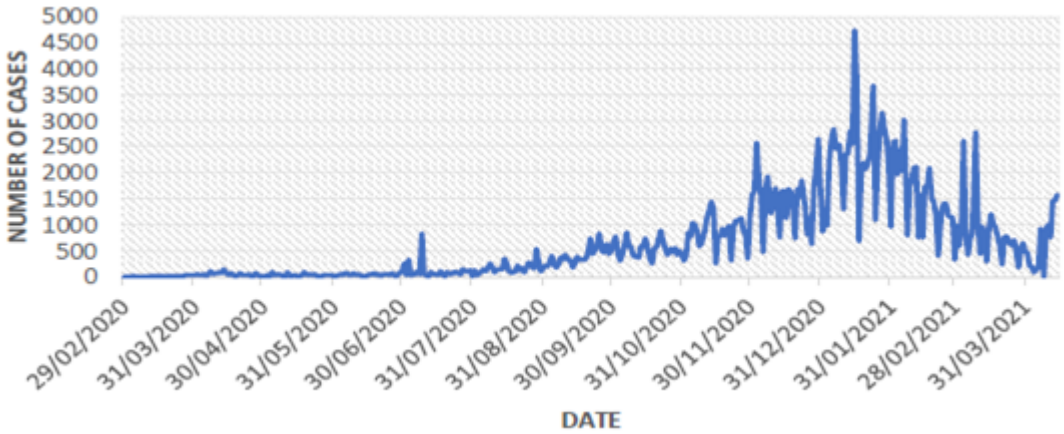

Figure 2. Graph of COVID-19 Cases in West Java

Initial calculations for pure premium in this paper relied on claim frequency derived from a negative binomial discrete distribution [2] [7]. This approach assumed uniform claim frequency and severity across all regions within the study area [8]. While this simplifies the initial analysis, it acknowledges the limitations of available data at the time. Severity was also assumed to be consistent geographically, reflecting a belief that treatment costs wouldn't vary significantly between sub-regions.
A random variable N has a Negative Binomial distribution if

$$N \sim NB(\beta, r) \tag{4}$$

It has a probability mass function of

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 1 Issue 2 December, 2024, pp. 65-70
https://idjams.com/index.php/jms/index

$$p_k = \Pr(N = k) = \binom{n+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k, \quad (5)$$

for $k = 0,1,2,\ldots, where\ r > 0, \beta > 0$. The mean and variance of negative binomial is given by mean $E(N) = r\beta$ and variance $Var(N) = r\beta(1+\beta)$. Fitting Negative Binomial distribution using fitdistrplus package in R.

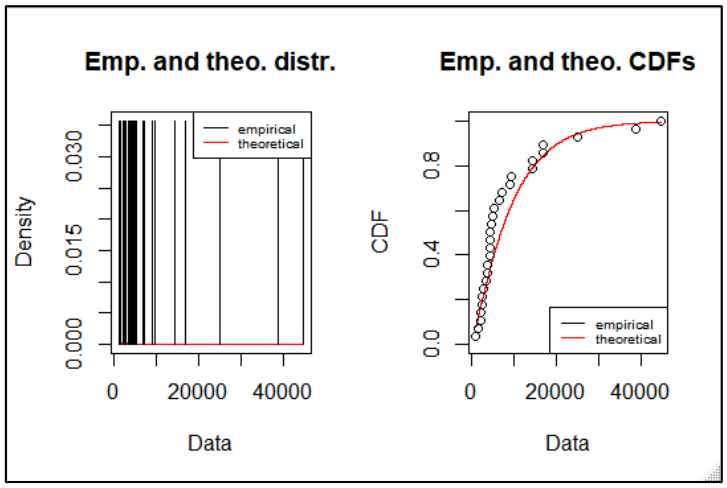

Figure 3. Plot of Negative Binomial Distribution Fit

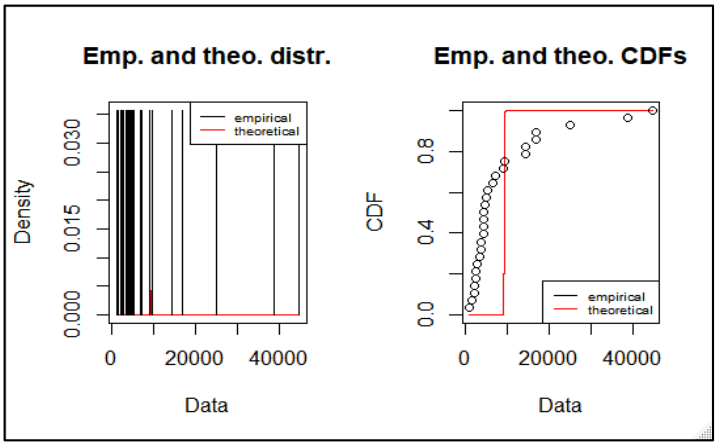

Figure 4. Plots of Poisson Fit to Cases

Comparing the two pure premium estimates, the model incorporating spatial random effects yields a demonstrably better assessment. This model, accounting for spatial variations, projects a total premium collection of 40,990,433 million Rupiah compared to 34,708,636 million Rupiah for the model without spatial effects. This translates to a significant increase, allowing the company to better prepare for unforeseen events compared to the non-spatial model.

To begin with, before fitting the various distributions we plot the histogram and empirical distribution function using the *plotdist* function of the package [9].

Indonesian Journal of Applied Mathematics and Statistics (IdJAMS)
ISSN 3064-4062
Volume 1 Issue 2 December, 2024, pp. 65-70
https://idjams.com/index.php/jms/index

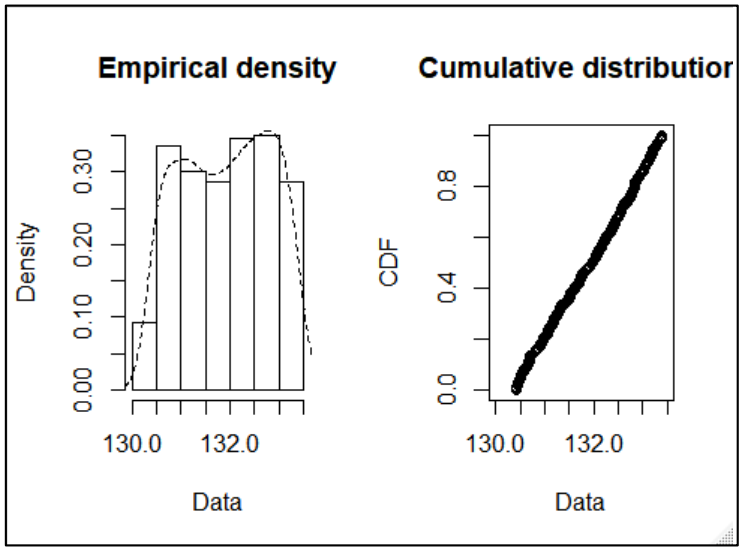

Figure 5. CDF and Histogram Plots from the Plotdist Function for Cost of Treatment

IV. Conclusion

To achieve desired financial stability: Applying this spatial model to an insurance company in West Java Province, Indonesia, would lead to higher premium collections, improving the company's ability to manage unforeseen events and maintain financial stability. However, limitations arise due to the criteria for defining neighboring regions, as administrative boundaries may not perfectly align with proximity or development similarities (biased estimate). Additionally, the unpredictable spread of pandemics complicates the modelling of average claims [10]. By accounting for regional variations, an insurer can set competitive rates that attract customers while ensuring financial sustainability under unseen circumstances. Future research could focus on refining the methods for defining neighboring regions and addressing issues in claim severity modelling to enhance the results further.

References

[1] S. P. Jati, R. T. Budiyanti, P. Ginandjar, and N. Nandini, "Cost Estimates Related to COVID-19 Treatment in Indonesia: What Should be Concerned?", *E3S Web Conf.*, vol. 202, pp. 1-8, 2020, DOI: 10.1051/e3sconf/202020212012

[2] S. Gschlößl and C. Czado, "Modelling Count Data with Overdispersion and Spatial Effects." *Stat. Pap.*, vol. 49, pp. 531-552, 2008.

[3] J. J. Faraway, "*Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models by Faraway*." CRC Press, 2016.

[4] S. Banerjee, B. Carlin, and A. Gelfand, "*Hierarchical Modeling and Analysis for Spatial Data*," Second Edition. CRC Press, 2015.

[5] A. N. Pettitt, I. S. Weir, and A. G. Hart, "A Conditional Autoregressive Gaussian Process for Irregularly Spaced Multivariate Data with Application to Modelling Large Sets of Binary Data," *Stat. Comput.*, vol. 12, no. 4, pp. 353– 367, 2002, DOI: 10.1023/A:1020792130229

[6] L. Yang, "Nonparametric Copula Estimation for Mixed Insurance Claim Data," *J. Bus. Econ. Stat.* , vol. 40, no. 2, pp. 537-546, 2022, DOI: 10.1080/07350015.2020.1835668

[7] O. Tufvesson, J. Lindström, and E. Lindström, "Spatial Statistical Modelling of Insurance Risk: a Spatial Epidemiological Approach to Car Insurance," *Scand. Actuar. J.*, vol. 2019, no. 6, pp. 508–522, 2019, DOI: 10.1080/03461238.2019.1576146

[8] T. M. Brown, W. H. Pogorzelski, and I. M. Giammanco, "Evaluating Hail Damage using Property Insurance Claims Data," *Weather Clim. Soc.*, vol. 7, no. 3, pp. 197-210, 2015, DOI: 10.1175/WCAS-D-15-0011.1

[9] M. L. Delignette-Muller and C. Dutang, "An R Package for Fitting Distributions," *J. Stat. Softw.*, vol. 64, no. 4, pp. 1-34, 2015, DOI: 10.18637/jss.v064.i04

[10] H. H. Wang and H. Zhang, "On the Possibility of a Private Crop Insurance Market: A Spatial Statistics Approach." *J. Risk Insur.*, vol. 70, no. 1, pp. 111-124, 2003.