UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Noviembre 2017

Estimación de un modelo lineal generalizado mixto para datos de conteo con exceso de ceros.

Estimation of a generalized linear mixed model for counting data with excess zeros.

Edinson Javier Díaz Prieto[a]
edinsondiaz@usantotomas.edu.co

Dagoberto Bermúdez Rubio[b]
dagobertobermudez@usantotomas.edu.co

Wilmer Pineda Ríos[c]
wilmerpineda@usantotomas.edu.co

Resumen

En este documento se presentan los análisis realizados a partir del ejercicio de modelación de la cantidad de incendios forestales registrados en los 212 municipios que componen los departamentos de Antioquia y Santander durante el año 2013. Para dicho análisis se empleó el método de modelos lineales generalizados mixtos inflados de ceros en el que se estimó el efecto de diferentes variables climáticas y geográficas medidas en cada municipio. El proceso de estimación se realizó mediante modelos aditivos generalizados de ubicación, escala y forma, en el cual se obtuvieron los parámetros mediante estimaciones lineales generalizas, estimaciones no paramétricas y una estimación de retardo espacial mediante modelos SAR.

***Palabras clave*:** Modelos Mixtos, Modelos Lineales Generalizados, Incendios Forestales, Datos de Conteo, Modelos Cero Inflados.

Abstract

This document presents the analyzes carried out in the 212 municipalities that make up the departments of Antioquia and Santander during 2013. For this analysis, the method of zero inflated generalized linear models was used in which the effect of different climatic and geographical variables measures in each municipality, is estimated. The estimation process was carried out by generalized additive models of location, scale and shape, in which the parameters were obtained through generalized linear estimates, non-parametric estimates and an estimate of spatial delay by SAR models.

***Keywords*:** Mixed Model, Linear Generalized Models, Forest Fires, Count Data, Zero-Inflated Models.

1. Introducción

Los incendios son una problemática que afecta cientas de hectáreas de terreno forestal cada año, por lo que una medición adecuada de esta información, puede proporcionar herramientas a los organismos de gestión y prevención de riesgo de desastres y control ambiental, para disponer de insumos que permita generar estrategias orientadas a la mitigación del riesgo relacionado con este tipo de desastres ambientales. Dicho esto, una de las alternativas metodológicas para tratar esta información es un modelo estadístico,

[a]Estudiante Estadística Universidad Santo Tomás
[b]Profesor Facultad de Estadística U. Santo Tomás
[c]Profesor Facultad de Estadística U. Santo Tomás

que permita identificar, aislar y cuantificar los factores que registren mayor incidencia sobre la cantidad de incendios forestales. Dentro del marco metodológico se planea estudiar los siniestros presentados en los municipios de los departamentos de Santander y Antioquia durante el año 2013. En la literatura se presentan diferentes estrategias para el modelado de este tipo de información, que van desde modelos lineales para datos de conteo, hasta la elección que se realizó para este estudio; Un Modelo Lineal Generalizado Mixto para datos de conteo.

El Ministerio de Ambiente y Desarrollo Sostenible de la República de Colombia resalta la importancia de las áreas boscosas en el país y la prevención de los incendios forestales, a través de investigaciones relacionadas con las causas de los siniestros, y proyectos de inclusión de las comunidades en la lucha contra los incendios forestales. Es por esto que se creó la ley 1523 de 2012, "por la cual se adopta la política nacional de gestión del riesgo de desastres y se establece el Sistema Nacional de Gestión del Riesgo de Desastres y se dictan otras disposiciones", sin embargo a nivel nacional no se ha implementado una estrategia que permita determinar los factores que incrementan el riesgo de que se presenten desastres relacionados con los incendios, por esto, el objetivo de esta investigación, es determinar a través de un modelo lineal generalizado mixto los factores climáticos y de ambiente que inciden en la cantidad de incendios forestales registrados en los departamentos de Santander y Antioquia a nivel municipal durante el año 2013.

De acuerdo con el IDEAM y al Ministerio de Ambiente y Desarrollo Sostenible de la República de Colombia gran parte de estos departamentos se encuentran en grado Alto y Muy Alto de vulnerabilidad de incendios a en relación a las coberturas vegetales del suelo. (Ver Anexo 2).

El proceso de estimación se realiza mediante GAMLSS (Generalized Additive Models for Location, Scale and Shape) que son modelos que se flexibilizan para permitir estimar, no solo la media, sino otros parámetros de la distribución de la variable respuesta, como funciones paramétricas lineales y/o aditivas no paramétricas de variables explicativas y/o efectos aleatorios. Para ajustar a los modelos se usa la estimación de máxima verosimilitud (penalizada).

El documento se desarrolla de acuerdo al siguiente orden: en la primera parte se da una introducción a los modelos lineales generalizados con profundización en los modelos Poisson, a continuación se presentan las bases teóricas de los modelos mixtos y los modelos cero inflados, después se hace la unión de todos los conceptos expuestos, expresando teóricamente la estructura de un modelo lineal generalizado mixto Poisson inflado con ceros en donde los efectos aleatorios están dados a través de estimaciones no paramétricas realizadas con splines penalizados (P-Splines) y con efectos de autocorrelación espacial dados a través de una estimación SAR (Modelos Autoregresivos Simultáneos Espaciales).

2. Modelo Lineal Generalizado

Los GLM *(Generalized Linear Models)* se concibieron como una extensión de los modelos lineales, que permiten el uso de distribuciones diferentes a la normal en las variables respuestas, como Binomiales, Poisson, Gamma y otras distribuciones pertenecientes a la familia exponencial.

La especificación de un GLM se realiza en tres partes:

1. **La componente aleatoria:** Que corresponde a la variable Y, que sigue una distribución de la familia exponencial; normal, lognormal, Poisson, Gamma, etc.

2. **La componente sistemática:** La cual es también conocida como predictor lineal, correspondiente a una combinación lineal de los parámetros desconocidos y las variables explicativas (covariables).

3. **Función de enlace:** o función Link, la cual relaciona la esperanza matemática de la variable dependiente Y con el predictor lineal. La función de enlace debe ser monótona y doblemente diferenciable.

2.1. Definición

Sean $Y_1, Y_2, \cdots, Y_n$ variables aleatorias independientes, cada una con función de densidad o función de probabilidad dada por:

$$f(Y_i; \theta_i, \phi) = exp\left[\phi\left\{y_i\theta_i - b(\theta_i)\right\} + c(y_i\phi)\right] \tag{1}$$

Es posible demostrar que, bajo condiciones de regularidad:

$$E\left\{\partial \log \frac{f(Y_i; \theta_i, \phi)}{\partial \theta_i}\right\} = 0$$

y

$$E\left[\partial^2 \log \frac{f(Y_i; \theta_i, \phi)}{\partial \theta_i^2}\right] = -E\left\{\frac{\partial \log f(Y_i; \theta_i, \phi)}{\partial \theta_i}\right\}^2 = 0,$$

para todo i, tal que $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ y $\mathrm{Var}(Y_i) = \phi^{-1}V(\mu_i)$, en donde $V_i = V(\mu_i) = d\mu_i/d\theta_i$ es la función de varianza y $\phi^{-1} > 0$ ($\phi > 0$) es el parámetro de dispersión (precisión).

La definición de los modelos lineales generalizados está dada por la descripción de la función de densidad de la familia exponencial descrita en la ecuación (1) y por la parte sistemática, descrita a continuación:

$$g(\mu_i) = \eta_i, \tag{2}$$

En donde $\eta_i = X_i^T\beta$ es el predictor lineal, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$, $p < n$, es un vector de parámetros desconocidos a ser estimados, $xi = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ representa los valores de variables explicativas y $g(\cdot)$ es una función monótona creciente y doblemente diferenciable, denominada función de enlace. (Paula (2004))
Para efectos de este caso de estudio, se va a tener en cuenta únicamente el caso particular de la distribución Poisson:

Caso Particular de la Distribución Poisson Sea $Y = Y_1, Y_2, \cdots, Y_n$ variables aleatorias con distribución Poisson de media μ, ($Y \sim P(\mu)$), cuya función de probabilidad está dada por:

$$P(y) = \frac{e^{-\mu}\mu^y}{y!} = exp\left\{y\log\mu - \mu - \log y!\right\}, \tag{3}$$

en donde $\mu > 0$ y $y = 0, 1, ...$ Haciendo $\log\mu = \theta$, $b(\theta) = e^\theta$, $\phi = 1$ y $c(y, \phi) = -\log y!$ obtenemos la ecuación (1).Por lo tanto $\mathrm{V}(\mu) = \mu$.

2.2. Funciones de enlace Canónicas

Suponiendo ϕ conocido, el logaritmo de la función de verosimilitud de un GLM con respuestas independientes, puede ser expresado por la forma:

$$L(\beta = \sum_{i=1}^{n} \phi\left\{y_i\theta_i - b(\theta_i)\right\} + \sum_{i=1}^{n} c(y_i, \phi) \tag{4}$$

Un caso particular ocurre cuando el parámetro canónico (θ) coincide con el predictor lineal, esto es cuando $\theta_i = \eta_i = \sum_{j=i}^{p} x_{ij}\beta_j$, en cuyo caso $L(\beta)$ está dada por:

$$L(\beta) = \sum_{i=1}^{n} \phi \left\{ \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j \right) \right\} + \sum_{i=1}^{n} c(y_i, \phi) \tag{5}$$

Definiendo la estadística $S_j = \phi \sum_{i=1}^{n} Y_i x_{ij}$, $L(\beta)$, es posible reexpresar esta función en la forma:

$$L(\beta) = \sum_{j=1}^{p} s_j\beta_j - \phi \sum_{i=1}^{n} b\left(\sum_{j=1}^{p} x_{ij}\beta_j \right) + \sum_{i=1}^{n} c(y_i, \phi) \tag{6}$$

Por lo tanto, por el teorema de la factorización la estadística $S = (S_1, S_2, \cdots, S_p)^T$ es suficiente minimal para el vector $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$. Las conexiones que corresponden a estas estadísticas son llamadas canónicas y desempeñan un papel importante en la teoría de los MLG. (Paula (2004))

La función de enlace canónica para un modelo Poisson es $\eta = \log \mu$ y para un modelo Binomial $\eta = logit(\mu) = \log\left\{ \frac{\mu}{1-\mu} \right\}$.

Una de las ventajas de usar enlaces canónicos es que estas garantizan la concavidad de $L(\beta)$ y en consecuencia muchos resultados asintóticos son obtenidos más fácilmente.

2.3. Modelos Inflados por Ceros

Los modelos inflados de ceros por acuerdo con Paula (2004) se caracterizan por la ocurrencia de ceros en dos situaciones:

I Ceros que se producen según la distribución de recuento

II Ceros estructurales

Esto significa que cuando una muestra aleatoria no cuenta con la característica que describe la variable aleatoria Y, se puede tratar como una muestra de ceros estructurales, esto se dá por ejemplo, cuando se quiere medir la cantidad de días que una familia consume un producto, las familias que no consumen dicho producto son ceros estructurales. El otro caso se presenta cuando la muestra aleatoria presenta la característica pero su valor es un cero, este caso se presenta cuando por algún motivo, hay más ceros de los cabría esperar en una distribución Poisson, en cuyo caso la función de probabilidad está expresada por:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i) f_z(0) & si \qquad y_i = 0 \\ \\ (1-\pi_i) f_z(y_i) & si \quad y_i = 1, 2, ... \end{cases} \tag{7}$$

para $i = 1, 2, \cdots, n$, $0 < \pi < 1$ es la probabilidad de que la observación de Y sea cero y $f_z(y)$ hace referencia para nuestro caso de estudio a la función de probabilidad de la distribución Poisson.

Puesto que $\sum_{y=1}^{\infty} f_z(y) = 1 - f_z(0)$ obtenemos $\sum_{y=0}^{\infty} P\{Y = y\} = \pi + (1-\pi) f_z(0) + (1-\pi)\{1 - f_z(0)\} = \pi + (1-\pi) = 1$.

Este modelo se denota como $Y \sim IZ(\pi, P(\lambda))$, cuya funcion de probabilidad está dada por:.

$$P(Y = y) = \begin{cases} \pi_i + (1-\pi_i) e^{-\mu} & si \qquad y_i = 0 \\ \\ (1-\pi_i) \frac{e^{-\mu}\mu^y}{y!} & si \quad y_i = 1, 2, ... \end{cases} \tag{8}$$

Para este modelo tenemos:

$$\begin{aligned} E(Y) &= \sum_{y=1}^{\infty} y(1-\pi) f_z(y) \\ &= (1-\pi) \sum_{y=1}^{\infty} y f_z(y) \\ &= (1-\pi) E(Z) \end{aligned}$$

y

$$\begin{aligned} E(Y^2) &= \sum_{y=1}^{\infty} y^2(1-\pi) f_z(y) \\ &= (1-\pi) \sum_{y=1}^{\infty} y^2 f_z(y) \\ &= (1-\pi) E(Z^2) \end{aligned}$$

Por lo tanto:

$$\begin{aligned} Var(Y) &= E(Y^2) - E^2(Y) \\ &= (1-\pi)E(Z^2) - (1-\pi)^2 E^2(Z) \\ &= (1-\pi)\{E(Z^2) - (1-\pi)E^2(Z)\} \end{aligned}$$

Y en el caso particular de la distribución Poisson tenemos:

$$E(Y) = (1-\pi)\mu$$

y dado que $E(Z) = \mu$ y $Var(Z) = \mu$, tenemos:

$$\begin{aligned} Var(Y) &= (1-\pi)\{E(Z^2) - (1-\pi)E^2(Z)\} \\ &= (1-\pi)\{\mu + \mu^2 - (1-\pi)\mu^2\} \\ &= (1-\pi)\{\mu + \mu^2 - \mu^2 + \pi\mu^2\} \\ &= (1-\pi)\mu(1+\pi\mu) \end{aligned}$$

Ahora tenemos que para una variable aleatoria $Z_i \sim P(\lambda_i)$ con $\lambda_i = e^{x_i^T \beta}$ y $\log\{\frac{\pi_i}{1-\pi_i}\} = e^{u_i^T \gamma}$, de esta manera el valor esperado de Y está dado por:

$$\begin{aligned} \mu_i =& (1-\pi)E(Z_i) \\ &= (1-\pi)\lambda_i \\ &= \left\{1 - \frac{e^{u_i^T\gamma}}{1-e^{u_i^T\gamma}}\right\} e^{x_i^T\beta} \\ &= \frac{e^{x_i^T\beta}}{1+e^{u_i^T\gamma}} \end{aligned}$$

Donde X_i son las covariables que explican λ y u_i son las covariables que explican π. La función de enlace está dada por:

$$\log \mu_i = \eta_i = x_i^T\beta - \log(1 + e^{u_i^T\gamma}) \tag{9}$$

2.4. Estimación en los Modelos Lineales Generalizados

En el caso de la familia exponencial, dado un vector de observaciones $y = (y_i, y_2, \cdots, y_n)'$. El logaritmo de la verosimilitud es

$$l(\theta \mid y) = \sum_{i=1}^{n}((y_i\theta_i - b(\theta_i)/a(\phi) + c(y_i, \phi)) \tag{10}$$

Cuando usamos el link canónico: $\theta = \eta = X'\beta$, de modo que podemos estimar los parámetros de interés β: Por lo tanto, la función score:

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta}$$

es posible probar que:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n}\frac{(y_i - \mu_i)}{g'(\mu_i)V_i}x_i$$

donde $V_i = Var(y_i) = a(\phi)b''(\theta_i)$.
Necesitamos igualar esa ecuación a cero, pero no existe una solución exacta, y utlizamos una versión de algoritmo de Newton-Rapson, llamado *Fisher Scoring Algorithm*. Es un algoritmo iterativo cuya solución es:

$$\beta_{new} = (X'WX)^{-1}X'Wz$$

donde $z = X\beta_{old} + (y - \mu_{old})g'(\mu_{old})$ (llamada *variable de trabajo*), y $\boldsymbol{W}$ es una matriz diagonal, cuyos elementos $w_{ii} = 1/g'(\mu_i)$.

3. Modelo Lineal Mixto

Antes de mostrar el modelo mixto es preciso aclarar algunos conceptos básicos en relación a los modelos de regresión, los cuales tienen que ver con los efectos fijos y aleatorios con los que cada modelo cuenta. Correa Morales & Salazar Uribe (2016) define estos términos así:

- **Efectos Aleatorios:** Se trata de efectos que son tomados al azar de una población de posibles niveles, son los factores de los cuales no es objetivo principal obtener una estimación, otra forma de verlos es el efecto que desea ser marginalizado por algún motivo.
- **Efectos Fijos:** Son los efectos que se determinan a propósito, es decir son los efectos que centran el objetivo principal del análisis, de los cuales, se desea obtener una estimación del impacto en cada posible nivel.

De acuerdo con Stroup (2012) un modelo estadístico debe cumplir con tres elementos:

1. La observación en la variable respuesta.
2. La parte sistemática o determinista del proceso que da lugar a la observación
3. La parte aleatoria, que incluye una declaración de la distribución de probabilidad asumida

Un modelo mixto permite analizar una variable aleatoria Y modelizando simultáneamente el valor esperado del fenómeno estudiado y su variabilidad, es decir se modela un efecto fijo a todos los sujetos de estudio y otro efecto aleatorio asociado a cada uno de los sujetos.

Es posible obtener un modelo mixto de la agrupación de un modelo en dos etapas, que no es mas que la adaptación al modelado lineal simple que permite aproximar los perfiles asociados a cada sujeto así: La primera parte es ajustar una regresión asociada a cada sujeto por separado y la segunda es ajustar una regresión a los coeficientes asociados a cada sujeto en función de variables conocidas (efectos fijos) así:

- **Primera Etapa:** Suponga $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ una variable aleatoria para las orbervaciones del i-ésimo sujeto tomado en el tiempo X_{ij}, el modelo $Y_i = Z_i\beta_i + \epsilon_i$ representa la varianza entre sujetos, siendo Z_i una matriz de orden n_i x q de covariables conocidas, β_i un vector con dimensión q de coeficientes de la regresión y $\epsilon_i \sim N(0, \Sigma_i)$, con Σ_i, la matriz de varianzas y covarianzas.
- **Segunda Etapa:** En esta etapa se modelan los $\beta_i's$ con covariables conocidas, obteniendo; $\beta_i = K_i\beta + b_i$, siendo K_i una matriz de orden q x p de covariables conocidas, β un vector con dimensión p de coeficientes de la regresión y $b_i \sim N(0, D)$, con D, la matriz de varianzas y covarianzas.

De esta manera un modelo en dos etapas dado por:

$$\begin{aligned} Y_i &= Z_i\beta_i + \epsilon_i \\ \beta_i &= K_i\beta + b_i \end{aligned} \tag{11}$$

Se puede resumir en un sólo modelo mixto (MLM) dado por:

$$\begin{aligned} Y_i &= Z_iK_i\beta + Z_ib_i + \epsilon_i \\ &= X_i\beta + Z_ib_i + \epsilon_i \end{aligned} \tag{12}$$

para este modelo tenemos $Z_iK_i = X_i$, $b_i \sim N(0, D)$ y $\epsilon_i \sim N(0, \Sigma_i)$, $b_1, b_2, ..., b_m$ y $\epsilon_1, \epsilon_2, ..., \epsilon_m$ son mutuamente independientes, en donde:

- β: Hace referencia a los efectos fijos.
- b_i y ϵ_i: hace referencia a los efectos aleatorios.
- D y Σ: Contienen las componentes de varianza.

Matricialmente:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix}; \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}; \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}; \quad Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}; \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

La forma matricial de un modelo Lineal Mixto queda definida como:

$$Y = X\beta + Zb + \epsilon \tag{13}$$

Donde $E(\epsilon) = 0$ y

$$\Sigma_\epsilon = \begin{pmatrix} \epsilon_1 & 0 & \cdots & 0 \\ 0 & \epsilon_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \epsilon_m \end{pmatrix}$$

$$\Sigma_b = \begin{pmatrix} D & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D \end{pmatrix} = I_m \otimes D$$

3.1. Modelo Lineal Generalizado Mixto

Sea Y el vector de respuesta de una variable aleatoria condicionada en b y perteneciente a la familia exponencial, tenemos:

$$f_{yi|b}(y_i|b,\beta,\phi) = exp\left\{\frac{y_i\eta_i + c(\eta_i)}{a(\phi)} + d(y_i,\phi)\right\} \tag{14}$$
$$b \sim f_b(b|D)$$

El modelo lineal generalizado mixto (MLGM) esta dado por:

$$\eta_i = \mathbf{X}_i'\beta + \mathbf{z}_i'b \tag{15}$$

Con $\mathbf{X}_i'$ la i-ésima fila de $\mathbf{X}$ y $\mathbf{z}_i'$ la i-ésima fila de $\mathbf{Z}$.

3.1.1. Estimación de un Modelo Lineal Generalizado Mixto

La verosimilitud para la ecuación (14) es:

$$L(\beta,\phi,\mathbf{D}|\mathbf{Y}) = \int \prod_{i=1}^{n} f_{y_i|b}(y_i|\mathbf{b},\beta,\phi) f_b(\mathbf{b}|\mathbf{D}) d\mathbf{b} \tag{16}$$

Esta función no puede evaluarse en forma cerrada y requiere de métodos numéricos. El método de la cuadratura de Gauss puede ser una opción adecuada para esta evaluación.

De acuerdo con Correa Morales & Salazar Uribe (2016) una aproximación presentada por McGilchrist (1994), asumiendo un modelo lineal mixto con la siguiente estructura.

$$\begin{aligned} Y &= \eta + \epsilon \\ &= X\beta + Zb + \epsilon \end{aligned}$$

donde $\epsilon \sim N(0,\sigma^2\mathbf{D})$ y con $\mathbf{D}$ conocida. El vector de respuesta media η depende de una componente fija $X\beta$ con X la matrix de diseño con constantes conocidas y β el vector de parámetros poblacionales desconocidos, además $\mathbf{Zb}$ se puede particionar como:

$$\begin{aligned} Z &= (Z_1, Z_2, \cdots, Z_k) \\ b' &= (b_1', b_2', \cdots, b_k') \end{aligned}$$

donde b_j tiene v_j componentes y $b_j \sim N(0,\sigma_j^2 A_j)$ y además es independiente de otras componentes b. Y con $\sigma_j^2 = \sigma^2\theta_j$, entonces:

$$A = \begin{bmatrix} \theta_1 A_1 & 0 & \cdots & 0 \\ 0 & \theta_2 A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_k A_k \end{bmatrix}$$

El procedimiento BLUP consiste en la maximización de la suma de las componentes de log-verosimilitud:

- Sea l_1 la log-verosimilitud para $\mathbf{Y}$ dado $\mathbf{b}$ condicionalmente fijo,

- sea l_2 la log-verosimilitud para $\mathbf{b}$ y
- $l = l_1 + l_2$

Así l representa la log-verosimilitud basada en la distribución conjunta de $\mathbf{Y}$ y $\mathbf{b}$. El procedimiento BLUP (mejor estimador lineal insesgado) selecciona estimaciones de β, b, σ^2 y los parámetros de $\mathbf{A}$ que maximicen l. Resolviendo el sistema de ecuaciones de obtiene:

$$\begin{bmatrix} X'D^{-1}X & X'D^{-1}Z \\ Z'D^{-1}X & Z'D^{-1}Z + A^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{b} \end{bmatrix} = \begin{bmatrix} X'D^{-1}Y \\ Z'D^{-1}Y \end{bmatrix}$$

4. Modelo Lineal Generalizado Mixto Poisson Cero Inflado

Para lograr que la estimación del modelo para la variable de estudio sea un modelo mixto, es necesario, como se mostró anteriormente que este tenga una parte de efectos fijos y otra parte de fectos aleatorios, para tal fin se propone incluir la relación espacial de los siniestros en cada municipio, es decir que los incendios se pueden dar más fácilmente en algunas regiones dadas sus características de terreno, tipo de bosques, temperatura, etc.
Dicho efecto aleatorio, está dado por dos factores:

I Inclusión de estimaciones no paramétricas mediante Splines penalizados **(P-Splines)**.

II Adición de la relación espacial de la cantidad de incendios en los municipios mediante una estimación autoregresiva espacial (SAR).

4.1. Splines Penalizados - P-Splines

De acuerdo con Durbán (2009) los modelos P-Splines, reúnen las mejores características de los enfoques de modelos de suavizado mediante splines; *smooting splines* y *regression splines*, ya que utilizan menos parámetros que los splines de suavizado, pero la selección de los nodos no es tan determinante como en los splines de regresión, veamos:

Para un modelo de n pares de datos (x_i, y_i), dado por:

$$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \tag{17}$$

siendo $f(\cdot)$ una función suave de los datos.
La penalización a los splines se realiza utilizando mínimos cuadrados para ajustar un modelo a una base $\mathbf{B}$ construída con k nodos, en donde la función a minimizar es:

$$S(a; y) = (y - Ba)'(y - Ba) \Rightarrow \hat{a} = (B'B)^{-1}B'y$$

La penalización introducida por Eilers & Marx (1996) se basa en las diferencias entre los coeficientes adyacentes de las bases B-Splines y se añade a la función de mínimos cuadrados, resultando la siguiente función de mínimos cuadrados penalizados:

$$S(a; y) = (y - Ba)'(y - Ba) + \lambda a' P_d a \Rightarrow \hat{a} = (B'B + \lambda P_d)^{-1}B'y$$

la función del parámetro λ es controlar la suavidad de la curva, pero aquí lo que hace es penalizar los coeficientes que están muy separados entre sí, y cuanto mayor sea λ, más se aproximarán los coeficientes a cero, de modo que si $\lambda \to \infty$ nos aproximamos a un ajuste polinómico. Por el contrario, cuando $\lambda \to 0$

estaremos utilizando mínimos cuadrados ordinarios. (Durbán (2009)).

La selección el parámetro de suavizado se hace mediante criterios de comparación de modelos: Criterio de información de Akaike (AIC), Criterio de información bayesiano (BIC), Validación cruzada generalizada (GCV), ect..

4.2. Autocorrelación Espacial (Modelos SAR)

Un modelo autoregresivo espacial consiste en una versión espacialmente rezagada de Y, así:

$$y = \rho \mathbf{W}\mathbf{y} + \varepsilon \tag{18}$$

Este es un modelo similar a u modelo de regresión estándar, donde la el primer término se construye a partir de una matrix nxn de correlación espacial W predefinida aplicada a la variable observada y junto con un parámetro de autorregresión espacial ρ que típicamente tiene que ser estimado a partir de los datos. Esencialmente, un modelo de desfase espacial expresa la noción de que el valor de una variable en una ubicación dada se relaciona con los valores de la misma variable medida en ubicaciones cercanas, lo que refleja algún tipo de efecto de interacción. La matriz de ponderaciones espaciales, W, casi siempre está estandarizada de manera que sus filas sumen a 1, por lo que efectivamente incluye un promedio ponderado de los valores vecinos en la ecuación de regresión.
Para una observación individual, la ecuación básica de autoregresión retardada espacial es simplemente:

$$y_i = \rho \sum_j W_{ij} y_j + \varepsilon_i$$

Este modelo se puede reescribir como:

$$y = (1 - \rho W)^{-1} + \varepsilon$$

y así obtener la expresión de la varianza de y, dada por:

$$\begin{aligned} \mathbf{VAR}(y) = \mathbf{E}(yy^T) &= (1 - \rho W)^{-1} \mathbf{E}(\varepsilon\varepsilon^T)(1 - \rho W^T)^{-1} \\ &= \mathbf{E}(yy^T) = (1 - \rho W)^{-1} C (1 - \rho W^T)^{-1} \end{aligned}$$

donde C, es la matriz de varianzas y covarianzas. (De Smith et al. (2007))
esta derivación no tiene supuestos distribucionales respecto a la variable respuesta y o sobre los errores. En este modelo, la matriz de pesos espacial$W$.

Dicho esto, para lograr el efecto espacial deseado sobre la estimación, en el modelo se incluye un operador de retardo espacial, para tal fin se da una suma de los pesos en todos los valores de una vecindad dada, obteniendo los valores asociados a cada municipio, mediante la multiplicación de la variable observada en la vecindad i con los pesos asociados a la matriz de peso espacial W. resultando así:

$$W_i y = \sum_j w_{ij} y_j,$$

para todo j que pertenezca al conjunto J.

Así, el modelo a aplicar formalmente está dado por:

Sea Y una variable aleatoria tal que $Y \sim IZ(\pi, P(\lambda))$, con función de probabilidad dada por:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda} & si \quad y_i = 0 \\ (1-\pi_i)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & si \quad y_i = 1, 2, ... \end{cases} \tag{19}$$

Para el cual la función de enlace está dada por:

$$\eta_i = x_i^T\beta - \log(1 + e^{u_i^T\gamma}) + f(z_i) + W_s y \tag{20}$$

En donde $x_i^T\beta$ hace referencia a los efectos fijos asociados a la media de la variable de estudio, $u_i^T\gamma$ hace referencia a la probabilidad de que el valor de Y sea un cero y $f(z_i)$ hace referencia a la estimación del modelo de suavizado por P-Splines y $W_s y$ es el operador de retardo espacial. Estos dos últimos son los efectos aleatorios dados bajo la hipótesis que los incendios en los municipios pueden tener relaciones no lineales con alguna covariable y que además tienen relación espacial.

5. Estimación de Parámetros

Para la estimación de los parámetros se utilizará el método de GAMLSS: Modelos Aditivos Generalizados de Ubicación, Forma y Escala o generalized additive model for location, scale and shape por sus siglas en inglés.
Los modelos aditivos generalizados para ubicación, escala y forma (GAMLSS) fueron introducidos por Rigby y Stasinopoulos (2001, 2005) como una forma de superar algunas de las limitaciones asociadas con los Modelos Lineales Generalizados (GLM) y los Modelos Aditivos Generalizados (GAM) (Stasinopoulos et al. (2017)).

En los modelos GAMLSS, la suposición de distribución de la familia exponencial para la variable de respuesta (y) se relajó y se reemplazó por una familia de distribución general, incluidas las distribuciones altamente asimétricas y/o kurtóticas. La parte sistemática del modelo se amplía para permitir modelar no solo la media (o ubicación) sino otros parámetros de la distribución de y, como funciones paramétricas lineales y/o aditivas no paramétricas de variables explicativas y/o efectos aleatorios.
La estimación de máxima verosilimitud (penalizada) se usa para ajustar los modelos mediante los algoritmos CG y RS.

El primero, el algoritmo CG, es una generalización del algoritmo de Cole y Green (1992) (y utiliza la primera (esperada o aproximada) segunda y derivadas cruzadas de la función de verosimilitud con respecto a los parámetros θ). Sin embargo, para muchas funciones de probabilidad (densidad) de población $f(y|?)$ los parámetros θ son información ortogonal (ya que los valores esperados de las derivadas cruzadas de la función de verosimilitud son 0). En este caso, es más adecuado el algoritmo RS más simple, que es una generalización del algoritmo que utilizaron Rigby y Stasinopoulos para ajustar los modelos aditivos medios y de dispersión (y no usa derivadas cruzadas) (Rigby & Stasinopoulos (2005)).

En los modelos GAMLSS se asume que las observaciones y_i para $i = 1, 2, \cdots, n$ con función de densidad $f(y_i|\theta^i)$ condicional el θ^i donde $\theta^i = (\theta_{i1}, \theta_{i2}, \cdots, \theta_{ip})$ es un vector de p parámetros, cada uno relacionado con las variables explicativas, para $p = 4$, la implementación en `R` denota estos parámetros como $\mu_i, \sigma_i, \nu_i, \tau_i$. los primeros dos parámetros suelen caracterizarse como parámetros de locación y escala y los restantes son parámetros de forma.

Sea $y^T = (y_1, y_2, \cdots, y_n)$ un vector de tamaño n de la variable respuesta. También para k_j Se debe conocer la función de enlace monónota $g_k(.)$ que relaciona el k-ésimo parámetro θ_k con las variables

explicativas mediante modelos aditivos semiparamétricos dados por:

$$\begin{aligned} g(\mu) &= \eta_1 = x_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \\ g(\sigma) &= \eta_2 = x_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}(x_{j2}) \\ g(\nu) &= \eta_3 = x_3\beta_3 + \sum_{j=1}^{J_3} h_{j3}(x_{j3}) \\ g(\tau) &= \eta_4 = x_4\beta_4 + \sum_{j=1}^{J_4} h_{j4}(x_{j4}) \end{aligned} \tag{21}$$

Donde la función h_{jk} es una función aditiva no paramétrica de la variable explicativa X_{jk}. Este modelo se ha ampliado para permitir que los términos de efectos aleatorios se incluyan en el modelo para los parámetros:

$$\begin{aligned} g(\mu) &= \eta_1 = x_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}(\gamma_{j1}) \\ g(\sigma) &= \eta_2 = x_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2}(\gamma_{j2}) \\ g(\nu) &= \eta_3 = x_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}(\gamma_{j3}) \\ g(\tau) &= \eta_4 = x_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}(\gamma_{j4}) \end{aligned} \tag{22}$$

Donde γ_{jk} tienen distribuciones normales con $\gamma_{jk} \sim N_{q_{jk}}(0, G_{jk}^{-1})$ y G_{jk}^{-1} es la matriz inversa generalizada. Los vectores paramétricos β_k y los parámetros de efectos aleatorios λ_{jk} se estiman dentro del marco GAMLSS maximizando una función de verosimilitud penalizada ℓ_p data por:

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \lambda_{jk}\gamma_{jk}' G_{jk}\gamma_{jk} \tag{23}$$

donde $\ell = \sum_{i=1}^{n} log(f(y_i|\theta^i))$ es la función de log-verosimilitud.

6. Resultados y Discusión

En el capítulo anterior se establecieron las bases teóricas para la estimación del modelo, a continuación se muestran los resultados obtenidos y las estimaciones de los parámetros mediante la aplicación a los datos reales.
Lo primero es presentar el contexto de la información analizada:

6.1. Análisis Descriptivo

La información analizada en el modelo se obtuvo a través del Instituto de Hidrología, Meteorología y Estudios Ambientales (**IDEAM**), con fines exclusivamente académicos, del cual se tomó el número de siniestros registrados por incendios forestales a nivel municipal.

La información recolectada mostraría un estudio a nivel total país, sin embargo, no se registró información de las covariables en todos los municipios de Colombia, por lo que se tomó la determinación de estudiar únicamente los municipios de los departamentos de Santander y Antioquia, adicionalmente se decidió trabajar la información recolectada para el año 2013, ya que fué el año que registró la mayor cantidad de siniestros y que presentaba la mayor completitud de los datos. La información recolectada se presenta a continuación:

1. **Variable respuesta:** Número de incendios Forestales registrados en cada uno de los municipios durante el año 2013.
2. **Área quemada:** Cantidad en hectáreas de terreno boscoso consumido por el fuego en todos los siniestros presentados durante el año.
3. **Área Bosques:** Cantidad en hectáreas de bosques por municipio.
4. **Precipitación máxima:** Precipitación máxima mensual en el año medida en mm o L/m^2.
5. **Temperatura media:** Promedio anual de temperatura diaria, medida en ℃.
6. **Días con precipitación:** Número de días al año en los que se presentó precipitación en cada municipio.
7. **Precipitación total:** Volumen total de precipitación al año medida en mm o L/m^2.
8. **Área oficial:** Tamaño del municipio en Km^2.
9. **Área cosechada:** Superficie cosechada de los principales productos agrícolas, medida en hectáreas por municipio.
10. **Poblacional Total:** Cantidad total de habitantes de cada municipio, incluye cabecera municipal y zonas rurales.

A continuación, se presenta el comportamiento de la variable respuesta en los departamentos en mención.

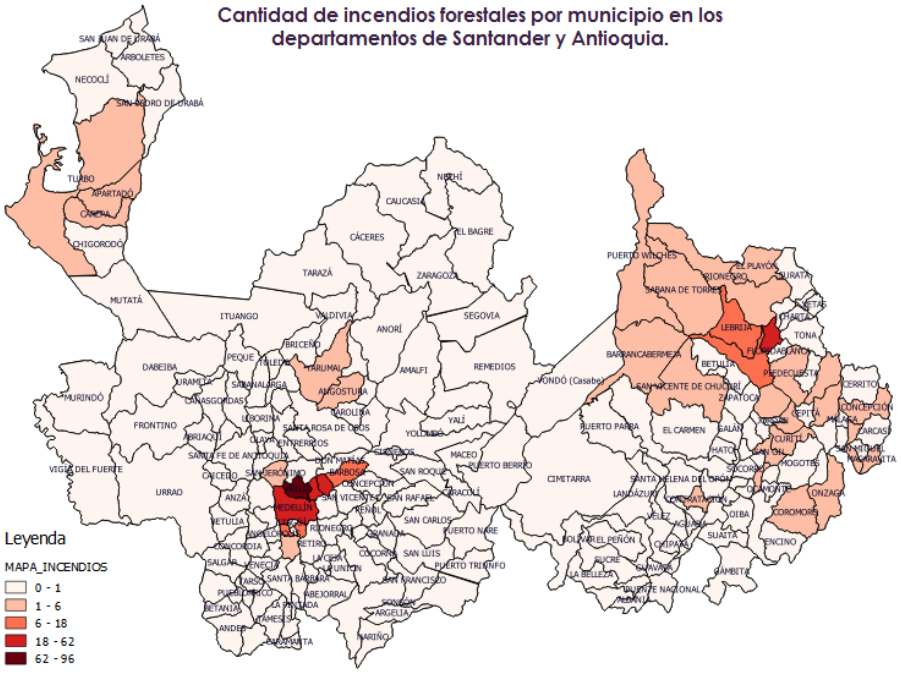


Figura 1: Incendios forestales por municipio.

Los intervalos representados en la gama de colores del anterior mapa cloroplético se construyó a partir de las rupturas naturales de Jenks el cual se basa en la minimización de la varianza en cada intervalo creado. En el se evidencia la concentración de siniestros relacionados con incendios forestales, mediante una escala colorométrica que indica colores claros donde se registran la menor candidad de incendios, e inversamente colores oscuros en los municipios que registran mayor candidad de siniestros.
La observación más sobresaliente es que las mayores concentraciones de incendios se encuentran en las áreas metropolitanas de las capitales de los dos departamentos, este hecho puede tener relación con la concentración de población registrada en estas áreas, ya que es sabido que buena parte de los citados incidentes ambientales son causados por la mano humana.
De acuerdo con el planteamiento teórico del modelo, la variable de estudio debe estar relacionada espacialmente, es decir que para regiones vecinas existe una asociación estadísticamente significativa de valores parecidos o discordantes. Esta hipótesis se corrobora con el Test de Morán (*I de Moran*), el cual con un tamaño de muestra grande se distribuye asintóticamente como una distribución Normal $Z(I)$, así, cuando el valor de $Z(I)$, es positivo, indica que existe autocorrelación espacial positiva, es decir, valores parecidos de la variable de interés en regiones cercanas. Inversamente cuando $Z(I)$ es negativo indica correlación espacial negativa, es decir, precencia de valores disímiles para la variable Y en regiones cercanas.
El índice de Moran se define como:

$$I = \frac{N}{S_0} \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, i \neq j$$

donde y_i refleja el valor de la variable y en la región i, $\overline{y}$ es la media muestral, w_{ij} son los pesos de la matriz W, N es el tamaño muestral y $S_0 = \sum_i \sum_j w_{ij}$.

El sistema de hipótesis planteado es:

H_0: No existe correlación espacial en la cantidad de incendios forestales por municipio.
H_a: Existe correlación espacial en la cantidad de incendios forestales por municipio.

(a) (b)

Figura 2: A la izquierda se encuentra el diagrama de dispersión de Morán, en el cual cada punto representa en el eje X las observaciones de la variable de estudio y en eje Y las observaciones de la misma variable en ubicaciones vecinas, indicando la relación positiva entre la cantidad de incendios de cada municipio con sus municipios vecinos.
A la derecha está la distribución de los datos construída con base a permutaciones para determinar qué tan probable sería observar el valor I de Moran de una distribución real en condiciones de aleatoriedad espacial. La línea verde indica el Pseudo P-Valor y la línea negra indica la región de rechazo de la distribución, se puede observar que el estadístico de Moran se encuentra en la región de rechazo.

A continuación se muestra el juzgamiento del estadístico realizado con la función `MTest` del software

estadístico R.

Tabla 1: Test de Morán para la Cantidad de Incendios por Municipio

Moran I Test			
Standard Deviate:	8.0054	P-Value:	1.191e-15
Null Hypothesis:		Two Sided	
Sample Estimates:	Moran I statistic:	Expectation:	Variance:
	0.274791710	-0.004739336	0.001219264

Los resultados obtenidos indican que el P-Valor es menor que el nivel de significancia del $\alpha = 5\,\%$, además el I de moran es positivo, por lo cual se puede determinar que existe suficiente evidencia estadística, para rechazar la hipótesis nula y afirmar que existe autocorrelación espacial es decir los valores en las observaciones de cada municipio están más agrupadas espacialmente de lo que se esperaría si los procesos espaciales vecinos fueran aleatorios.
El valor del estadístico de Moran es positivo, lo que indica que la relación espacial de la variable de Incendios forestales es directa, es decir que hay valores similares de la variable en regiones vecinas para cada municipio.

A partir de este resultado se garantiza que hay condiciones de trabajo suficientes para la aplicación de la relación espacial propuesta en el modelo, por otro lado, se tiene la variable de estudio que se contiene datos de conteo con exceso de ceros, lo que nos dá el escenario para avanzar en la propuesta de estimación.

6.2. Aplicación del modelo

La aplicación del modelo teórico a los datos de incendios forestales mencionados se realiza a través del software estadístico R, principalmente a través del uso del paquete `gamlss`, el cual permite realizar estimaciones para modelos aditivos mediante procesos iterativos.
Antes de incluir las variables al modelo se calculó W, la matriz de pesos basada en las vecindades desde el punto de vista espacial, con el fin de utilizar el operador de rezago espacial dentro de la estimación del modelo.
Por otro lado se realizó un análisis de las variables que podrían ser candidatas a la estimación por suavizamiento no paramétrico, teniendo en cuenta las que no tenían una relación lineal con la variable de estudio, determinando así que las variables **Temperatura media** y **Área Cosechada** son las que mejor se ajustan a este comportamiento.

A continuación veremos la comparación de varios modelos estimados mediante la metodología de GAMLSS.

Tabla 2: Tabla Comparativa entre modelos estimados

Estimación de la media de la variable	Estimación de $P(X=0)$	Iteraciones	Global Deviance	AIC	Observaciones
pb(Temp) + Ar. Cos + Pres. + Pres. d + Area	Esp. Eff + Pob.T	23	832.642	872.8874	
pb(Temp) + pb(Ar. Cos) + Pres. + Pres. d + Area	Esp. Eff + Pob.T	100	609.5804	668.5282	No converge
pb(Temp) + pb(Ar. Cos) + Pres. + Pres. d + Area + Area.B	Esp. Eff + Pob.T	100	634.7279	694.414	No converge, variables no significativas
pb(Temp) + Ar. Cos + Pres. + Pres. d + Area	Esp. Eff	18	854.4892	891.1326	
pb(Temp) + Ar. Cos + Pres. + Pres. d + Area	Area + Esp. Eff	47	853.7735	890.7997	
(Temp) + Ar. Cos + Pres. + Area	Area + Esp. Eff + Pob.T)	100	1696.955	1712.955	
pb(Temp) + pb(Ar. Cos) + Pres. + Area + pb(Pob.T)	Area + Esp. Eff + Pob.T	100	630.0635	659.0392	No converge, variables no significativas
pb(Temp) + pb(Ar. Cos) + Esp. effec Pob + Pres. + Pres. d + Area	Esp. Eff + Pob.T	88	637.717	690.2505	

De acuerdo a la tabla se evidencia que el modelo con el menor AIC que converge es el último, cuya ecuación está dada por:

Fitted Model
Model Equation:
`mu.formula=CANT_INCENDIOS~0+pb(Temp)+pb(Ar. Cos)+Esp. effec Pob+Pres+Pres. d+Area,sigma.formula=~0+Esp.Eff+Pob.T`

Method: `Mixed`
Enlaces:
`Mu link function: log`
`Sigma link function: logit`

Los resultados del modelo se muestran a continuación:

Tabla 3: Resultados Estimación del Modelo

Mu Coefficients:					
Variable	Estimate	Std. Error	t value	Pr(> |t|)	Transfomada
pb(MEAN_TEMP)	2.224e-01	7.681e-03	28.955	<2e-16 ***	
pb(AREACOSECHADA)	-4.899e-05	9.537e-06	-5.136	7.04e-07 ***	
TOT_PREC	6.479e-05	1.530e-05	4.235	3.58e-05 ***	1.0000616
DIAS_PREC	-1.309e-02	5.022e-04	-26.076	<2e-16 ***	0.9873258
AREABOSQUES	-2.167e-05	5.588e-06	-3.878	0.000146 ***	0.9999787
POBLACION_TOTAL	1.207e-06	1.217e-07	9.920	<2e-16 ***	1.0000012
Sigma Coefficients:					
Variable	Estimate	Std. Error	t value	Pr(> |t|)	Transfomada
ESP_EFF	9.805e-02	5.062e-02	1.937	0.0541 .	0.104719699475
POBLACION_TOTAL	-7.349e-06	3.577e-06	-2.055	0.0412 *	-0.000006912372

Se evidencia que con este modelo, todas las variables son estadísticamente significativas son un nivel de confianza del 94 %.
En cuanto al efecto espacial, se observa que para cada municipio, los municipios cercanos presentan un 10 % de riesgo adicional de que no haya incendios. Por otro lado, el efecto causado por las variables **Precipitación total**, **Días con precipitación**, **Área total de bosques** y **Población total por municipio** presentan efectos similares en relación con la cantidad de incendios en cada municipio, indicando que con un incremento en una unidad de cada una de las variables, el efecto sobre la cantidad de incendios es cercana al 1 %.
A continuación se puede observar algunos supuestos sobre los residuales del modelo.

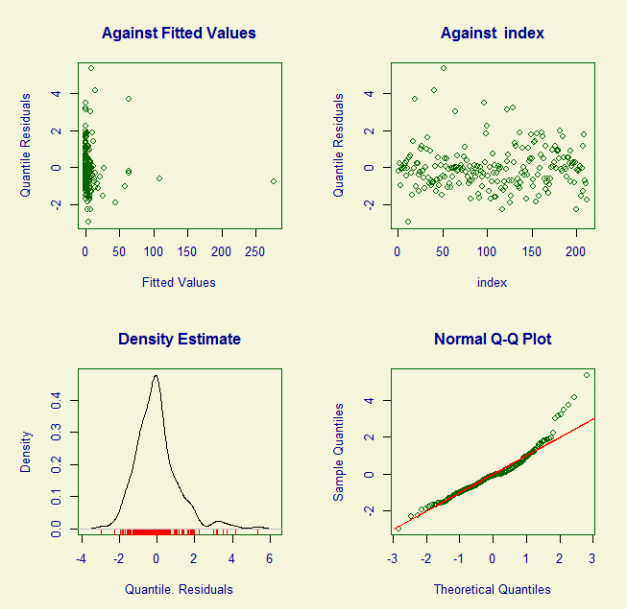
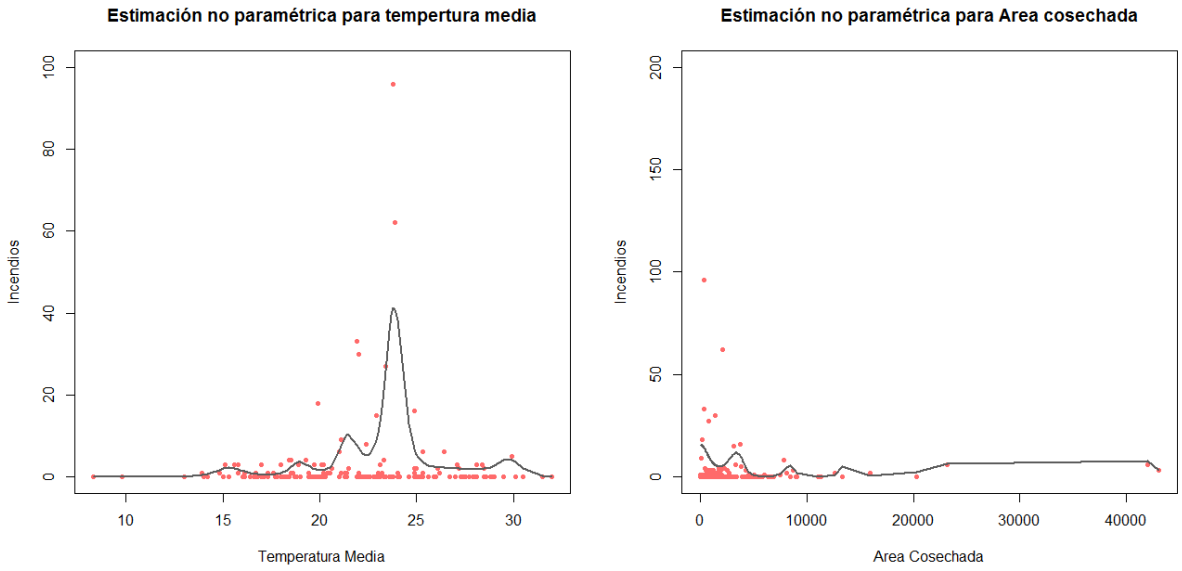


Figura 3: Incendios forestales por municipio.

De acuerdo con el anterior gráfico de residuales, se puede determinar que el modelo es bueno, sin embargo, el supuesto de normalidad en los residuales demuestra que cuando la cantida de incendios es excesivamente grande, el modelo tiene problemas para registrarlo.

Las variables *Temperatura Media* y *Area Cosechada* fueron estimadas mediante métodos de suavizamiento con P-Splines y los resultados de estas estimaciones se muestran a continuación:



(a) (b)

Figura 4: A la izquierda, está la estimación P-Spline para Temperatura Media, en la cual se observa la relación no lineal entre la temperatura media por municipio y la cantidad de incendios presentados en cada uno de ellos. De acuerdo al gráfico, la mayor cantidad de incendios se presenta en temperaturas de entre los 20 ℃y los 25 ℃. Además, se evidencia que en temperaturas menores a los 20 ℃es poco probable encontrar incendios forestales.
A la derecha se encuentra la estimación P-Spline para la superficie cosechada de los principales productos agrícolas en cada municipio, en el cual se observa que el mayor volumen de incendios forestales se presenta en municipios cuya área cultivada es menos a 4000 hectáreas. La curva muestra incrementos en la cantidad de incendios cuando las áreas cultivadas están al rededor de 8000 hectráreas y cerca a las 13000 hectráreas, sin embargo, de evidencia que en grandes extenciones cultivadas mayores a las 20000 hectáreas la cantidad de incendios no aumenta.

En los dos casos se evidenció que en la estimación del modelo existe una ganancia significativa en términos de ajuste utilizando este método de estimación en las variables en mención.

7. Conclusiones

1. Las variables analizadas tienen efectos significativos sobre la cantidad de incendios registrados en cada municipio. Ya que sólo se evaluó un año, es posible expander esta investigación, tanto a otros años, como a otras regiones del país.

2. La metodología usada de modelos mixtos, incluyendo estimaciones no paramétricas y de autocorrelación espacial, demostró ser mejor que un modelo lineal generalizado Poisson cero inflado, ya que las estimaciones presentaron mejores resultados en los criterios de comparación.

3. Las variables analizadas son estadísticamente significativas, lo que sugiere que las estancias gubernamentales encargadas de la gestión del riesgo, pueden enfocar sus esfuerzos en función de estas variables hacia la prevención de futuros siniestros ambientales.

Agradecimientos

A mi esposa Fanny Patricia Ovalle y a mi hija Juliana Valeria, por su invaluable apoyo, acompañamiento y porque no dejaron de creer en mí durante este proceso, a mis padres por enseñarme que la perseverancia y el esfuerzo al final siempre serán premiados, a los profesores de la facultad, quienes compartieron su conocimiento, experiencia y valores y a mis compañeros por las inolvidables experiencias que tuvimos durante los últimos cinco años.

Referencias

Agresti, A. (2015), *Foundations of linear and generalized linear models*, John Wiley & Sons.

Correa Morales, J. C. & Salazar Uribe, J. C. (2016), *Introducción a Los Modelos Mixtos*, Universidad Nacional de Colombia-Sede Medellín.

Cressie, N. (1991), *Statistics for Spatial Data*, Wiley-Interscience.

De Smith, M. J., Goodchild, M. F. & Longley, P. (2007), *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*, Troubador Publishing Ltd.

Durbán, M. (2009), 'An introduction to smoothing with penalties: P-splines', *Boletín de Estadística e Investigación Operativa* **25**(3), 195–205.

Eilers, P. H. & Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical science* pp. 89–102.

Paula, G. A. (2004), *Modelos de regressão: com apoio computacional*, IME-USP São Paulo.

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3), 507–554.

Shelton, A. O., Thorson, J. T., Ward, E. J. & Feist, B. E. (2014), 'Spatial semiparametric models improve estimates of species abundance and distribution', *Canadian Journal of Fisheries and Aquatic Sciences* **71**(11), 1655–1666.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. & De Bastiani, F. (2017), *Flexible Regression and Smoothing: Using GAMLSS in R*, CRC Press.

Stroup, W. W. (2012), *Generalized linear mixed models: modern concepts, methods and applications*, CRC press.

Vanegas, L. H. & Paula, G. A. (2015), 'A semiparametric approach for joint modeling of median and skewness', *Test* **24**(1), 110–135.

8. Anexos

8.1. Anexo 1: Código fuente en el programa R para los cálculos del modelo.

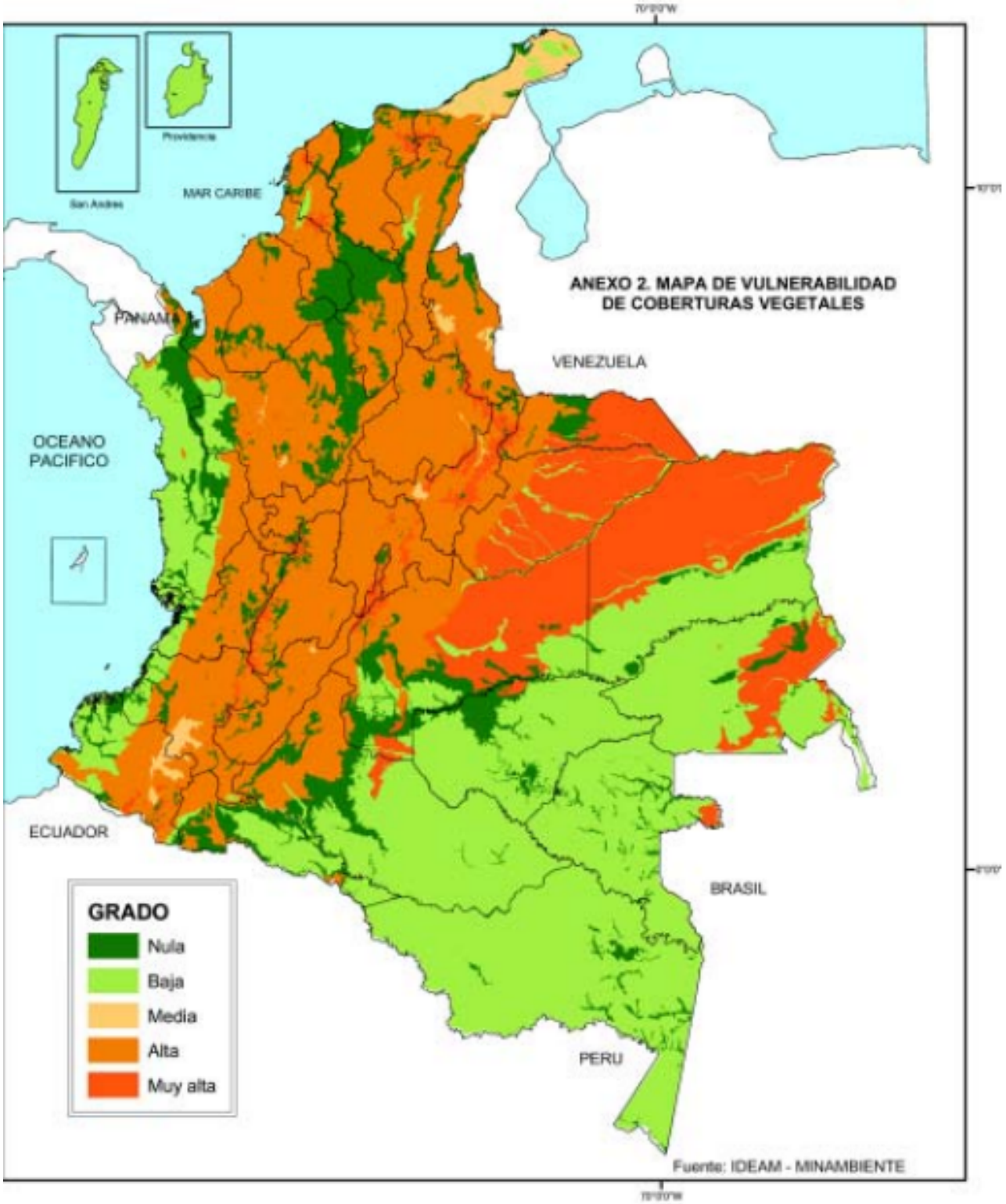
```
#############################################################################################
############################################## ANÁLISIS ESPACIAL ############################
#############################################################################################

library(coda)
library(spdep)
library(rgdal)
library(hglm)
library(maptools)
library(GISTools)
library(HistogramTools)
library(pgirmess)
library(strucchange)
library(car)
library(sqldf)
library(ggplot2)
library(dplyr)
library(stats)
library(pscl)
library(gamlss)
library(stats)
library(readxl)

##########################Lectura de los datos#############################
setwd("C:/Users/Familia Díaz Ovalle/Google Drive/Personal Javier/USTA/Proyecto/Bases")


base_total <- read.csv("C:/Users/Familia Díaz Ovalle/Google Drive/Personal Javier/USTA/I

n <- dim(base_total)[1]
names(base_total)
base_total$CODIGO_MUNICIPIO <- as.factor(base_total$CODIGO_MUNICIPIO)
base_total$CODIGO_DEPARTAMENTO <- as.factor(base_total$CODIGO_DEPARTAMENTO)
str(base_total)
attach(base_total)
##### capa municipios

municipios<-readOGR(dsn="C:/Users/Familia Díaz Ovalle/Google Drive/Personal Javier/USTA/
#municipios<- subset(municipios, municipios$COD_DEPTO == c("05","68"))
#municipios<- subset(municipios, municipios$ID_ESPACIA != 99999)

#merge de la bases y eliminacion de duplicados

data<- merge(municipios,base_total, by.x="ID_ESPACIA",by.y="CODIGO_MUNICIPIO")
data <-  data[!duplicated(data$ID_ESPACIA),]

#Conversion a coordenadas UTM
crs.geo <- CRS("+proj=tmerc +lat_0=4.599047222222222 +lon_0=-74.08091666666667 +k=1 +x_0
proj4string(data) <- crs.geo
final.utm=spTransform(data,CRS("+init=epsg:3724 +units=km"))
plot(final.utm)

#Centroides de las areas
```

```
centros=getSpPPolygonsLabptSlots(final.utm)
centroids <- SpatialPointsDataFrame(coords=centros, data=final.utm@data,
proj4string=CRS("+init=epsg:3724 +units=km"))
plot(centroids,add=T,pch=1,col=2)

#Create a k=4 nearest neighbor set
us.nb4<-knearneigh(coordinates(final.utm), k=4)
us.nb4<-knn2nb(us.nb4)
us.wt4<-nb2listw(us.nb4, style="W")

# Matriz de distancias entre los centriodes
W=dist(centros,up=T)
head(W)

# Matriz W de vecindades
final.nb=poly2nb(final.utm,queen=T)
plot(final.nb,coordinates(final.utm))
head(final.nb)

# Martiz W estilos

final.lw=nb2listw(final.nb)
final.lw
names(final.lw)
final.lw$weights


final.lwb=nb2listw(final.nb,style="B")
final.lwb
names(final.lwb)


final.lwb$weights

final.lwc=nb2listw(final.nb,style="C")
final.lwc
names(final.lwc)
final.lwc$weights

final.lwu=nb2listw(final.nb,style="U")
final.lwu
names(final.lwu)
final.lwu$weights

#Mapa con datos completos
marr_seq <- seqdef(biofam3c$married, start = 15)
attr(marr_seq, "cpal") <- c("#AB82FF", "#E6AB02", "#E7298A")

coorde <- merge(final.utm, centroids, by.x=ID_ESPACIA,by.y="ID_ESPACIA")
spplot(coorde, "CANT_INCENDIOS" , do.log = T, colorkey = TRUE)

########### Prueba de Correlacion
```

```
#Test Moran
MTest=moran.test(final.utm$CANT_INCENDIOS, final.lw, alternative="two.sided")
moran.plot(final.utm$CANT_INCENDIOS, final.lw)
#Test C Geary
geary.test(CANT_INCENDIOS, final.lw, alternative="two.sided")

##########################################################################################################
################################################## MODELO MIXTO ##########################################
##########################################################################################################

### OPERADOR DE RETARDO ESPACIAL
MATRIZ_VECINOS<-nb2mat(final.nb)
ESP_EFF<-(MATRIZ_VECINOS %*% CANT_INCENDIOS)

m1 <- gamlss(CANT_INCENDIOS~#0+
pb(MEAN_TEMP)
#+pb(TOT_PRES)
+pb(AREACOSECHADA)
#+ESP_EFF_2
#+MEAN_TEMP
#+MAX_PRES
+(TOT_PRES)
+DIAS_PRES
#+(AREAOFICIAL)
+AREABOSQUES
+(POBLACION_TOTAL)
,sigma.formula = ~-1
#+AREABOSQUES
#+AREACOSECHADA
#+MEAN_TEMP
#+AREAOFICIAL
+(ESP_EFF)
+(POBLACION_TOTAL)
,family=ZIP,method = mixed(1,100))

summary(m1)

qplot(CANT_INCENDIOS,
geom="histogram",
binwidth=5,
main="                    Histograma de la cantidad de incendios",
xlab="Incendios forestales",
fill=I("indianred1"),
col=I("red"))

m2 <- gamlss(CANT_INCENDIOS~0+
pb(MEAN_TEMP)
#+pb(TOT_PRES)
+pb(AREACOSECHADA)
+ESP_EFF_2
#+MEAN_TEMP
#+MAX_PRES
+(TOT_PRES)
```

```
+DIAS_PRES
+(AREAOFICIAL)
#+AREABOSQUES
#+pb(POBLACION_TOTAL)
,sigma.formula = ~0+
#+AREABOSQUES
#+AREACOSECHADA
#+MEAN_TEMP
#+pb(AREAOFICIAL)
+(ESP_EFF)
#+ESP_EFF_2
#+pb(MEAN_TEMP)
+(POBLACION_TOTAL)
#+MAX_PRES
#,nu.formula = ~1+ESP_EFF
#gmrf(CODIGO_MUNICIPIO, neighbour=final.nb)
#final.lw
,family=ZIP,method = mixed(1,100))

summary(m2)

exp(m1$mu.coefficients)

centiles.fan(m1, xvar=AREACOSECHADA, cent=c(3,10,25,50,75,90,97),
colors="terrain",ylab="Incendios", xlab="Temperatura Media"                #xlim=c(0,28)
)
windows()
plot(CANT_INCENDIOS~MEAN_TEMP, col="indianred1",pch=20,ylim=c(0,100),
ylab="Incendios", xlab="Temperatura Media",
main="Estimación no paramétrica para tempertura media")
lines(fitted(m1)[order(MEAN_TEMP)]~
MEAN_TEMP[order(MEAN_TEMP)],col="gray36",lwd=2)

windows()
plot(CANT_INCENDIOS~AREACOSECHADA, col="indianred1",pch=20,ylim=c(0,200),
ylab="Incendios", xlab="Area Cosechada",
main="Estimación no paramétrica para Area cosechada")
lines(yest[order(AREACOSECHADA)]~
AREACOSECHADA[order(AREACOSECHADA)],col="gray36",lwd=2)

windows()
term.plot(m1, pages=1, ask=FALSE)

windows()
plot(m1)
```

8.2. Anexo 2: Mapa de Vulnerabilidad de Coberturas vegetales



Vulnerabilidad de coberturas vegetales de Colombia.
Fuente: IDEAM - MINAMBIENTE

8.3. Anexo 3: Carta Autorización de la Utilización Información

Señores
Comité de opción de grado.
Facultad de Estadística
Universidad Santo Tomás.

Cordial saludo,

Los datos utilizados para este estudio provienen de varias institucionales gubernamentales encargadas de su administración.

- La información de Incendios forestales por municipio fue suministrada por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), mediante solicitud con radicado número 20179050051402 del 04 de Julio de 2017.
- La información de la base de datos de climatología por municipio fue suministrada por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), mediante solicitud con radicado número 20179050012812. del 13 de Marzo de 2017.
- Las cifras de población para los departamentos de Antioquia y Santanter, se tomaron de la publicación de proyecciones del DANE hasta el año 2020, basadas en el censo de población realizado en el 2005.

Estos datos son de carácter público, por lo cual se da autorización para su divulgación.

Edinson Javier Díaz Prieto
C.C. 1014217705
edinsondiaz@usantotomas.edu.co