Resumen de trabajo de tesis

Errores en las variables regresoras del modelo Besag-York-Mollié

M. Fernanda Vallejos López*

Graduada de Magíster en Bioestadística, Escuela de Salud Pública
Facultad de Medicina, Universidad de Chile

El aumento de la información georreferenciada y el desarrollo de potencia computacional y de *softwares* estadísticos ha permitido la implementación de metodologías mas complejas para estimación en áreas pequeñas (SAE por sus siglas en inglés: *Small Area Estimation*). tal es el caso del modelo Besag-York-Mollié (BYM), un modelo jerárquico bayesiano utilizado para estimar el riesgo relativo en áreas pequeñas, caracterizado por considerar la autocorrelación espacial en la descripción del modelo. Muchas veces, junto a la estimación del riesgo relativo, puede ser de interés explorar la asociación de la variable respuesta con ciertas covariables y evaluar su efecto en el modelo, para lo cual se asume que las variables están medidas sin error. Sin embargo, esto puede no ser cierto. En este trabajo se exploró el efecto que tiene la adición de error en las covariables en la estimación del riesgo relativo del modelo BYM2. El modelo BYM y BYM2 (Simpson et al., 2017), una reparametrización del BYM, son ampliamente utilizados en disciplinas como epidemiología y ecología (Seaton et al., 2024).

El modelo BYM es un modelo log-lineal compuesto por tres niveles. En el primer nivel, el número de casos observados en un área específica i (y_i) se modela como una variable discreta de tipo conteo utilizando una distribución de Poisson (Blangiardo & Cameletti, 2015):

$$y_i \sim \text{Poisson}(E_i\theta_i), \tag{1}$$

donde E_i representa el número esperado de casos en el área i, basado en la población regional como referencia, y θ_i es el riesgo relativo verdadero y desconocido en el área i.

En el segundo nivel, se incorporan efectos aleatorios para modelar la heterogeneidad espacial estructurada y no estructurada. El logaritmo del riesgo relativo se representa como:

$$\log(\theta_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + u_i + v_i, \tag{2}$$

donde β_0 es la tasa promedio del riesgo relativo (intercepto fijo), $\mathbf{x}_i$ es el vector de covariables del área i, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión (efectos fijos), u_i es el efecto espacial aleatorio correlacionado (heterogeneidad estructurada), y v_i es el efecto aleatorio no correlacionado (heterogeneidad no estructurada).

Los efectos aleatorios tienen distribuciones específicas: v_i se asume como una variable con distribución normal de media 0 y varianza σ_v^2, mientras que u_i sigue una distribución autorregresiva intrínseca (ICAR) dependiente de la matriz de adyacencia, que refleja la relación de vecindad entre áreas.

Para evaluar la robustez del modelo BYM2 en la estimación de riesgo relativo frente a errores en las covariables, se ajustó un modelo BYM en un conjunto de datos real proveniente de repositorios públicos nacionales, en el cual una de las covariables tenia un efecto positivo significativo y la otra tenía un efecto no significativo. Para ello se realizó una simulación en paralelo y se ajustó el modelo mediante el software R y Rstudio (RStudio Team, 2020) con el paquete INLA (Martino & Rue, 2009), el cual ajusta modelos jerárquicos bayesianos mediante aproximaciones anidadas integradas de Laplace. Con el objetivo de evaluar el impacto de los errores en una variable significativa y en otra no significativa para la estimación, se diseñó un estudio de simulación factorial con 9 escenarios, cada uno con diferentes grados de error adicional en las covariables. En cada escenario, se simularon 5000 conjuntos de datos. El error adicional está basado en los hallazgos del análisis exploratorio de datos, principalmente en la desviación estándar de cada una de las covariables, de modo que se tendrá un escenario sin error, otros en los que tenemos un error adicional con una desviación estándar similar a la encontrada en el conjunto de datos inspiracional y otro con un grado de error un poco más extremo, como se visualiza en la Figura 1.

Los principales hallazgos tienen que ver con la distribución de muestreo de los coeficientes de efectos fijos del modelo, los cuales quedaron subestimados en todos los escenarios. El impacto de los errores en la variable con un efecto significativo (covariable 1) es notorio, en cambio el efecto de la variable que no tiene un efecto significativo (covariable 2) no es claro.

En los efectos aleatorios, la precisión, equivalente al inverso de la varianza, fue mayor en los modelos con menor error en la covariable significativa. Y el porcentaje de variabilidad atribuible al componente espacialmente estructurado fue similar en todos los escenarios de simulación (Tabla 1).

La estimación de riesgo relativo fue similar en todos los escenarios de simulación. Como se observa en la Figura 2, el escenario 9, el cual tiene un mayor error adicional en las covariables, la media estimada para

*Matrona, Unidad de Medicina Reproductiva, Clínica Alemana de Santiago, fda.vallejos@gmail.com.

el riesgo relativo de ese escenario fue similar a la obtenida en el 1 el cual no tiene error adicional en sus covariables.

Nivel de error covariable 1 \ Nivel de error covariable 2	0	0.002	0.005
0	Escenario 1	Escenario 2	Escenario 3
4	Escenario 4	Escenario 5	Escenario 6
6	Escenario 7	Escenario 8	Escenario 9

Figura 1: Diseño de escenarios de simulación: cada escenario tiene distintos niveles de error adicional en sus covariables, escenario 1 es simulado sin componente de error adicional, escenario 9 es diseñado con una variable de error adicional con una media 0 y desviación estándar de 6 en su covariable 1 y un componente de error adicional de media 0 y desviación estándar de 0.005 en su covariable 2.

Se evaluó el desempeño del modelo para la estimación del riesgo relativo mediante el sesgo absoluto, error cuadrático medio (MSE) y cobertura empírica del intervalo de credibilidad (de 95 % de credibilidad). Para el sesgo y el MSE no se observan grandes diferencias entre escenarios, independiente del escenario de simulación, los errores tienden a concentrarse en las mismas comunas y en un grado similar, la cobertura empírica del intervalo sucede algo similar, se observan diferencias de cobertura entre comunas y no entre los escenarios de simulación.

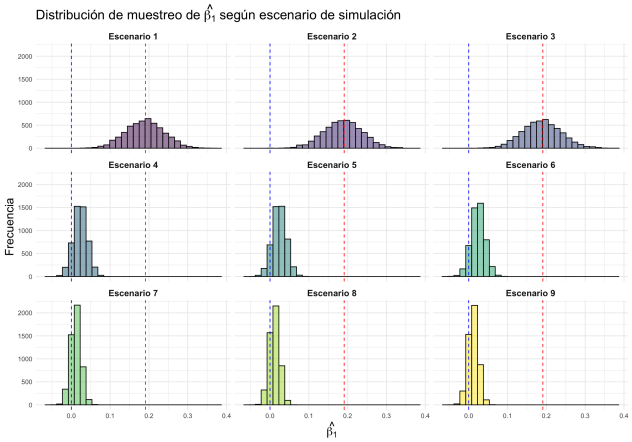

Figura 2: Distribución de muestreo de coeficiente de covariable 1 ($\hat{\beta}_1$) según escenario de simulación. Línea roja representa el valor del coeficiente en el conjunto de datos real; línea azul indica el punto donde el coeficiente toma el valor de cero, correspondiente a un efecto nulo de la covariable.

Tabla 1: Media estimada de los parámetros de efectos aleatorios del modelo BYM2: τ (precisión) y ϕ (proporción de variabilidad atribuida al componente espacial) calculada a partir de conjuntos de datos simulados.

Escenario	τ	ϕ
1	4.34	0.614
2	4.34	0.614
3	4.34	0.617
4	3.24	0.624
5	3.25	0.619
6	3.24	0.620
7	3.19	0.629
8	3.20	0.629
9	3.18	0.630

El desempeño estadístico del modelo BYM2 medido con el sesgo relativo, MSE y cobertura empírica del valor real, del modelo BYM2 en cuanto a la estimación del riesgo relativo es buena en el contexto de error en las variables regresoras, no hay grandes diferencias entre lo obtenido en escenarios sin errores de los escenarios con errores. Sin embargo la interpretación a los efectos fijos puede estar errada. En conclusión el modelo BYM es un modelo robusto frente a los errores en las variables regresoras, sin embargo esos errores pueden llevar a mal interpretar el efecto de las covariables en el modelo, quedando subestimadas.

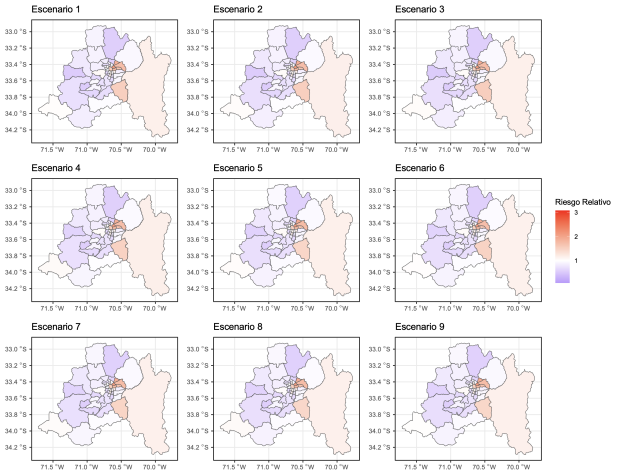

Figura 3: Riesgo relativo de postergación de la maternidad en comunas de la Región Metropolitana según escenario de simulación, se observan valores similares en todos los escenarios.

Literatura Recomendada

- Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman and Hall/CRC.
- Morris, Tim P., White, Ian R. y Crowther, Michael J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11), 2074–2102. doi: 10.1002/sim.8086.

Información adicional

Director: Prof. Felipe Medina Marín. Programa de Bioestadística, Escuela de Salud Pública, Universidad de Chile.

Fecha de la graduación: 20 de noviembre de 2024.

Referencias

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Martino, S., & Rue, H. (2009). *Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program*. Department of Mathematical Sciences, NTNU, Norway. http://www.math.ntnu.no/hrue/GMRFLib

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

Seaton, F. M., Jarvis, S. G., & Henrys, P. A. (2024). Spatio-temporal data integration for species distribution modelling in R-INLA. *Methods in Ecology and Evolution*, *15*, 1221-1232. https://doi.org/10.1111/2041-210X.14356

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1-28. https://doi.org/10.1214/16-STS576