

Inferencia en datos espaciales con respuestas censuradas o faltantes usando ecuaciones diferenciales parciales estocásticas

PROBLEMA

En 2024, Ecuador registró 5.815 incendios forestales, destruyendo 83.323,33 hectáreas de cobertura vegetal y causando pérdidas humanas, animales y materiales, además de un grave impacto ambiental (Secretaría Nacional de gestión de Riesgos (Infografía Nacional No.87), 2024), adicionalmente, la modelación estadística enfrenta desafíos como la presencia de datos faltantes o censurados y altos costos computacionales, complicando la gestión de riesgos.

OBJETIVO GENERAL

Desarrollar un modelo estadístico espacial para la predicción de incendios forestales en Ecuador, que aborde los problemas de datos faltantes o censurados, y que reduzca los altos costos computacionales. Este modelo incorporará una nueva metodología para analizar grandes volúmenes de datos, superando las limitaciones actuales en el estudio de este fenómeno.

PROPUESTA

Se propone una nueva metodología, empleando algoritmos del tipo EM (Expectation-Maximization) para lidiar con datos censurados o faltantes y una aproximación del campo aleatorio gaussiano (GRF) $Y(S)$ por el campo aleatorio gaussiano markoviano (GMRF) $\tilde{Y}(S)$, obtenido a partir de la solución a la ecuación diferencial parcial estocástica (SPDE)

$$(\phi^{-2} - \nabla)Z(S) = 4\pi W(S)$$

para optimizar el análisis de grandes volúmenes de datos.

Modelo propuesto

$$\tilde{Y}(S) = \mathbf{X}\beta + \tilde{Z}(S) + \epsilon(S)$$

$$\mathbf{X} \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^{p \times 1}$$

$$\tilde{Z}(S) \sim N_n(\mathbf{0}, \sigma^2 \gamma \mathbf{A} \mathbf{Q}_\phi^{-1} \mathbf{A}^\top)$$

$$\epsilon(S) \sim N_n(\mathbf{0}, \sigma^2 (1 - \gamma) \mathbf{I}_n)$$

$$\mathbf{Q}_\phi = \frac{\phi^2}{4\pi} \left[\frac{1}{\phi^4} \mathbf{D} + \frac{2}{\phi^2} \mathbf{G}_1 + \mathbf{G}_2 \right]$$

Donde

$$\tilde{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \Psi)$$

$$\text{con } \Psi = [\gamma \mathbf{A} \mathbf{Q}_\phi^{-1} \mathbf{A}^\top + (1 - \gamma) \mathbf{I}_n]$$



Fotografía 1. Incendio forestal activo en la cumbre del Pichincha (tomado de El Comercio <https://www.elcomercio.com/actualidad/quito/incendio-forestal-quito-activo-faldas-pichincha.html>).

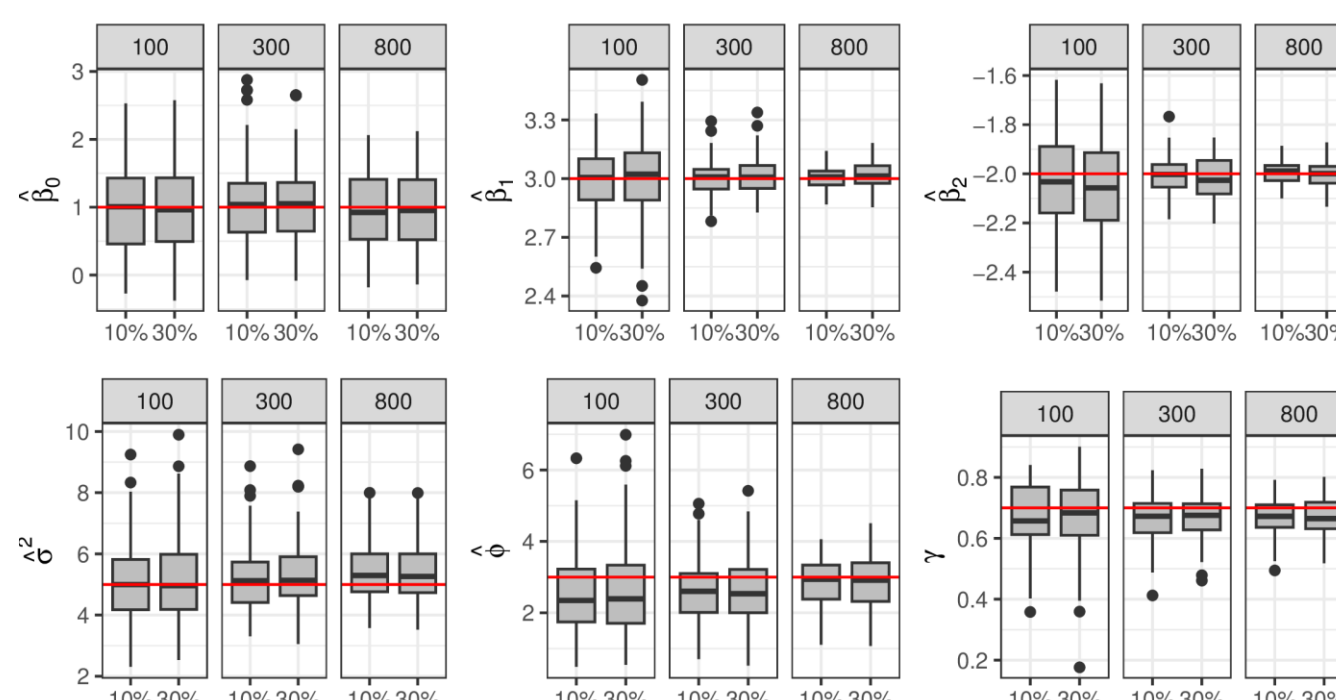


Figura 1. Boxplot de las estimaciones de los parámetros del modelo, considerando 100 muestras de tamaño $n = 100, 300$ y 800 con 10% y 30% de censura a la izquierda y valores faltantes. La línea roja representa el verdadero valor del parámetro.

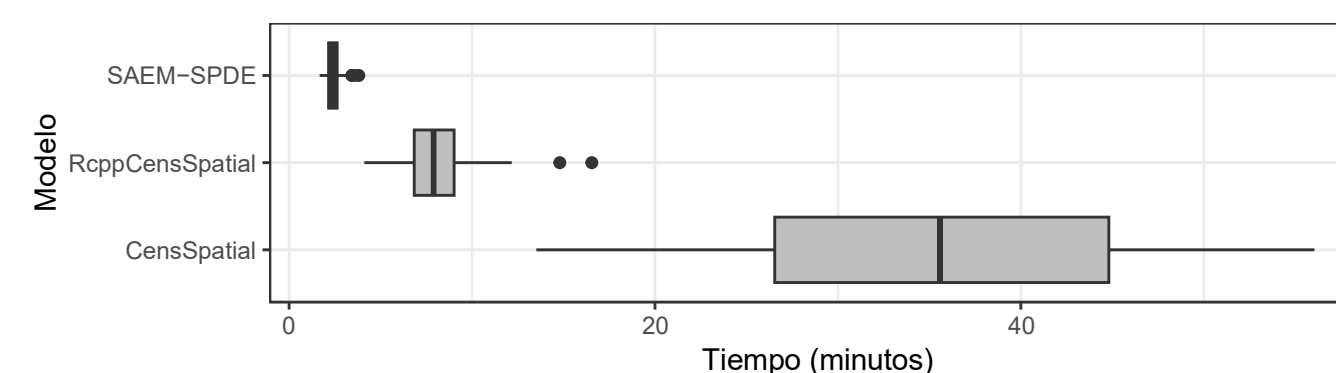


Figura 2. Boxplot de los tiempos de ejecución (en minutos) del modelo propuesto (SAEM-SPDE) y las propuestas en las librerías de R, RcppCensSpatial y CensSpatial, para estimar los parámetros del modelo espacial con respuestas censuradas en muestras de tamaño 300 con 10% de censura a la derecha.

RESULTADOS

Con la metodología propuesta se logró estimar un modelo, empleando las coordenadas como covariables para los niveles de afectación por incendios forestales con alrededor de un 10% de datos faltantes (ver Tablas 1 y 2). Además, las predicciones generadas para la provincia de Pichincha permitieron identificar las zonas de mayor afectación (ver Figura 3), proporcionando un marco metodológico para la implementación de medidas preventivas basadas en evidencia y una gestión eficiente de los recursos frente a la problemática de los incendios forestales.

Parámetro	Estimado	SE
β_0	-62.948	26.289
β_1	0.048	0.021
β_2	-0.017	0.020
σ^2	10.996	1.081
ϕ	7.814	3.61
γ	0.3	0.068

Tabla 1. Parámetros estimados del modelo espacial y error estándar (SE) obtenida a partir de la matriz de Información estimada.

Criterio	Valor
LogLik	-4492.453
AIC	8996.907
BIC	9030.604

Tabla 2. Criterios de selección de modelos.

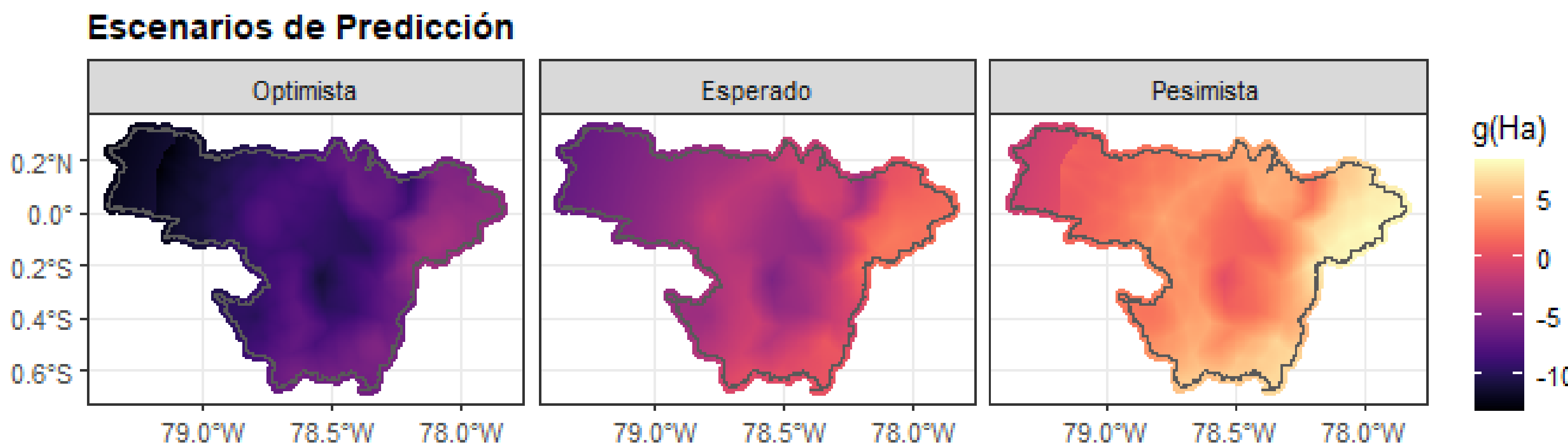


Figura 3. Predicción de niveles de afectación por incendios forestales. El escenario optimista (izquierda) está representado por el límite inferior del intervalo de predicción, el escenario esperado (centro) es representado por los valores predichos, mientras que el escenario pesimista (derecha) está representado por el límite superior del intervalo de predicción. Se utilizó un nivel de credibilidad del 95%.

CONCLUSIONES

- El modelo propuesto se destaca frente a las metodologías actualmente disponibles para lidiar con datos incompletos, como RcppCensSpatial y CensSpatial.
- Las regiones con mayores niveles de afectación se concentran mayoritariamente en la frontera oeste de la provincia de Pichincha, donde hay mayor densidad arbórea dada su conexión con la Amazonía y gran extensión de parques forestales.

El valor estimado del parámetro de escala ϕ nos indica que a distancias mayores de 31.25 km, los niveles de afectación tienen una correlación menor a 0.05.

Los niveles de afectación presentan una variabilidad total de 10.996 unidades cuadradas. De esta, un 30% es atribuible a la distribución natural y local de los niveles de afectación. Por otro lado, el 70% de la variabilidad restante corresponde a factores externos, vinculados a la aleatoriedad no espacial del evento.