UNIVERSIDAD NACIONAL DE COLOMBIA

Modelación longitudinal de casos de dengue en Colombia, mediante modelos de conteo Poisson y ZIP de efectos Mixtos

David Esteban Morales Suarez

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2022

Modelación longitudinal de casos de dengue en Colombia, mediante modelos de conteo Poisson y ZIP de efectos Mixtos

David Esteban Morales Suarez

Trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en Ciencias Estadística

Director:
Edilberto Cepeda Cuervo, Ph.D
Doctor en Matemática

Línea de Investigación:
Datos Longitudinales

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2022

A mis padres y hermanos que siempre me han brindado su apoyo.

Agradecimientos

A Dios y a mi familia por su apoyo incondicional en todo este proceso.

Al profe Edilberto Cepeda por su orientación en la realización del presente documento.

A mis profesores y compañeros del departamento de Estadística.

Resumen

Modelación longitudinal de casos de dengue en Colombia, mediante modelos de conteo Poisson y ZIP de efectos Mixtos

En el presente trabajo de tesis de maestría se lleva a cabo la modelación de casos de dengue en Colombia por municipio mediante Modelos Lineales Generalizados de Efectos Mixtos (GLMM) aplicados a datos longitudinales. Partiendo de que la naturaleza de los datos son especificamente datos de conteo, se consideran los modelos de regresión Poisson, binomial negativo, Poisson inflado de ceros (ZIP) y binomial negativo inflado de ceros (ZINB) de efectos mixtos; para el ajuste de los modelos se tomaron 266 municipios del país que contaban con las siguientes variables explicativas: temperatura, precipitación, densidad poblacional e índice de calidad del agua (IRCA); dichos modelos se abordan desde una perspectiva clásica. Adicionalmente, se realiza el estudio de bondad de ajuste de estos modelos, su análisis de residuales así como el respectivo análisis de diagnóstico de influencia, teniendo presente que en este caso cada observación se encuentra anidada dentro de un municipio. Además, se clasifican los municipios en clústeres según la cantidad de casos acumulados siguiendo la metodología kml (k-means para datos longitudinales) presentando en cada clúster tendencias generales de los casos de dengue en el trascurso del tiempo respecto al rango que asumen las variables explicativas. Se encontró que el modelo que presentó el mejor ajuste es el modelo ZIP de efectos mixtos, las variables explicativas consideradas resultaron ser significativas en el modelo, esto es, ejercen un efecto sobre la cantidad de casos de dengue.

Palabras clave: Datos de conteo, Modelos Lineales Generalizados de Efectos Mixtos, Datos influyentes, Dengue.

Abstract

Longitudinal modeling of dengue cases in Colombia, using Mixed effects Poisson and ZIP counting models

In this master's thesis, the modeling of dengue cases in Colombia by municipality is carried out using Generalized Linear Mixed Effects Models (GLMM) applied to longitudinal data. Assuming that the nature of the data is specifically counting data, the Poisson, negative binomial, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models of mixed effects are considered; For the adjustment of the models, 266 municipalities of the country were taken into account that had the following explanatory variables: temperature, precipitation, population density and water quality index (IRCA); These models are approached from a classical perspective. Additionally, the goodness-of-fit study of these models, their residual analysis as well as the respective influence diagnostic analysis is carried out, bearing in mind that in this case each observation is nested within a municipality. In addition, the municipalities are classified into clusters according to the number of accumulated cases following the kml methodology (k-means for longitudinal data), presenting in each cluster general trends of dengue cases over time with respect to the range assumed by the explanatory variables. It was found that the model that presented the best fit is the ZIP model of mixed effects, the explanatory variables considered turned out to be significant in the model, that is, they exert an effect on the number of dengue cases.

Keywords: Count data, Generalized Linear Mixed Effects Models, Influential data, Dengue.

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

El dengue es una enfermedad endémica causada por una familia de virus conocida cómo *arbovirus* cuyo vector de transmisión son los artrópodos, específicamente mosquitos de género Aedes. Actualmente se considera el dengue como un problema de salud pública debido a la rápida expansión geográfica del virus y severidad de la progresión de la enfermedad, afectando más de 100 países en el mundo, principalmente zonas como Asia y América.

Diversos factores ambientales, como cambios climáticos, la temperatura y la humedad influyen en el desarrollo larvario y la replicación del virus; además, aspectos sociales, culturales y económicos de los que se destaca por ejemplo la acelerada urbanización y alta densidad poblacional [Simmons et al., 2012] así como la falta de medidas de control de los mosquitos que también desarrollan un papel importante en la rápida transmisión de la enfermedad.

El dengue se caracteriza por presentar 4 serotipos (DENV-1, DENV-2, DENV-3, DENV-4) identificados, los cuales pueden producir la enfermedad y la muerte debido a su variabilidad genética [Álvarez and Vargas, 2019]. La vía de transmisión de la enfermedad se produce mediante picadura de la hembra del mosquito. Este vector pone sus huevos en agua limpia siendo este un tipo de mosquito que se reproduce en cualquier recipiente artificial o natural que contenga agua. En cuanto a los factores ambientales que propician la proliferación del vector se encuentran temperaturas de 15 a 40 grados y una humedad relativa alta; dentro de los factores sociales, una densidad poblacional de moderada a alta, así como agua almacenada por más de 7 días y la intermitente disponibilidad del agua, entre otros. Su ciclo de vida se puede completar en un periodo aproximado de 7 a 10 días, siendo la hembra responsable de trasmitir la enfermedad porque usa la sangre como medio de desarrollo de sus óvulos (OPS/OMS) [Ochoa et al., 2015].

Existen varios métodos para la detección del virus del dengue, por ejemplo métodos que permitan detectar los niveles de inmunoglobulinas específicas en el cuerpo, que aumentan durante el curso de la infección [Ochoa et al., 2015]. Para que la enfermedad evolucione, debe existir un incremento en los focos del vector, una alta densidad de las hembras adultas, disponibilidad de huésped y una susceptibilidad innata a la infección. Por los anteriores motivos, es común que la enfermedad se presente especialmente en lugares de bajos recursos y pobreza, pues el vector puede proliferar fácilmente [Velandia and Castellanos, 2011]. El

déficit de los programas de control del vector, la urbanización no planificada, el crecimiento acelerado de la población y la existencia de una infraestructura de salud deficiente en la mayoría de los países ha permitido que prolifere la enfermedad a gran escala.

En Colombia, el primer caso de dengue se registró en 1971 [Mendez et al., 2012]. Según la Pan American Health Organization (PAHO), en Latinoamérica se reportaron 2,100,473 (394.2 por 100,000 habitantes) casos de dengue durante el año 2020. Para el año 2019 se presentó la mayor incidencia con 3,140,872 casos. Las regiones colombianas donde predominan casos severos de dengue son la región central, costa atlántica y Orinoquía; sin embargo, en otras regiones no se poseen los recursos suficientes para acceder a la salud [Gutierrez et al., 2020].

El Sistema Nacional de Vigilancia en Salud Pública - Sivigila, realiza de forma constante el seguimiento y monitoreo de los casos de dengue, debido entre otros aspectos a las características geográficas del país y la presencia del vector en la mayoría de los municipios. Con base en esta información, el Instituto Nacional de Salud presenta constantemente boletines epidemiológicos en los que realiza un análisis descriptivo detallado por departamento y resaltando además los municipios con mayor incidencia.

Por su parte, [Castrillón et al., 2015] realizaron un análisis de la incidencia de dengue en Colombia entre los años 2004 y 2013, en el que resaltan los periodos y los departamentos en los que se reporta el mayor número de casos. Consideraron además, las variables climatológicas precipitación y temperatura y hallaron su correlación con la incidencia de dengue mediante correlación de Pearson. Encontraron que dicha infección presenta un comportamiento cíclico que posiblemente se repite cada 3 o 4 años y que puede ser atribuido a cambios sociales y climatológicos, sin embargo, no encontraron significancia estadística entre las variables climáticas estudiadas y la incidencia de dengue.

[Rúa et al., 2013] desarrollaron un modelo ARIMA que describe los camabios en el comportamiento epidemiológico del dengue para la ciudad de Medellín. De las variables meteorológicas que consideraron, mostraron que la precipitación fue la única variable que influyó en los cambios de incidencia. De igual manera, para la ciudad de Montería [Cassab et al., 2010] emplearon un modelo de regresión Poisson tomado como variable respuesta los casos de dengue promedio por mes y como variables explicativas la pluviosidad en mm promedio mensual, la temperatura en grados centígrados promedio mensual y la humedad relativa promedio en porcentaje, tomando como medida de bondad de ajuste el coeficiente R^2. Hallaron que la temperatura no influyó en los casos de dengue en Montería y lo asocian a que la temperatura óptima de transmisión ocurre en temperaturas superiores a 30 °C mientras que la temperatura promedio encontrada en Montería es de 27 °C. La pluviometría y la humedad relativa tampoco resultaron tener efecto sobre los casos de dengue.

Entre otros trabajos se destacan [Vélez et al., 2006] y [López et al., 2012] quienes implementaron modelos matemáticos para representar la dinámica de la trasmisión del virus mientras que [Medina and Ramos, 2017] adaptaron un modelo matemático para explicar la afectación e identificar patrones relevantes de la difusión del dengue en la zona urbana del municipio de Neiva.

El propósito del presente trabajo es aplicar una metodología que permita modelar el comportamiento de la cantidad de casos de contagiados de dengue nominal a partir del año 2007 hasta el año 2018 en Colombia por municipio y así poder explicarlo a través de distintas variables climatológicas como la precipitación promedio y la temperatura promedio, variables geográficas como la densidad poblacional y sociales como el Índice de Riesgo para la Calidad del Agua potable (IRCA), esta última, permite tener una idea del saneamiento básico de cada municipio.

El documento está compuesto de 6 capítulos y 3 apéndices. En el capítulo 2 se presenta el marco teórico de los modelos Lineales Generalizados de Efectos Mixtos (GLMM) centrando la atención especialmente en el modelo Poisson de efectos mixtos y ZIP de efectos mixtos así como su respectiva estimación de los parámetros. En el capítulo 3 se abordan los métodos de diagnóstico haciendo énfasis en la metodología que se aplica para los modelos GLMM. En el capítulo 4 se ilustra la metodología kml (k-means para datos longitudinales) con la cual se clasifican los municipios en clústeres según la cantidad de casos acumulados de dengue y se presentan tendencias generales de la variable respuesta a través del tiempo a partir de distintos intervalos de cada variable explicativa por clúster. En el capítulo 5 se presenta la respectiva modelación con su correspondiente análisis de influencia e interpretación de resultados. Finalmente, en el capítulo 6 se presentan las conclusiones y recomendaciones para futuros trabajos.

2. Modelos de Regresión Poisson y ZIP

En este capítulo, se presenta en primer lugar el Modelo de Regresión Poisson y el Modelo de Regresión Poisson inflado de ceros (ZIP) así como la estimación clásica de los parámetros con base en el desarrollo presentado por autores como [McCullagh and Nelder, 1989] y [Lambert, 1992] en el contexto de Modelos Lineales Generalizados (GLM). Luego, se presentan dichos modelos para datos longitudinales que incluyen además, efectos aleatorios, esto es, se presentan modelos de efectos mixtos para conteos los cuales se emplearán posteriormente en el ajuste de los datos; estos modelos hacen parte de los Modelos Lineales Generalizados de efectos Mixtos (GLMM).

2.1. Modelos Lineales Generalizados

En numerosos estudios, la variable de interés se centra en la frecuencia con la que ocurre un comportamiento específico, por ejemplo, el número de niños que nacen con alguna patología determinada, el número de fallas de un máquina en cierto proceso, etc. Naturalmente en estas situaciones el comportamiento probabilístico es distinto al de la distribución normal. Por tal motivo, [McCullagh and Nelder, 1989] introducen los modelos lineales generalizados, en los cuales precisamente la variable de interés está relacionada por ejemplo con conteos (el foco de nuestro estudio); se hace ecesario además, una distribución de probabilidad que explique el comportamiento de tales variables como lo es la distribución Poisson. A continuación se presenta el Modelo de Regresión Poisson así como la estimación de parámetros correspondiente.

2.1.1. Modelo de Regresión Poisson

Sea $Y_1, ..., Y_n$ variables aleatorias independientes con $Y_i \sim P(\mu_i)$ $i = 1, ..., n$ y sea $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ip})'$ el i-ésimo $(p+1)$ vector de covariables, se define el modelo de regresión Poisson de la siguiente manera,

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta},$$

donde $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$ es un vector $(p+1)$ de parámetros desconocidos y $g(\cdot)$ es la función de enlace (link), que en este caso corresponde al enlace canónico, es decir $\eta_i = log(\mu_i)$.

Estimación

La función de verosimilitud para n observaciones de la variable de interés es:

$$L = \prod_{i=1}^{n} \frac{e^{-\mu_i} \cdot \mu_i^{y_i}}{y_i!}$$

Así, la función de logverosimilitud es

$$l = \log L = \sum_{i=1}^{n} [-\mu_i + y_i \log(\mu_i) - \log(y_i!)]$$

En consecuencia, sin tener en cuenta el término $log(y_i!)$ tenemos:

$$l = \sum_{i=1}^{n} [-\mu_i + y_i \log(\mu_i)]$$

Hallando la primera derivada parcial con respecto a β_k y simplificando se tiene:

$$\begin{aligned} \frac{\partial l}{\partial \beta_k} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_k} \\ &= \sum_{i=1}^{n} \left(\frac{y_i}{\mu_i} - 1 \right) \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ik} \end{aligned}$$

La matriz de información de Fisher para $\boldsymbol{\beta}$ está dada por:

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial \beta_k \partial \beta_m} \right) &= -E\left(\sum_{i=1}^{n} -\frac{y_i}{\mu_i^2} \cdot \frac{\partial \mu_i}{\partial \beta_k} \cdot \frac{\partial \mu_i}{\partial \beta_m} \right) \\ &= -E\left(\sum_{i=1}^{n} -\frac{y_i}{\mu_i^2} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ik} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{im} \right) \\ &= -E\left(\sum_{i=1}^{n} -\frac{y_i}{\mu_i^2} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \cdot x_{ik} \cdot x_{im} \right) \\ &= \sum_{i=1}^{n} \frac{1}{\mu_i} \cdot (\mu_i)^2 \cdot x_{ik} \cdot x_{im} \\ &= \sum_{i=1}^{n} \mu_i \cdot x_{ik} \cdot x_{im} \end{aligned}$$

Entonces la matriz de información de Fisher es: $Inf = \boldsymbol{X'WX}$, donde $\boldsymbol{W}$ es la matriz diagonal de pesos definida como:

$$\boldsymbol{W} = diag(\mu_i)$$

Por otra parte, empleando procedimientos iterativos como el algoritmo de Newton-Raphson o el algoritmo de Fisher se puede obtener la estimación de parámetros $\hat{\boldsymbol{\beta}} = \boldsymbol{(X'WX)^{-1}X'WZ}$, donde la variable dependiente $\boldsymbol{Z}$ tiene elementos:

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

para lo cual, a partir de valores iniciales, se determina la estimación de máxima verosimilitud que satisface $\hat{\boldsymbol{\beta}}$ y resolviéndola iterativamente.
Un estimador de la matriz de varianza-covarianza de $\hat{\boldsymbol{\beta}}$, $Var(\hat{\boldsymbol{\beta}})$, es el inverso de la matriz de información, $I(\hat{\boldsymbol{\beta}})$, esto es, como se mostró ateriormente $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'WX})^{-1}$.

2.1.2. Modelo Poisson inflado de ceros ZIP

En algunas situaciones prácticas donde se analizan datos de conteo, se puede presentar un exceso de ceros mucho mayor de lo esperado que el supuesto por la distribución Poisson, esta característica se denomina "inflado de cero". Estos modelos se pueden ver como una mixtura con una distribución degenerada cuya masa se concentra en cero [Rose et al., 2013].

Distintos autores han propuesto modificaciones del modelo Poisson para tratar estos ceros adicionales. Por ejemplo, en modelos que incluyen covariables, [Mullahy, 1986] y [King, 1989] propusieron modelos de regresión Hurdle (obstáculo) de Poisson. [Heilbron, 1989], introdujo un modelo de Poisson cero-alterado ZAP (zero-altered Poisson). Por su parte, [Lambert, 1992] propuso un modelo de Poisson inflado de ceros, ZIP (zero-inflated Poisson) el cual es una mixtura de modelos de regresión logística y de regresión Poisson donde la parte logística contribuye a la probabilidad de un recuento de cero y la parte Poisson contribuye a la frecuencia de las demas observaciones. A continuación se presenta precisamente el modelo Poisson inflado de ceros (ZIP).

Modelo

Sea $Y_1, ..., Y_n$ variables aleatorias independientes, tales que $Y_i \sim ZIP(\pi_i, \mu_i)$, esto es, la variable aleatoria Y_i tiene una distribución Poisson inflada de ceros definida como:

$$f(y_i) = P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i} & \text{si } y_i = 0 \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{si } y_i \geq 1 \end{cases} \tag{2.1.1}$$

Donde π_i es la probabilidad de exceso de ceros, el valor esperado y la varianza de Y_i están dados por: $E(Y_i) = (1-\pi_i)\mu_i$ y $Var(Y_i) = \mu_i(1-\pi_i)(1+\pi_i\mu_i)$. Además, incluyendo covariables se usan las funciones de enlace *logit* y *log* para π_i y μ_i respectivamente:

$$\begin{aligned} logit(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\omega}_i'\boldsymbol{\gamma} \\ \log(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}, \end{aligned} \tag{2.1.2}$$

donde $\boldsymbol{\omega}_i = (1, \omega_{i1}, ..., \omega_{is})'$ y $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ip})'$ son vectores de covariables de tamaño $(s+1)$ y $(p+1)$ respectivamente y $\boldsymbol{\gamma} = (\gamma_0, ..., \gamma_s)'$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$ son vectores de parámetros desconocidos de tamaño $(s+1)$ y $(p+1)$ respectivamente.

Estimación

Para la estimación por máxima verosimilitud, siguiendo a autores como [Lambert, 1992], [Cameron and Trivedi, 2013] y [Hossain and Howlader, 2015], se tiene que la función de verosimitud está dada por:

$$L(\boldsymbol{\mu}, \boldsymbol{\pi}; \boldsymbol{Y}) = \prod_{y_i=0} P(Y_i = y_i) \prod_{y_i>0} P(Y_i = y_i)$$

y la función de logverosimilitud correspondiente es:

$$l(\boldsymbol{\mu}, \boldsymbol{\pi}; \boldsymbol{Y}) = \log(L(\boldsymbol{\mu}, \boldsymbol{\pi}; \boldsymbol{Y})) = \sum_{y_i=0} \log(\pi_i + (1-\pi_i)e^{-\mu_i}) + \sum_{y_i>0} \log\left((1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right) \tag{2.1.3}$$

Teniendo en cuenta las expresiones dadas en (2.1.2), se deduce que $\pi_i = \frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}$ y $\mu_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$ con lo cual tenemos:

$$\begin{aligned}
\log(\pi_i + (1-\pi_i)e^{-\mu_i}) &= \log\left(\frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}} + \left(1 - \frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right) e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right) \\
&= \log\left(\frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}} + \frac{e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right) \\
&= \log\left(\frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}} + e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right) \\
&= \log(e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}} + e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}) - \log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})
\end{aligned}$$

y

$$\begin{aligned}
\log\left((1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right) &= \log(1-\pi_i) - \mu_i + y_i\log(\mu_i) - \log(y_i!) \\
&= \log\left(1 - \frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right) - e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \log(y_i!) \\
&= \log\left(\frac{1}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right) - e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \log(y_i!) \\
&= -\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}) - e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \log(y_i!)
\end{aligned}$$

Reemplazando las anteriores expresiones en la función de logverosimilitud dada en (2.1.3), se tiene:

$$\begin{aligned} l(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{Y}) &= \sum_{y_i=0}\left[\log(e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}+e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}})-\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})\right]+\sum_{y_i>0}\left[y_i\boldsymbol{x}_i'\boldsymbol{\beta}-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}-\log(y_i!)-\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})\right] \\ &= \sum_{y_i=0}\left[\log(e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}+e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}})\right]+\sum_{y_i>0}\left[y_i\boldsymbol{x}_i'\boldsymbol{\beta}-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}-\log(y_i!)\right]-\sum_{i=1}^{n}\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}) \end{aligned} \tag{2.1.4}$$

La maximización de $l(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{Y})$ no es sencilla especialmente por la suma de exponenciales del primer término que involucra tanto a $\boldsymbol{\beta}$ como a $\boldsymbol{\gamma}$. Para resolver este inconveniente, [Lambert, 1992] propone emplear el algoritmo EM para maximizar la función de logverosimilitud. Para ello, en primer lugar supone que se conoce cuáles ceros provienen del estado perfecto y cuáles del estado Poisson; de este modo, incluye una variable latente U_i, tal que $U_i = 1$ cuando Y_i proviene del estado perfecto cero y $U_i = 0$ cuando Y_i proviene del estado Poisson. Luego, formula una expresión para la logverosimilitud de los datos completos.

$$\begin{aligned} l_c(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{U}) &= \log\left[\prod_{i=1}^{n}P(Y_i=y_i,U_i=u_i|\boldsymbol{x}_i,\boldsymbol{\omega}_i,\boldsymbol{\gamma},\boldsymbol{\beta})\right] \\ &= \log\left[\prod_{i=1}^{n}P(Y_i=y_i|U_i=u_i,\boldsymbol{x}_i,\boldsymbol{\omega}_i,\boldsymbol{\gamma},\boldsymbol{\beta})\cdot P(U_i=u_i|\boldsymbol{x}_i,\boldsymbol{\omega}_i,\boldsymbol{\gamma},\boldsymbol{\beta})\right] \\ &= \log\left[\prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right)^{u_i}\left(\frac{1}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\cdot\frac{e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}e^{\boldsymbol{x}_i'\boldsymbol{\beta}y_i}}{y_i!}\right)^{u_i-1}\right] \\ &= \sum_{i=1}^{n}\log\left(\frac{e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}}\right)^{u_i}+\sum_{i=1}^{n}\log\left(\frac{e^{-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}e^{y_i\boldsymbol{x}_i'\boldsymbol{\beta}}}{(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})y_i!}\right)^{1-u_i} \\ &= \sum_{i=1}^{n}u_i\boldsymbol{\omega}_i'\boldsymbol{\gamma}-u_i\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})+\sum_{i=1}^{n}(1-u_i)\left[y_i\boldsymbol{x}_i'\boldsymbol{\beta}-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}-\log(y_i!)-\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})\right] \\ &= \sum_{i=1}^{n}u_i\boldsymbol{\omega}_i'\boldsymbol{\gamma}-u_i\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}) \\ &+\sum_{i=1}^{n}(1-u_i)(y_i\boldsymbol{x}_i'\boldsymbol{\beta}-e^{\boldsymbol{x}_i'\boldsymbol{\beta}})-\sum_{i=1}^{n}(1-u_i)log(y_i!)-\sum_{i=1}^{n}(1-u_i)\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}}) \\ &= \sum_{i=1}^{n}\left(u_i\boldsymbol{\omega}_i'\boldsymbol{\gamma}-\log(1+e^{\boldsymbol{\omega}_i'\boldsymbol{\gamma}})\right)+\sum_{i=1}^{n}(1-u_i)(y_i\boldsymbol{x}_i'\boldsymbol{\beta}-e^{\boldsymbol{x}_i'\boldsymbol{\beta}})-\sum_{i=1}^{n}(1-u_i)log(y_i!) \\ &= l_c(\boldsymbol{\gamma};\boldsymbol{Y},\boldsymbol{U})+l_c(\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{U})-\sum_{i=1}^{n}(1-u_i)log(y_i!) \end{aligned} \tag{2.1.5}$$

De este modo, es posible maximizar la logverosimilitud por separado. Para esto, se implementa el algoritmo EM de tal forma que la logverosimilitud dada en (2.1.3) se determina

maximizando iterativamente. En primer lugar, se dan valores iniciales de $(\boldsymbol{\beta},\boldsymbol{\gamma})$, luego, para el Paso E, se calcula la esperanza de U_i mediante los valores iniciales y se toma este valor como estimador de U_i. Para el Paso M, se toma la estimación de U_i para maximizar $l_c(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{U})$ lo cual da un estimador de máximo verosimilud sin restricción para $\boldsymbol{\beta}$ y para $\boldsymbol{\gamma}$ y con estos valores se vuelve al paso E. Una vez que los valores esperados de U_i convergen, las estimaciones de $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ convergen y la iteración se detiene. Las estimaciones de la iteración final son los estimadores de máximo verosimilitud $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$. El desarrollo más detallado de este método se presenta en [Lambert, 1992].

A partir de las funciones de enlace dadas en (2.1.2) y haciendo $\boldsymbol{x}_i'\boldsymbol{\beta} = \sum_j x_{ij}\beta_j$ y $\boldsymbol{\omega}_i'\boldsymbol{\gamma} = \sum_m w_{im}\gamma_m$ se obtienen las siguientes expresiones de las derivadas parciales siguiendo a [Hossain and Howlader, 2015]:

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_k\partial\beta_h} &= \sum_{y_i=0} \frac{\left(-e^{-e^{\sum_j x_{ij}\beta_j}} e^{\sum_j x_{ij}\beta_j}\cdot x_{ik}\right)\left(-e^{\sum_j x_{ij}\beta_j}\cdot x_{ih} + x_{ih}\right)\left(e^{\sum_m w_{im}\gamma_m} + e^{-e^{\sum_j x_{ij}\beta_j}}\right)}{\left(e^{\sum_m w_{im}\gamma_m} + e^{-e^{\sum_j x_{ij}\beta_j}}\right)^2} \\
&\quad - \sum_{y_i=0} \frac{\left(-e^{-e^{\sum_j x_{ij}\beta_j}} e^{\sum_j x_{ij}\beta_j}\cdot x_{ik}\right)\left(-e^{-e^{\sum_j x_{ij}\beta_j}} e^{\sum_j x_{ij}\beta_j}\cdot x_{ih}\right)}{\left(e^{\sum_m w_{im}\gamma_m} + e^{-e^{\sum_j x_{ij}\beta_j}}\right)^2} + \sum_{y_i>0} -e^{x_{ij}\beta_j}\cdot x_{ik}x_{ih} \\
&= \sum_{y_i=0} \frac{\left(-e^{-\mu_i}\mu_i\cdot x_{ik}\right)\left(-\mu_i\cdot x_{ih} + x_{ih}\right)\left(\frac{\pi_i}{1-\pi_i} + e^{-\mu_i}\right) - \left(-e^{-\mu_i}\mu_i\cdot x_{ik}\right)\left(-e^{-\mu_i}\mu_i\cdot x_{ih}\right)}{\left(\frac{\pi_i}{1-\pi_i} + e^{-\mu}\right)^2} \\
&\quad + \sum_{y_i>0} -\mu_i\cdot x_{ik}x_{ih}
\end{aligned}$$

Simplificando se tiene:

$$\frac{\partial^2 l}{\partial\beta_k\partial\beta_h} = \sum_{y_i=0} \frac{-e^{-\mu_i}\mu_i(1-\pi_i)\left[(1-\mu_i)\pi_i + (1-\pi_i)e^{-\mu_i}\right]x_{ik}x_{ih}}{\left[\pi_i + (1-\pi_i)e^{-\mu_i}\right]^2} + \sum_{y_i>0}(-\mu_i)x_{ik}x_{ih}$$

De manera análoga:

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_k\partial\gamma_r} &= \sum_{y_i=0}\left[\frac{e^{-\mu_i}\mu_i\cdot x_{ik}w_{ir}}{\left(\pi_i + (1-\pi_i)e^{-\mu_i}\right)^2}\right] \\
\frac{\partial^2 l}{\partial\gamma_r\partial\gamma_q} &= \sum_{y_i=0}\left[\frac{-(1-e^{-\mu_i})^2\cdot w_{ir}w_{iq}}{\left(\pi_i + (1-\pi_i)e^{-\mu_i}\right)^2}\right] + \sum_{y_i>0}\left[\frac{-w_{ir}w_{iq}}{(1-\pi_i)^2}\right]
\end{aligned}$$

De este modo, la matriz de información para el modelo ZIP esta dada por:

$$I(\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{bmatrix} -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) & -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}}\right) \\ -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}}\right) & -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right) \end{bmatrix}$$

2.2. Modelos Lineales Generalizados de efectos Mixtos

Según [Hedeker and Gibbons, 2006], una limitación importante del modelo ZIP es que no tiene en cuenta los efectos aleatorios, además, el modelo de regresión Poisson supone independencia en los datos y, en el contexto de datos longitudinales este supuesto no se cumpliría. Por otra parte, cuando se supone un efecto aleatorio en los modelos de mixtura y, en particular en los modelos ZIP, además de disminuir una posible sobredispersión, tambien aborda la correlación inherente [Monod, 2014].

Diversos autores se han enfocado precisamente en el problema de incluir efectos aleatorios, por ejemplo, [Breslow, 1984] propuso un modelo de regresión Poisson con efectos aleatorios que siguen una distribución normal, mientras que [Lawless, 1987] propuso para la distribución Poisson agregar un factor multiplicativo aleatorio de la media el cual seguía una distribución gamma. Posteriormente, [Siddiqui, 1996] desarrolló un modelo de regresión Poisson de efectos mixtos para datos de conteo agrupado y comparó modelos con efectos aleatorios que tenían distribución normal y distribución gamma.

2.2.1. Modelo de Regresión Poisson de efectos mixtos

Sea Y_{ij} observaciones no negativas con $i = 1, ..., n$ individuos, $j = 1, ..., n_i$ observaciones repetidas para el i-ésimo individuo y $N = \sum_{i=1}^{n} n_i$ el total de observaciones y sea además $\boldsymbol{x}_{ij} = (1, x_{ij1}, ..., x_{ijp})'$ un vector de covariables de tamaño $(p + 1)$, el modelo de regresión Poisson de efectos mixtos está dado por:

$$g(\mu_{ij}) = \log(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \nu_i, \tag{2.2.1}$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ es un vector de parámetros de tamaño $(p + 1)$ y $g(\cdot)$ corresponde a la función de enlace canónica, ν_i es un efecto aleatorio, tal que es independiente del vector de covariables $\boldsymbol{x}_{ij}$ y se asume que $\nu_i \sim N(0, \sigma^2)$.

Estimación

Siguiendo a autores como [Fitzmaurice et al., 2011] y [Molenberghs and Verbeke, 2005], dado que los efectos aleatorios no son observados, la inferencia sobre β y σ^2 se basa en la función de verosimilitud marginal que se obtiene integrando respecto de los efectos aleatorios. Para ello, en primer lugar tenemos que la función de densidad condicional de las n_i respuestas individuales para el i-ésimo individuo es:

$$\begin{aligned} f(\boldsymbol{y}_i|\nu_i, \boldsymbol{\beta}) &= \prod_{j=1}^{n_i} f(y_{ij}|\nu_i, \boldsymbol{\beta}) \\ &= \prod_{j=1}^{n_i} \frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} \end{aligned} \tag{2.2.2}$$

De este modo, la contribución del i-ésimo individuo a la verosimilitud (función de densidad marginal de $\boldsymbol{y}_i$) está dada por:

$$f(\boldsymbol{y}_i|\boldsymbol{\beta},\sigma^2) = \int f(\boldsymbol{y}_i|\nu_i,\boldsymbol{\beta})f(\nu_i|\sigma^2)d\nu_i \tag{2.2.3}$$
$$= \int \prod_{j=1}^{n_i} f(y_{ij}|\nu_i,\boldsymbol{\beta})f(\nu_i|\sigma^2)d\nu_i$$

Luego, la función de verosimilitud marginal es:

$$L(\boldsymbol{\beta},\sigma^2) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{\beta},\sigma^2) \tag{2.2.4}$$
$$= \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f(y_{ij}|\nu_i,\boldsymbol{\beta})f(\nu_i|\sigma^2)d\nu_i$$
$$= \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} \frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} f(\nu_i|\sigma^2)d\nu_i \tag{2.2.5}$$

Los estimadores de máxima verosimilitud son los valores que maximizan L; sin embargo, en este caso no existen soluciones simples de forma cerrada [Fitzmaurice et al., 2011], para esto, es necesario realizar técnicas de integración numérica para maximizar la función de verosimilitud.

Una de las técnicas de integración numérica que se pueden emplear se conoce como *cuadratura gaussiana*, con la cual se realiza una aproximación a la integral que se presenta en la función de verosimilitud marginal (2.2.5) mediante una suma ponderada de la siguiente manera:

$$L(\boldsymbol{\beta},\sigma^2) \approx \prod_{i=1}^{n} \sum_{k=1}^{K} f(\boldsymbol{y}_i|\delta_i = z_k)w_k$$
$$\approx \prod_{i=1}^{n} \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} \frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ij}}}{y_{ij}!}\right] w_k \tag{2.2.6}$$

Donde $\delta_i = \nu_i/\sigma$ de tal forma que $\delta_i \sim N(0,1)$, $\mu_{ijk} = e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\sigma z_k}$ y además, z_k y w_k corresponden a los puntos de evaluación y los pesos en k respectivamente, conocidos como los valores de la cuadratura de Gauss-Hermite, para $k = 1, ..., K$. El número de valores de cuadratura (K) determinan el grado de precisión de la aproximación, entre más puntos se empleen más precisa será dicha aproximación, pero la carga computacional también aumenta. Los puntos de evaluación z_k y los pesos w_k se muestran en tablas, además, en [Press et al., 1992] se presenta un algoritmo para calcular todos los z_k y w_k para cualquier valor K.

Un problema que tiene el método de cuadratura es que a medida que aumenta el número de efectos aleatorios, puede implicar la suma de un gran número de puntos. Por ejemplo, si se tiene solo un efecto aleatorio ($r = 1$) la solución con cuadratura requiere la suma sobre K puntos en relación con la solución de efectos fijos, mientras que si tiene más efectos aleatorios ($r > 1$) la cuadratura se realiza sobre K^r puntos lo que lo vuelve computancionalmente pesado [Hedeker and Gibbons, 2006].

Para abordar el problema de escoger el número de puntos de cuadratura necesarios para realizar la estimación, varios autores han propuesto el método conocido como *cuadratura adaptativa* el cuál emplea un número de puntos reducido por dimensión (alrededor de 3) que se adaptan a la ubicación y dispersión de la distribución a integrar. Para un desarrollo más detallado se puede consultar [Pinheiro and Bates, 1995].

A partir de la expresión dada en (2.2.6), la logverosimitud es aproximadamente:

$$\begin{aligned}\log(L(\boldsymbol{\beta}, \sigma^2)) &\approx \log \prod_{i=1}^{n} \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ij}}}{y_{ij}!} \right] w_k \\ &\approx \sum_{i=1}^{n} \log \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ij}}}{y_{ij}!} \right] w_k \end{aligned} \tag{2.2.7}$$

Las primeras derivadas de la logverosimilut son:

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{h(\boldsymbol{y}_i)} \sum_{k=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \mu_{ijk}) \boldsymbol{x}_{ij} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ij}}}{y_{ij}} w_k \tag{2.2.8}$$

$$\frac{\partial \log L}{\partial \sigma} = \sum_{i=1}^{n} \frac{1}{h(\boldsymbol{y}_i)} \sum_{k=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \mu_{ijk}) \delta_i \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ij}}}{y_{ij}} w_k \tag{2.2.9}$$

Con $h(\boldsymbol{y}_i) \approx \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ij}}}{y_{ij}!} \right] w_k$, la densidad marginal de $\boldsymbol{y}_i$.

Los parámetros del modelo se pueden estimar igualando las anteriores derivadas a cero y resolviendo de forma iterativa utilizando el método de puntuación de Fisher o Newton-Raphson.

Estimación de los efectos aleatorios

Estimar los efectos aleatorios δ_i es de frecuente interés ya que, por ejemplo, reflejan la variabilidad entre sujetos lo que permite identificar perfiles especiales o grupos de individuos que evolucionan de manera diferente a través del tiempo y además, si el interés se centra en la predicción de individuos específicos, es necesario determinar las estimaciones de los

efectos aleatorios [Molenberghs and Verbeke, 2005].

Una vez se tienen los estimadores de máxima verosimilitud de $\boldsymbol{\beta}$ y σ^2 los efectos aleatorios δ_i para cualquier individuo particular puede ser estimado mediante la media condicional de δ_i dado $\boldsymbol{y}_i$ y $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}_i^2$ esto es:

$$\begin{aligned}\hat{\delta}_i &= E(\delta_i|\boldsymbol{y}_i, \hat{\boldsymbol{\beta}}, \hat{\sigma}_i^2) \\ &= \frac{1}{h(\boldsymbol{y}_i)} \int \delta_i l(\boldsymbol{y}_i|\delta_i) g(\delta) d\delta\end{aligned}$$

El cuál coincide con el estimador empírico de Bayes. De nuevo, la integral puede ser aproximada usando la cuadratura de Gauss-Hermite.

2.2.2. Modelo ZIP de efectos mixtos

Como antecedentes a este modelo, [Hall, 2000] extendió el modelo Poisson inflado de ceros ZIP propuesto por [Lambert, 1992] (que se presentó en la sección 2.1.2) en el cual incluyó efectos aleatorios en un modelo ZIP para el análisis de datos longitudinales. Sin embargo, los efectos aleatorios solo los incluyó en la componente Poisson de la mixtura; [Dobbie and Welsh, 2001] aplicaron modelos marginales mediante ecuaciones de estimación generalizadas para ambas partes de un modelo Hurdle mientras que [Yau and Lee, 2001] agregaron un par de efectos aleatorios normales no correlacionados para ambas componentes de un modelo Hurdle y [Min and Agresti, 2005] discuten modelos de conteo de efectos aleatorios para medidas repetidas donde además comparan los modelos Hurdle con los modelos ZIP.

Por su parte, [Hur et al., 2002] describen los modelos ZIP de efectos aleatorios incluyendo un solo efecto aleatorio en ambas partes de la mixtura y asumiendo que dichos efectos distribuían normalmente. Finalmente, [Zhu et al., 2015] presenta los modelos ZIP para datos longitudinales con efectos aleatorios heterogéneos mostrando que ignorar la heterogeneidad específica de las covariables puede presentar estimaciones sesgadas.

Modelo

Sea $\mathbf{Y} = (\mathbf{Y}_1', ..., \mathbf{Y}_n')'$ el vector de respuestas de n grupos independientes, donde $\mathbf{Y_i} = (Y_{i1}, ..., Y_{in_i})'$, tales que $Y_{ij}|\nu_i, b_i \sim ZIP(\pi_{ij}, \mu_{ij})$, con $i = 1, ..., n$ individuos $j = 1, ..., n_i$ observaciones repetidas para el individuo i y $N = \sum_{i=1}^{n} n_i$ el total de observaciones; de este modo, Y_{ij} condicionado a los efectos aleatorios, tiene una distribución Poisson inflada de ceros definida como:

$$f(y_{ij}|\nu_i, b_i) = \begin{cases} \pi_{ij} + (1-\pi_{ij})e^{-\mu_{ij}} & \text{si } y_{ij} = 0 \\ (1-\pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} & \text{si } y_{ij} \geq 1 \end{cases} \tag{2.2.10}$$

donde $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in_i})'$ y $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{in_i})$ con funciones de enlace loglineal y logística definidas por:

$$\log(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \nu_i \tag{2.2.11}$$
$$logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\omega}'_{ij}\boldsymbol{\gamma} + b_i$$

Con $\boldsymbol{x}_{ij} = (1, x_{ij1}, ..., x_{ijp})'$ y $\boldsymbol{\omega}_{ij} = (1, \omega_{ij1}, ..., \omega_{ijs})'$ vectores de covariables de tamaño $(p+1)$ y $(s+1)$ respectivamente y, $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$ y $\boldsymbol{\gamma} = (\gamma_0, ..., \gamma_s)'$, vectores de parámetros desconocidos de tamaño $(p+1)$ y $(s+1)$ respectivamente. Se supone que los efectos aleatorios ν_i y b_i están distribuidos normalmente, esto es $\nu_i \sim N(0, \sigma_1^2)$ y $b_i \sim N(0, \sigma_2^2)$ y además son independientes de los vectores de covariables $\boldsymbol{x}_{ij}$ y $\boldsymbol{\omega}_{ij}$ respectivamente.

Para poner en contexto, en el presente estudio se pueden presentar distintos conjuntos de respuesta cero, por ejemplo, en aquellos municipios que por sus características bien sea climatológicas, geográficas, sociales, etc., no son propicias para el desarrollo del dengue y por tanto no se registran casos a través del tiempo, y un grupo que corresponde a aquellos municipios cuyas características son favorables para la posible presencia de casos de dengue pero que no se han registrado casos durante un periodo de tiempo determinado. Por tal motivo, ajustar un modelo Poisson clásico a estos datos subestimaría la probabilidad teórica de cero en el modelo Poisson.

Estimación

A continuación se presenta la estimación de máxima verosimilitud basada en los trabajos de [Hur et al., 2002] y [Hedeker and Gibbons, 2006]. En primer lugar, se define $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)'$ donde $\boldsymbol{\eta}_1 = (\boldsymbol{\beta}, \sigma_1)$ y $\boldsymbol{\eta}_2 = (\boldsymbol{\gamma}, \sigma_2)$. Además, se representa el modelo ZIP dado en (2.2.10) como:

$$f(y_{ij}|\nu_i, b_i) = (1-\pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ij}$$

Donde la función $I(y_{ij})$ es la función indicadora tomando el valor de 1 si la respuesta observada es cero $(y_{ij} = 0)$ y un valor de 0 si la respuesta observada es positiva $(y_{ij} > 0)$.

La función de verosimilitud para el indivuo i es:

$$L(\boldsymbol{y}_i|\nu_i, b_i) = \prod_{j=1}^{n_i}\left[(1-\pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ij}\right] \tag{2.2.12}$$

La función de logverosimilitud está dada por:

$$\log(L(\boldsymbol{y}_i|\nu_i, b_i)) = \sum_{j=1}^{n_i}\log\left[(1-\pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ij}\right] \tag{2.2.13}$$

Y la correspondiente función de densidad marginal de $\boldsymbol{y}_i$ es:

$$h(\boldsymbol{y}_i) = \int L(\boldsymbol{y}_i|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta},$$

donde $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\theta_1 = \nu_i/\sigma_1$ y $\theta_2 = b_i/\sigma_2$ de tal forma que $\theta_1 \sim N(0,1)$ y $\theta_2 \sim N(0,1)$ y $g(\boldsymbol{\theta})$ es la densidad normal estándar.

De manera análoga a la estimación realizada en el modelo de regresión Poisson de efectos mixtos presentada anteriormente y asumiendo que solo hay un efecto aleatorio en cada parte de la mixtura, la integración sobre la distribución de los efectos aleatorios puede ser aproximada usando la cuadratura de Gauss-Hermité de la siguiente manera:

$$\begin{aligned} h(\boldsymbol{y}_i) &\approx \sum_{k=1}^{K} L(\boldsymbol{y}_i|z_k)w_k \\ &\approx \sum_{k=1}^{K}\prod_{j=1}^{n_i}\left[(1-\pi_{ijk})\frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ijk}\right] w_k \end{aligned} \tag{2.2.14}$$

Donde $\pi_{ijk} = \frac{e^{\boldsymbol{\omega}_{ij}'\boldsymbol{\gamma}+\sigma_2 z_k}}{1-e^{\boldsymbol{\omega}_{ij}'\boldsymbol{\gamma}+\sigma_2 z_k}}$ y $\mu_{ijk} = e^{\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\sigma_1 z_k}$.

Así, a partir de la expresión dada en (2.2.14) las funciones de verosimilitud y logverosimitud marginal son respectivamente:

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma_1,\sigma_2) &= \prod_{i=1}^{n} h(\boldsymbol{y}_i) \\ &= \prod_{i=1}^{n}\sum_{k=1}^{K}\prod_{j=1}^{n_i}\left[(1-\pi_{ijk})\frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ijk}\right] w_k \end{aligned} \tag{2.2.15}$$

$$\begin{aligned} \log(L(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma_1,\sigma_2)) &= \sum_{i=1}^{n}\log(h(\boldsymbol{y}_i)) \\ &= \sum_{i=1}^{n}\log\sum_{k=1}^{K}\prod_{j=1}^{n_i}\left[(1-\pi_{ijk})\frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ijk}\right] w_k \end{aligned} \tag{2.2.16}$$

Las primeras derivadas de la logverosimilitud marginal (2.2.16) están dadas por:

$$\frac{\partial \log(L)}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{n}\frac{1}{h(\boldsymbol{y}_i)}\cdot\frac{\partial h(\boldsymbol{y}_i)}{\partial \boldsymbol{\eta}} \tag{2.2.17}$$

Donde

$$\begin{aligned} \frac{\partial h(\boldsymbol{y}_i)}{\partial \boldsymbol{\eta}} &= \sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\eta}}(L(\boldsymbol{y}_i|z_k)w_k) \\ &= \sum_{k=1}^{K}\frac{\partial(\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta}}\cdot L(\boldsymbol{y}_i|z_k)w_k \end{aligned}$$

Las segundas derivadas se presentan a continuación, por comodidad en la notación se tiene $\frac{\partial h(\boldsymbol{y}_i)}{\partial \boldsymbol{\eta}} = d_i$:

$$\begin{aligned}\frac{\partial^2 \log(L)}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} &= \sum_{i=1}^{n} \frac{1}{h(\boldsymbol{y}_i)} \cdot \frac{\partial d_i}{\partial \boldsymbol{\eta}'} + d_i \left(\frac{-1}{h^2(\boldsymbol{y}_i)}\right) \frac{h(\boldsymbol{y}_i)}{\partial \boldsymbol{\eta}'} \\ &= \sum_{i=1}^{n} \frac{1}{h(\boldsymbol{y}_i)} \cdot \frac{\partial d_i}{\partial \boldsymbol{\eta}'} - \frac{1}{h^2(\boldsymbol{y}_i)} d_i \cdot d_i' \\ &= \sum_{i=1}^{n} \frac{1}{h^2(\boldsymbol{y}_i)} \left[h(\boldsymbol{y}_i) \cdot \frac{\partial d_i}{\partial \boldsymbol{\eta}'} - d_i \cdot d_i' \right] \end{aligned} \tag{2.2.18}$$

Donde:

$$\begin{aligned}\frac{\partial d_i}{\partial \boldsymbol{\eta}'} &= \sum_{k=1}^{K} \frac{\partial (\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta}} \cdot \frac{\partial L(\boldsymbol{y}_i|z_k) w_k}{\partial \boldsymbol{\eta}'} + \frac{\partial^2 (\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} L(\boldsymbol{y}_i|z_k) w_k \\ &= \sum_{k=1}^{K} \left[\left(\frac{\partial (\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta}} \right) \left(\frac{\partial \log L(\boldsymbol{y}_i|z_k)}{\partial \boldsymbol{\eta}'} \right)' L(\boldsymbol{y}_i|z_k) w_k + \frac{\partial^2 (\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} L(\boldsymbol{y}_i|z_k) w_k \right] \\ &= \sum_{k=1}^{K} \left[\left(\frac{\partial (\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta}} \right) \left(\frac{\partial \log L(\boldsymbol{y}_i|z_k)}{\partial \boldsymbol{\eta}'} \right)' + \frac{\partial^2 (\log L(\boldsymbol{y}_i|z_k))}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} \right] L(\boldsymbol{y}_i|z_k) w_k \end{aligned}$$

Para $\boldsymbol{\eta}_1 = (\boldsymbol{\beta}, \sigma_1)$ y $\boldsymbol{\eta}_2 = (\boldsymbol{\gamma}, \sigma_2)$, las primeras derivadas de la logverosimilitud con respecto a $\boldsymbol{\eta}_1$ y $\boldsymbol{\eta}_2$ están dadas por:

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{y}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}_1} &= \sum_{j=1}^{n_i} \frac{\partial}{\partial \boldsymbol{\eta}_1} \log \left[(1-\pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} + I(y_{ij}) \pi_{ij} \right] \\ &= \begin{bmatrix} \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[(1-\pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \cdot \frac{\partial}{\partial \beta_0} \left(\log \left(\frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right) \right) \right] \\ \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[(1-\pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \cdot \frac{\partial}{\partial \beta_1} \left(\log \left(\frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right) \right) \right] \\ \vdots \\ \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[(1-\pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \cdot \frac{\partial}{\partial \beta_p} \left(\log \left(\frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right) \right) \right] \\ \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[(1-\pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \cdot \frac{\partial}{\partial \sigma_1} \left(\log \left(\frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right) \right) \right] \end{bmatrix} \end{aligned}$$

y

$$\frac{\partial \log L(\boldsymbol{y}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}_2} = \sum_{j=1}^{n_i} \frac{\partial}{\partial \boldsymbol{\eta}_2} \log \left[(1-\pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} + I(y_{ij})\pi_{ij} \right]$$

$$= \begin{bmatrix} \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[I(y_{ij}) - \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right] \frac{\partial \pi_{ij}}{\partial \gamma_0} \\ \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[I(y_{ij}) - \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right] \frac{\partial \pi_{ij}}{\partial \gamma_1} \\ \vdots \\ \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[I(y_{ij}) - \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right] \frac{\partial \pi_{ij}}{\partial \gamma_s} \\ \sum_{j=1}^{n_i} \frac{1}{f(y_{ij})} \left[I(y_{ij}) - \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right] \frac{\partial \pi_{ij}}{\partial \sigma_2} \end{bmatrix}$$

El cálculo de las segundas derivadas de la logverosimilitud marginal dadas en (2.2.18) es complejo. La solución de la función de logverosimilitud se puede obtener empleando el método iterativo de Newton-Raphson.

3. Métodos de diagnóstico

En este capítulo se abordan los distintos métodos de diagnóstico para la detección de datos atípicos en los Modelos Lineales Generalizados de efectos Mixtos GLMM. Es importante resaltar que si bien la aplicación de estos modelos se ha convertido en una práctica común, las herramientas de diagnóstico para evaluar dichos modelos no han sido ampliamente abordadas a diferencia de los modelos lineales o los modelos lineales generalizados. Además, las especificaciones incorrectas en los modelos GLMM no se pueden diagnosticar de manera confiable con gráficos residuales estándar como los residuales de Pearson, residuales estandarizados o residuales Deviance ya que son más difíciles de interpretar debido a que la distribución esperada de los datos (distribución predictiva) cambia con los valores ajustados [Hartig, 2020]. Esto conlleva a que los modelos GLMM no se estudien a fondo como se debería. No obstante, algunos autores se han enfocado precisamente en abordar estos problemas como [Hartig, 2020] y [Nieuwenhuis et al., 2012].

3.1. Análisis de diagnóstico de influencia

Al ajustar un modelo de regresión, todas las observaciones ejercen algún nivel de influencia sobre la estimación de los parámetros del modelo. Sin embargo, pueden haber observaciones que ejercen una influencia desproporcional en la estimación, bien sea por valores extremos en la variable respuesta, la existencia de diferencias entre los individuos respecto a sus variables explicativas o una mezcla de ambas situaciones, lo que implica que la inclusión o exclusión de esta observación pueda conducir a estimaciones de los coeficientes de regresión sustancialmente diferentes. Estas observaciones se conocen como **puntos influyentes**.

Es importante señalar que el análisis de residuales no se puede emplear para la detección de puntos influyentes [Crawley, 2013]; aquellas observaciones con residuos elevados o residuos estandarizados altos reciben el nombre de ***outliers*** (los cuáles por lo general son observaciones influyentes), sin embargo, un punto influyente no siempre es outlier.
Para la detección de puntos influyentes en modelos de efectos mixtos generalizados, se presentan a continuación dos medidas: DFBETAS y la distancia de Cook, expuestas en [Nieuwenhuis et al., 2012]. La primera es una medida estandarizada de la diferencia absoluta entre la estimación con una observación particular incluida y la estimación sin esa observación particular. La segunda, proporciona una medida general del cambio en todas las estimaciones de los parámetros.

3.1.1. DFBETAS

Teniendo en cuenta que los modelos de regresión de efectos mixtos consideran que las observaciones de los encuestados individuales están anidadas dentro de grupos de nivel superior, como colegios, Estados y países [Snijders and Bosker, 1999] o en el presente caso de estudio dentro de municipios, en primer lugar se debe medir la influencia que ejerce cada grupo de nivel superior, para esto, un camino consiste en eliminar todas las observaciones de los datos que están anidados dentro de un solo grupo de nivel superior (municipio) y a continuación volver a estimar el modelo para evaluar las diferencias existentes en la estimación de los parámetros.

De este modo, DFBETAS en el caso de los modelos mixtos mide la influencia que tiene una unidad de nivel superior en la estimación de un solo parámetro; para calcularlo, se toma la diferencia entre las estimaciones de los parámetros del modelo que incluye el caso del nivel superior y el modelo que lo excluye, dividido entre el error estándar de la estimación del parámetro excluyendo la correspondiente unidad de nivel superior, esto es:

$$DFBETAS_{pi} = \frac{\hat{\gamma}_p - \hat{\gamma}_{p(-i)}}{se(\hat{\gamma}_{p(-i)})}$$

Donde p corresponde al parámetro a estimar e i al grupo de nivel superior, $\hat{\gamma}_p$ representa la estimación original del parámetro p mientras que $\hat{\gamma}_{p(-i)}$ representa la estimación del parámetro p después de excluir el i-ésimo grupo.

Como valor de corte para DFBETAS se considera el valor $2/\sqrt{n}$ con n el número de individuos (grupos de nivel superior). Se considera que los valores que superen dicho corte son los que más influyen en la estimación.

3.1.2. Distancia de Cook

La distancia de Cook es una medida que resume la influencia que ejerce una unidad de nivel superior en todas las estimaciones de los parámetros de forma simultánea; se diferencia del DFBETAS en que este último proporciona un valor por cada parámetro. La distancia de Cook se calcula como:

$$C_i^{0F} = \frac{1}{r+1}(\hat{\gamma} - \hat{\gamma}_{(-i)})'\hat{\Sigma}_F^{-1}(\hat{\gamma} - \hat{\gamma}_{(-i)})$$

Donde $\hat{\gamma}$ representa el vector de estimaciones de parámetros originales, $\hat{\gamma}_{(-i)}$ representa las estimaciones de parámetros del modelo excluyendo la i-ésima unidad de nivel superior y $\hat{\Sigma}_F$ representa la matriz de covarianza excluyendo la i-ésima unidad de nivel superior y r corresponde al número de parámetros que se evalúan.

Para esta medida se considera como valor de corte $4/n$ siendo n nuevamente el número de individuos o grupos de nivel superior; los casos se consideran demasiado influyentes si el valor asociado de la distancia de Cook excede dicho valor.

Una vez se realiza el análisis de influencia respecto de los grupos de nivel superior, de manera análoga se realiza el análisis de influencia dentro de los grupos de nivel inferior, esto es, en nuestro estudio se refiere a las distintas medidas de cada municipio, esto en virtud que pueden haber observaciones influyentes dentro de los grupos que sean precisamente los que hacen que dicho grupo se considere influyente.

3.2. Análisis de diagnóstico de residuales

Como se mencionó en la introducción del presente capítulo, los gráficos de residuales estándar dificultan la identificación de problemas de especificación en los modelos GLMM. Por ejemplo, en el caso específico de los datos de conteo, para corregir la sobredispersión en un modelo con respuesta Poisson, se emplea habitualmente la distribución binomial negativa, sin embargo, una vez se corrige dicha sobredispersión, las violaciones de los supuestos de distribución no son detectables con las pruebas de sobredispersión estándar puesto que estas miran la dispersión total y prácticamente son imperceptibles con los gráficos de residuales estándar [Hartig, 2020].

Con el fin de resolver estos inconvenientes, [Hartig, 2020] propuso unos residuales para modelos lineales generalizados de efectos mixtos que están estandarizados a valores entre 0 y 1 con la característica que se pueden interpretar de manera análoga al modelo lineal. La idea central es emplear un método basado en simulación similar al bootstrap, que consiste en transformar los residuos a una escala estandarizada de la siguiente manera:

- Simular nuevos datos de respuesta del modelo ajustado para cada observación.
- Calcular la función de densidad acumulada empírica para las observaciones simuladas de cada observación.
- El residual se define como el valor de la función de densidad empírica en el valor de los datos observados.

De esta manera, un residual de 0 significa que todos los valores simulados son mayores que el valor observado y un residual de 0.5 significa que la mitad de los valores simulados son mayores que el valor observado. Aquellos casos donde todos los valores simulados sean menores o mayores que los datos observados, obteniendo residuales de 0 y 1 respectivamente se consideran como valores atípicos de simulación (los cuales además corresponden a los valores mínimo y máximo de los residuos).

Un aspecto a tener en cuenta es que debido a la naturaleza de la identificación de los valores atípicos, estos se deben interpretar con cuidado, puesto que no se conoce cuánto se desvían estos valores del modelo ajustado; además, el número de valores atípicos disminuye a medida que se aumenta el número de simulaciones. De este modo, el término *valor atípico* no indica la magnitud de la desviación residual sino que actúa como un signo dicotómico que señala si se encuentra fuera del rango simulado.
Además, para un modelo especificado correctamente se espera que los resiudales escalados sigan una distribución uniforme de forma asintótica. Precisamente la distribución uniforme es la única diferencia que se tiene con los residuales habituales calculados en una regresión lineal.

Paquete estadístico en R

El paquete estadístico a emplear se denomina DHARMA por sus siglas *Diagnostics for HierArchical Regression Models* (Diagnóstico para modelos de regresión jerárquica), el cual, como se mencionó anteriormente emplea un enfoque basado en simulación para crear residuales escalados para modelos lineales y lineales generalizados de efectos mixtos, además, permite procesar simulaciones creadas externamente como las simulaciones predictivas posteriores de *software* bayesinos como STAN o BUGS. Finalmente, el paquete también proporciona una serie de gráficos y pruebas para problemas de estimación tales como la sobredispersión y cero inflación.

Respecto a los gráficos, el paquete DHARMA mediante la función plot.DHARMa() genera dos gráficos:

- QQ plot: el primer gráfico corresponde a un qq plot para detectar desviaciones generales en la distribución esperada, además presenta algunas pruebas de bondad de ajuste sobre los residuales escalados como la prueba ks, prueba de dispersión y prueba de outliers las cuales se abordan más adelante.

- El segundo gráfico presenta los residuales contra el valor predicho. Los valores atípicos de simulación (puntos de datos que están fuera del rango de valores simulados) se resaltan como estrellas rojas.

A continuación se presentan algunas pruebas de bondad de ajuste que realiza el paquete DHARMA:

- testUniformity(): Esta función prueba la uniformidad general de los residuales simulados mediante una prueba ks.

- testDispersion(): Esta función realiza pruebas basadas en simulación para determinar la presencia de sobredispersión. Para ello, compara la varianza de los residuales observados

contra la varianza de los residuales simulados a través de razones. La prueba devuelve la razón de la varianza observada *vs.* la media simulada junto con un p valor basado en la distribución simulada. Una razón significativa > 1 indica sobredispersión mientras que una razón significativa < 1 indica subdispersión.

- testOutliers(): prueba si hay más valores atípicos de simulación de los esperados. La función implementa un procedimiento que usa el bootstrap para generar una simulación basada en la cantidad esperada para los valores atípicos. De igual manera, el método bootstrap prueba el exceso o la ausencia de valores atípicos, por defecto, la función testOutliers mira ambos, así, si se tiene un p valor significativo se debe revisar si se tienen muchos o pocos outliers. El exceso de valores atípicos debe interpretarse como demasiados valores fuera de la envolvente de simulación (posiblemente causado por sobredispersión), mientras que la ausencia de outliers puede ser causado por subdispersión.

Repecto a los modelos inflados de ceros, un aspecto a tener en cuenta es que la sobredispersión puede dar lugar a un exceso de ceros por lo que observar demasiados ceros no es una señal confiable de emplear precisamente un modelo inflado de ceros. Una diferencia confiable entre la sobredispersión y la inflación cero generalmente solo será posible cuando se comparen directamente modelos alternativos [Hartig, 2020] a través por ejemplo de la comparación de residuales o ajustando un modelo con inflación cero y mirando la estimación del parámetro para la inflación cero.

4. Análisis por clústeres de casos de dengue en Colombia

En esta sección se presenta el proceso de obtención de los datos así como la construcción de las bases, seguido de la metodología kml (k-means para datos longitudinales) mediante la cual se establecen los clústeres o agrupamientos de municipios con características similares y, finalmente, el análisis del comportamiento de los casos de dengue en los municipios de cada clúster.

Los datos fueron obtenidos a través del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) en donde se registra semanalmente el número de casos de distintas enfermedades por municipios (los registros corresponden a los casos notificados por los entes territoriales como "confirmados", ver http://portalsivigila.ins.gov.co/sivigila/index.php).

Respecto a las variables explicativas, inicialmente se realizó una estimación de la densidad poblacional tomando como punto de partida las proyecciones poblacionales ajustadas por cobertura desde el año 2007 hasta el año 2018, datos tomados del Departamento Administrativo Nacional de Estadística DANE desde https://www.dane.gov.co. Además, se incluyeron las variables climatológicas clima promedio y precipitación promedio (mm), estos datos se obtuvieron gracias al instituto de Hidrología, Meteorología y Estudios Ambientales IDEAM (http://www.ideam.gov.co/) quienes proporcionaron ambas mediciones mensuales desde enero de 2007 hasta diciembre de 2018 para aquellos municipios que cuentan con estaciones metereológicas.

Finalmente, se considera el Índice de Riesgo de la Calidad del Agua para el consumo humano (IRCA), definida en la Resolución 2115 de 2007 como *un indicador que determina la calidad del agua, por el grado de riesgo de ocurrencia de enfermedades relacionadas con el no cumplimiento de las características físicas, químicas y microbiológicas del agua para consumo humano, basado en análisis de características físicas, químicas y microbiológicas en muestras de agua*. La metodología para la obtención del indicador se puede consultar en la misma resolución.

Se seleccionaron únicamente los índices urbanos específicados de la siguiente manera: para valores entre 0 % y 5 % el agua distribuida es apta para consumo humano (sin riesgo), valores

entre 5.1 % y 14 % ya no es apta para consumo humano (nivel de riesgo bajo); valores entre 14.1 % y 35 % no es apta para consumo humano (nivel de riesgo medio); para valores del IRCA entre 35.1 % y 80 % el nivel de riesgo es alto y entre 80.1 % y 100 % el agua distribuida es inviable sanitariamente; esta información se obtuvo a partir de la página del Instituto Nacional de Salud desde http://www.ins.gov.co/sivicap/Paginas/sivicap.aspx.

Construcción de las bases de datos

La variable de interés corresponde a la cantidad de casos de dengue. Para la construcción de esta base, se tomó la información anual obtenida del portal SIVIGILA y se realizó la sumatoria anual de los casos reportados de dengue, a partir del año 2007 hasta el año 2018 para cada uno de los 1,122 municipios del país. Obteniendo así una base de tamaño 1,122 por 12 correspondiente a los 12 años que abarcan el periodo de interés.

La primera variable explicativa considerada es la densidad poblacional, como se mencionó anteriormente, se realizó una estimación de la densidad poblacional a partir de la proyección realizada por el DANE de la población para cada municipio y el área en km^2.

Otras variables explicativas de interés son las variables climatológicas precipitación y clima. Cabe señalar que la información proporcionada por el IDEAM son los registros de cada estación metereológica en el periodo de interés, con las siguientes características: por un lado, no traía la información de todos los municipios del país (puesto que no se tiene estaciones en todos los municipios), en contraste, hay municipios con más de una estación metereológica (principalmente en grandes ciudades) además, algunas estaciones traían registros diarios mientras que otras traían registros mensuales (temperatura promedio). Bajo estas características, se seleccionó una única estación por municipio y se calculó la precipitación promedio anual y la temperatura promedio anual desde el año 2007 hasta el año 2018. También es importante señalar que si bien hay registros de la precipitación promedio de 845 municipios y de la temperatura promedio de 354 municipios, en algunos casos solo se tiene información desde el 2007 hasta el 2010. En total se cuenta con 266 municipios que tienen la información climatológica completa.

La variable climatológica *humedad* también se solicitó al IDEAM pero la información venía para pocos municipios y tenía datos faltantes, razón por la cual no se consideró en el análisis.

Finalmente, se tuvo en cuenta el IRCA como otra variable explicativa. En principio, no fue posible obtener una base de datos con esta información, por lo cual, se tomaron los informes anuales presentados por el Ministerio de Salud en los que se presenta el IRCA para cada municipio por departamento. A partir de estos archivos, se construyó la base de datos con el IRCA anual para cada municipio.

4.1. k-means para datos longitudinales

En general k-means es un algoritmo que consiste en particionar un conjunto de n individuos en k grupos de la siguiente manera: en primer lugar se escogen los centroides de los grupos que minimicen las distancias de cada individuo a ellos, luego, cada individuo es asignado al grupo cuyo centroide esté más cercano a dicho individuo [Días and Morales, 2012]. El objetivo es reubicar los individuos de tal forma que se consigan grupos con la menor variabilidad posible.

Respecto a datos longitudinales, en [Genolini et al., 2015] se presenta la adaptación e implementación del método k-means para datos longitudinales, bien sea para una variable *trayectoria* (la misma variable medida repetidamente a través del tiempo) como para trayectorias conjuntas (evolución conjunta de varias variables trayectorias); además se presentan los paquetes kml y kml3d de R los cuales realizan de forma automática el algoritmo k-means. A continuación se presenta el desarrollo más detallado de este algoritmo.

4.1.1. Definición de distancia entre trayectorias conjuntas

Sea el conjunto S de n elementos, donde para cada elemento se miden p variables en t tiempos diferentes. Se define una trayectoria individual como $y_{i.A} = (y_{i1A}, y_{i2A}, ..., y_{itA})$ que representa cada uno de los valores de la variable A para el individuo i en los t tiempos. Además, se define la trayectoria conjunta como:

$$y_{i..} = \begin{pmatrix} y_{i.A} \\ y_{i.B} \\ \vdots \\ y_{i.P} \end{pmatrix} = \begin{pmatrix} y_{i1A} & y_{i2A} & \cdots & y_{itA} \\ y_{i.B} & y_{i2B} & \cdots & y_{itB} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i.P} & y_{i2P} & \cdots & y_{itP} \end{pmatrix}$$

El cual, representa las trayectorias individuales para las p variables.

Plantear una distancia para trayectorias conjuntas, implica trabajar con distancias entre matrices. Para esto, un método se basa en consideran las t columnas de las dos matrices, luego, calcular las t distancias entre las t parejas respectivas. Esto es, calcular la distancia entre $y_{1..}$ y $y_{2..}$, para cada j fijo se define la distancia entre $y_{1j.}$ y $y_{2j.}$ como $d_{j.}(y_{1j.}, y_{2j.}) = Dist(y_{2j.}, y_{2j.})$ la cual es la distancia entre la columna j en la matriz $y_{1..}$ y la columna j de la matriz $y_{2..}$ Se obtiene así, un vector de t distancias:

$$(d_{1.}(y_{11.}, y_{21.}), d_{2.}(y_{12.}, y_{22.}), ..., d_{t.}(y_{1t.}, y_{2t.}))$$

Luego, se combinan las t distancias mediante una norma, obteniendo:

$$d(y_{1..}, y_{2..}) = \|(d_{1.}(y_{11.}, y_{21.}), d_{2.}(y_{12.}, y_{22.}), ..., d_{t.}(y_{1t.}, y_{2t.}))\|$$

De manera análoga, otro método para calcular la distancia entre dos matrices se realiza tomando las filas correspondientes. Al tomar la norma como la p-norma estándar y la distancia *Dist* como la distancia de Minkowski, los dos métodos dan el mismo resultado. Así, se obtiene la siguiente expresión.

$$d(y_{1..}, y_{2..}) = \sqrt[p]{\sum_j \sum_X |y_{1jX} - y_{2jX}|^p}$$

Haciendo $p = 2$ se tiene la distancia euclidiana. que es la distancia por defecto de los paquetes kml y kml3d.

4.1.2. Selección de la cantidad de grupos

Se considera que una partición es buena cuando los grupos son compactos y están bien separados unos de otros. Los paquetes kml y kml3d traen incorporados distintos criterios los cuales toman valores altos para particiones de "alta calidad". El criterio seleccionado en este caso es el criterio de Calinski y Harabasz descrito como sigue:

$$C(k) = \frac{Traza(B)}{Traza(W)} \cdot \frac{n-k}{k-1}$$

En donde B es la matriz de covarianza entre grupos por lo cual para valores altos de $TrazaB$ se tiene grupos bien separados mientras que W es la matriz de covarianza dentro de grupos la cual para valores bajos corresponde a grupos compactos.

4.1.3. Configuración inicial del método k-means

El primer paso del algoritmo se basa en seleccionar una configuración inicial que forme un conjunto de k grupos. La escogencia de esta configuración es clave puesto que si es cercana a la mejor partición, el método convergerá más rápido. Los paquetes kml y kml3d ofrecen siete alternativas para elegir la configuración inicial. La configuración incial seleccionada es la siguiente.

a. Elegir un centro c_0 al azar entre los puntos de datos.

b. Para cada punto de datos x calcular la distancia $D(x)$ entre x y c_0.

c. Elegir un nuevo cntro c_1 al azar mediante una distribución de probabilidad ponderada proporcional a $D(x)^2$.

d. Eliminar el centro inicial c_0 de la lista de centros.

e. Comenzar un ciclo:

i. Para cada punto de datos x, calcular $D(x)$, la distancia entre x y el centro más cercano que ya ha sido elegido.

ii. Elegir aleatoriamente un punto de datos como el nuevo centro c_i, utilizando una distribución de probabilidad ponderada donde se elija un punto x con probabilidad proporcional a $D(x)^2$.

iii. Repetir los pasos i y ii hasta que se hayan elegido k centros.

4.1.4. Aplicación del método kml

En nuestro estudio, como se tiene una variable trayectoria que corresponde a la cantidad de casos anuales reportados de dengue desde el año 2007 hasta el año 2018, se ejecutó el algoritmo mediante el paquete kml. Para esto, se tomaron las frecuencias acumuladas, con el fin de obtener una mejor visualización de los clústeres.
Respecto a la cantidad de grupos, en la figura 4-1 se ilustran los resultados al aplicar el criterio de Calinski y Harabasz.

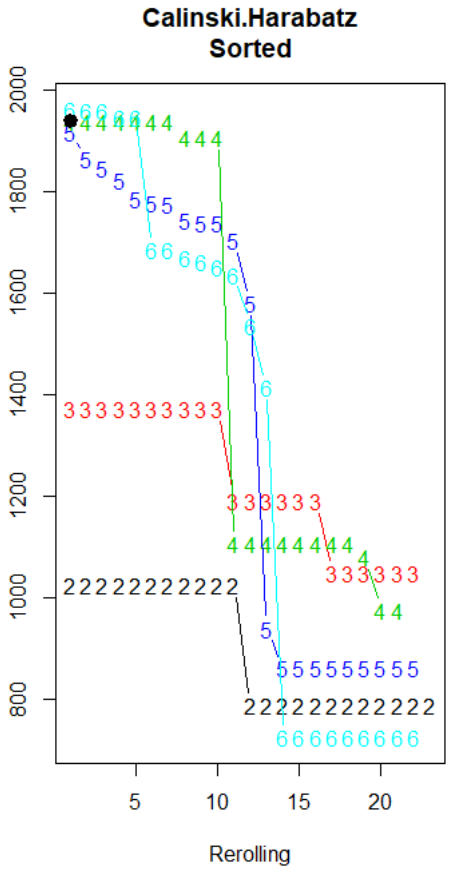

Figura 4-1.: Criterio de Calinski y Harabasz para determinar el número de grupos óptimo del algoritmo.

Se puede observar todas las particiones encontradas por kml; cada partición se representa por un número que indica la cantidad de grupos que tiene. La altura representa el valor del criterio de Calinsky-Harabatz. De todas las particiones que se encontraron, se escoge la que tenga el mayor valor, que en este caso corresponde a seis grupos.

Una vez se establece la cantidad óptima de grupos, se trazan las trayectorias acumuladas individuales de todos los municipios y se resalta la estructura de grupo de la partición seleccionada mediante colores. Los resultados se muestran en la figura **4-2**.

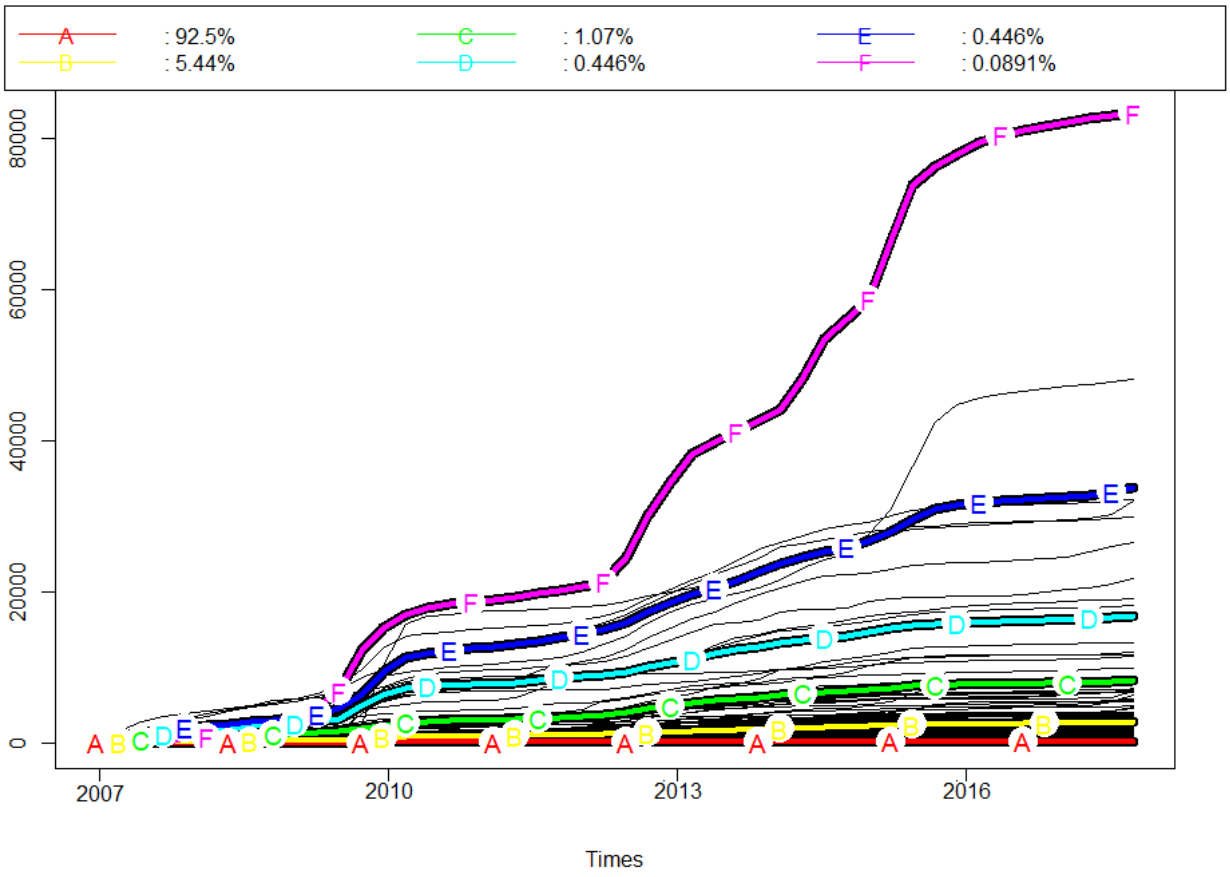

Figura 4-2.: Estructura de los 6 clústeres seleccionados.

Según los resultados obtenidos, en primer lugar, se puede observar que en el año 2010 se presentó un punto de inflexión en las estructuras de los grupos D, E y F, esto es, se presentó un aumento considerable en la cantidad de casos reportados; por otra parte, entre los años 2012 y 2013 se presenta un nuevo punto de inflexión para los grupos B, C, D, E y F en donde nuevamente se presenta un aumento considerable en la cantidad de casos reportados. Además, en el año 2016 la cantidad de casos se estabiliza para los grupos B, C, D y E es decir, el aumento en los casos no es tan marcado como en los otros años.

Por otra parte, se puede notar que en los grupos E y F se encuentran los municipios con la mayor cantidad de casos de dengue reportados en el país a lo largo de los 12 años. La estructura del grupo F, la presenta únicamente la ciudad de Cali, la estructura del grupo E la presentan las siguientes ciudades: Medellín, Bucaramanga, Cúcuta, Ibagué y Villavicencio; la estructura del grupo D la presentan Barranquilla, Neiva, Floridablanca, Armenia y Sincelejo mientras que la estructura del grupo B la tienen ciudades como Pereira, Valledupar y Yopal, entre otras. Finalmente la estructura del grupo A la presentan aquellos municipios que tienen la menor cantidad de casos acumulados, por lo cual se ve que es estable a través del tiempo y, además, es el clúster con la mayor cantidad de municipios.

Para complemetar este análisis, en la figura **4-3**, se presenta la distribución espacial de la cantidad de casos de dengue acumulados por municipio.

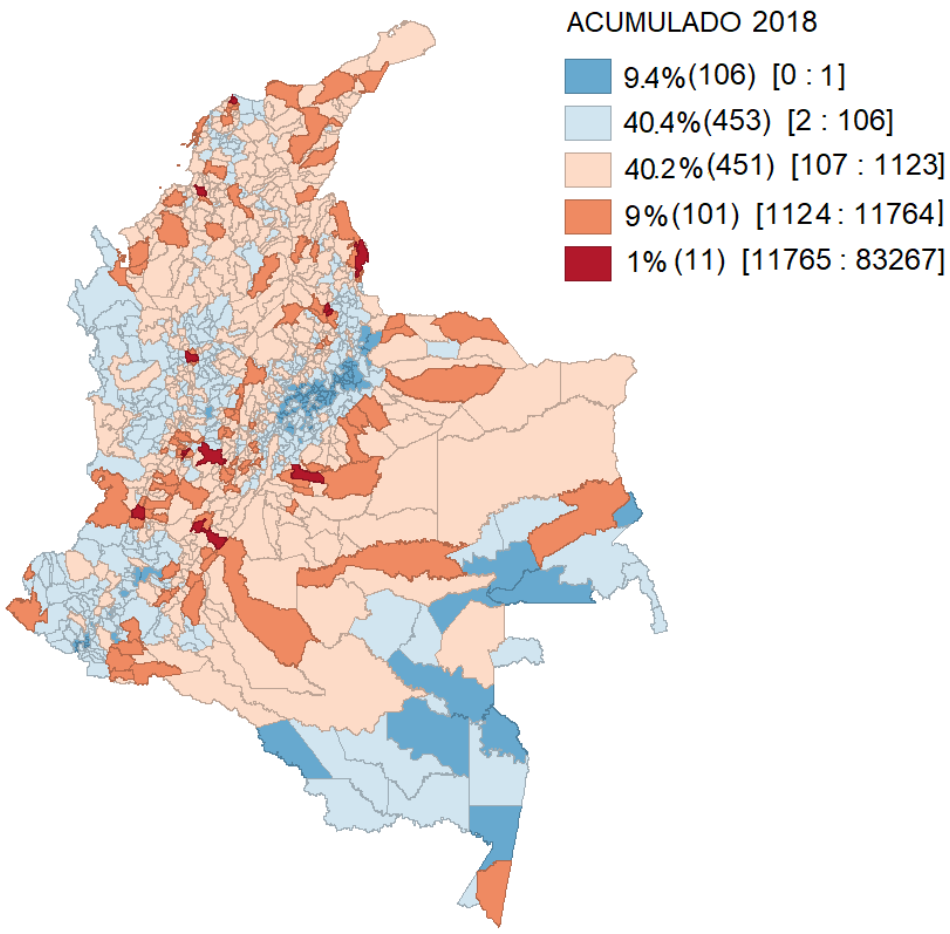

Figura 4-3.: Cantidad de casos acumulados de dengue registrados entre los años 2007 y 2018.

Como se puede observar en la figura **4-3**, los municipios que presentan menos casos de dengue, indicados con azul, se encuentran principalmente en la región pacífica y en el sur del país, los municipios que tienen un total acumulado entre 1,123 y 11,764 casos están indicados en color naranja (en total son 101 municipios que corresponden al 9 % del total), mientras que, los municipios indicados con rojo son aquellos que presentan un total acumulado superior a 11,764 casos (en total son 11 municipios que representan el 1 % del total). Si bien la distribución espacial de los municipios no coincide necesariamente con los clústeres obtenidos con la metodología kml, permite visualizar el comportamiento de la cantidad de casos acumulados de dengue en el país.

Por otra parte, en la Tabla **4-1** se presenta la media y la desviación estándar de los casos acumulados de dengue para los años 2007, 2011, 2015 y 2018 para cada clúster, a excepción del clúster F ya que corresponde únicamente a la ciudad de Cali.

Año	Clúster A		Clúster B		Clúster C		Clúster D		Clúster E	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
2007	7.3	19.6	108.3	123.6	458.8	227.9	1210	757.8	1675	1323.9
2011	65.58	116.3	915.2	647.4	3593	1836.5	6801	2563.4	13011	3663.1
2015	168.4	246.4	2177	929.9	7033	1829.9	14787	3437.8	26724	2557.9
2018	214.9	303.1	2736	1175	8235	1815.8	16839	4076.1	33743	8293.7

Tabla **4-1**.: Media y desviación estándar de la cantidad de casos acumulados de dengue por clúster.

Cabe resaltar que, mientras la cantidad promedio de casos acumulados de dengue hasta el año 2018 en el clúster A es de 215 aproximadamente, para el clúster E esta cantidad promedio es de 33743, lo que evidencia una diferencia importante en el comportamiento de la variable de interés.

4.2. Comportamiento de los casos de dengue por clúster

A continuación se presenta un análisis por clúster de la cantidad de casos de dengue en el periodo de estudio así como el comportamiento de la variable de interés a través del tiempo, considerando distintos intervalos de las variables explicativas en cada uno de los clústeres.

4.2.1. Clúster E

El clúster E lo conforman las ciudades de Medellín, Bucaramanga, Cúcuta, Ibagué y Villavicencio. En la figura 4-4 se presenta el comportamiento a través del tiempo de los casos de dengue reportados en estas ciudades.

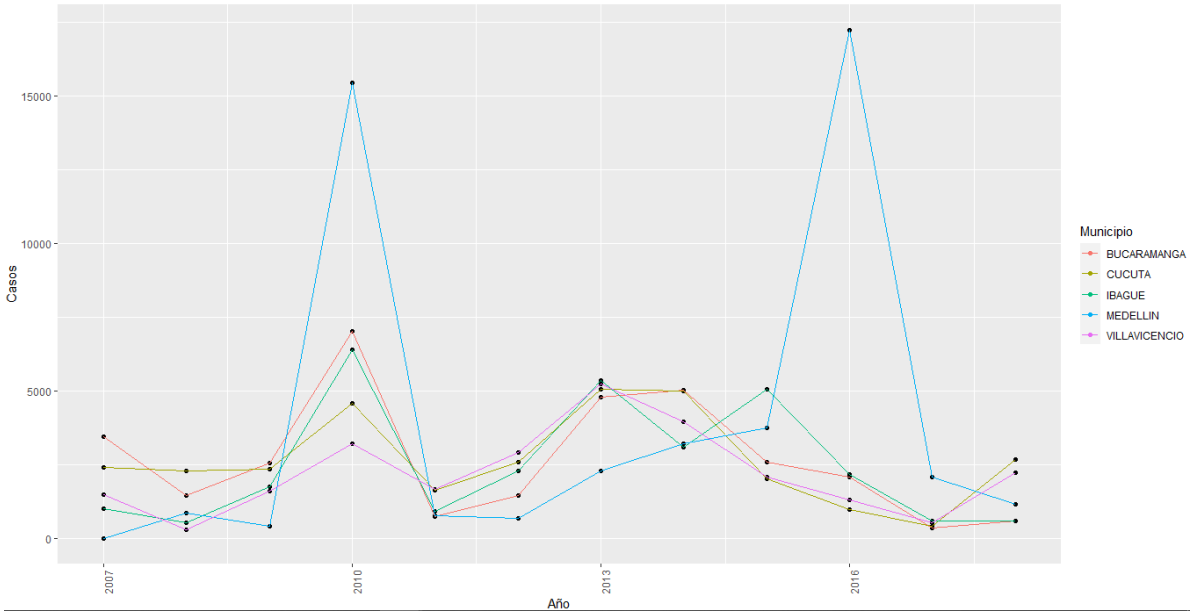

Figura 4-4.: Comportamiento de la cantidad de casos de dengue, clúster E.

Se puede observar que en el año 2010 se presentó un aumento considerable en los casos de dengue en los 5 municipios, resaltando principalmente la ciudad de Medellín ya que tanto en el año 2010 como en el año 2016 superó los 15,000 casos anuales (valores extremos en comparación con los otros años y los demás municipios).

En la Tabla 4-2 se presenta el rango que toman las variables explicativas para cada municipio del clúster E. Con relación al IRCA, teniendo en cuenta la clasificación del nivel de riesgo de calidad del agua presentada anteriormente, Medellín es la única ciudad en la cual todas las mediciones están por debajo del 5 % (sin riesgo); en Bucaramanga se presentan mediciones con un nivel de riesgo bajo (entre 5.1 % y 14.1 %), Cúcuta y Villavicencio presentan mediciones con un nivel de riesgo medio (entre 14.1 % y 35 %) mientras que Ibagué presenta mediciones con un nivel de riesgo alto (entre 35.1 % y 80 %).

Respecto a la densidad poblacional, Medellín y Bucaramanga presentan una mayor densidad en comparación con ciudades como Cúcuta, Ibagué y Villavicencio. Finalmente, respecto a las variables climatológicas, si bien son ciudades con comportamientos climatológicos similares, en Villavicencio se registró la precipitación promedio más alta (419.1), además, la temperatura promedio de Cúcuta es mayor frente a las otras ciudades mientras que la temperatura promedio de Medellín es menor a la de las demás ciudades del clúster.

Municipio	IRCA (%)		Densidad (hab/km^2)		Precipitación (mm)		Temperatura (°C)	
	Mín.	Máx.	Mín.	Máx.	Mín.	Máx.	Mín.	Máx.
Medellín	0.39	2.4	6016.12	6410.71	142.42	312.55	21.8	22.9
Bucaramanga	0	14	3360.65	3587.22	70.05	144.65	22.7	23.3
Cúcuta	0.5	34.42	531.52	605.2	0	129.6	26.8	27.8
Ibagué	8.1	59.5	349.17	368.06	90.7	198.4	24.6	26
Villavicencio	0.34	33.4	317.18	404.63	315.07	419.1	24.6	26.4

Tabla **4-2**.: Rango de las variables explicativas para cada municipio del clúster E.

A continuación se presentan tendencias globales de la cantidad de casos de dengue en el transcurso del tiempo con respecto a las variables explicativas. Para ello, se emplea la función *xyplot* la cual muestra un diagrama de dispersión entre dos variables numéricas separado por paneles para cada uno de los subgrupos de observaciones, según intervalos determinados de la variable explicativa de interés. De este modo, en la figura **4-5** se presenta el diagrama de dispersión con relación a la variable IRCA.

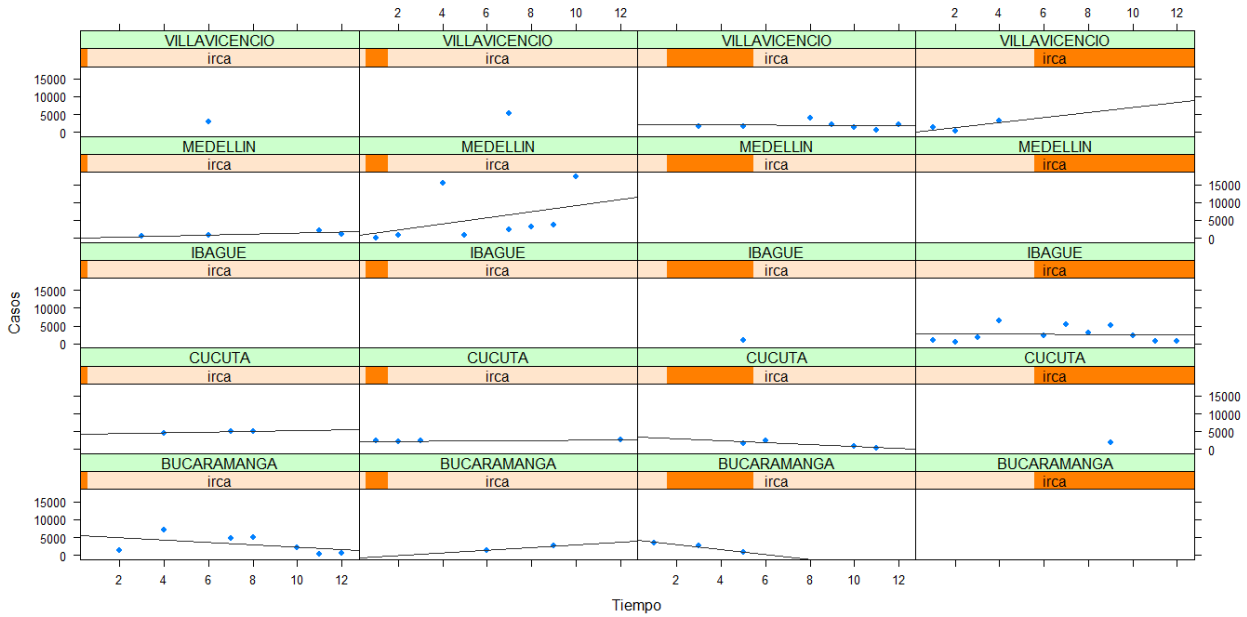

Figura 4-5.: Diagrama de dispersión según rango de valores de la variable IRCA, clúster E.

En este caso, el diagrama de dispersión presenta la cantidad de casos de dengue a través del tiempo separado en cuatro grupos (paneles) según los siguientes intervalos de la variable IRCA: [0 %, 1.42 %), [1.42 %, 6.29 %), [6.29 %, 25.29 %) y (25.29 %, 59.51 %), de tal forma que cada grupo quede con la misma cantidad de observaciones. Así, por ejemplo, en el primer panel se encuentran aquellas observaciones en las que la variable IRCA fue inferior al 1.42 % en cada municipio.

Algunas características generales que se pueden identificar es que en Medellín se presenta una tendencia creciente a pesar de ser una ciudad que tiene todas las mediciones del IRCA por debajo de 5 %, esto es, el agua potable no representa algún riesgo para el consumo humano; caso contrario a Bucaramanga en el cual se evidencia una tendencia decreciente en aquellas observaciones con mediciones de IRCA por debajo de 1.42 %. En contraste, en Ibagué a pesar de que la mayoría de las observaciones se dieron con valores de IRCA mayores a 25.9 % (niveles de riesgo medio y alto), no se presenta una tendencia marcada.

Respecto a la densidad poblacional, en la figura **4-6** se presentan la cantidad de casos de dengue respecto al tiempo, separado nuevamente en 4 grupos (paneles) según el intervalo que asume esta variable explicativa. En el primer panel se presentan las observaciones en las cuales la densidad poblacional toma valores entre 317.1 hab/km^2 y 361.8 hab/km^2, para el segundo panel toma valores entre 361.8 hab/km^2 y 555.51 hab/km^2, en el tercer panel toma valores entre 557.94 hab/km^2 y 3,534.51 hab/km^2 y en el último panel la densidad poblacional toma valores entre 3,534.51 hab/km^2 y 6,410.79 hab/km^2.

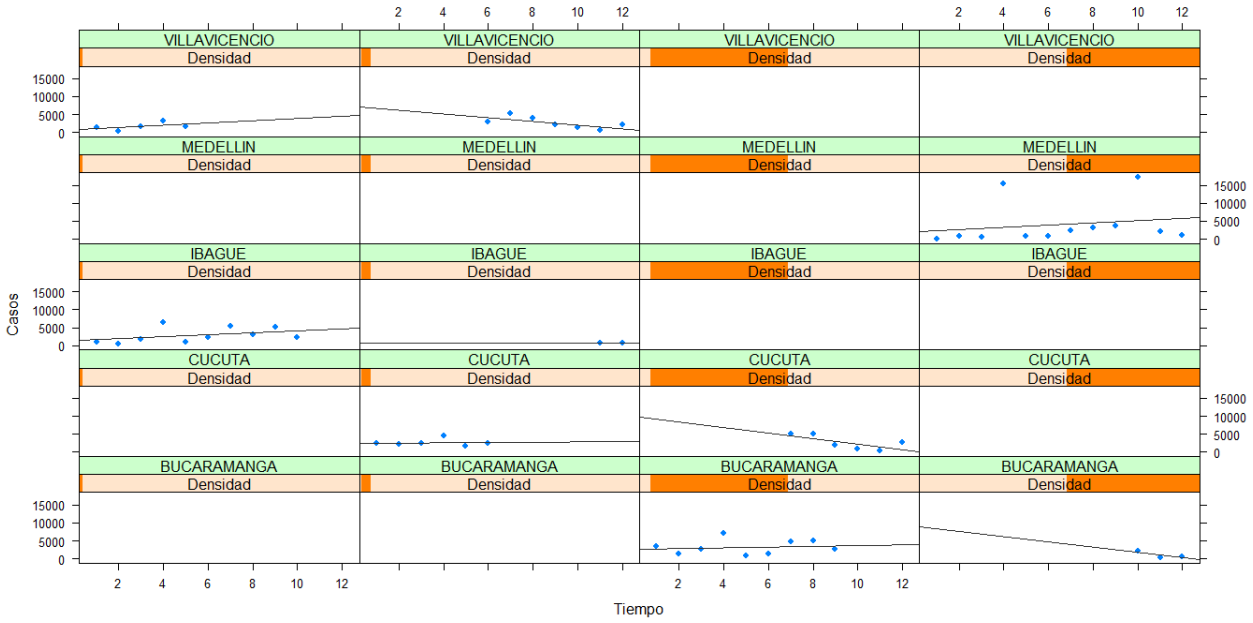

Figura 4-6.: Diagrama de dispersión según rango de valores de la variable Densidad poblacional, clúster E.

En Villavicencio la tendencia es decreciente cuando la densidad poblacional fue mayor a 361.8 hab/km^2, situación similar a Cúcuta en la cual la tendencia nuevamente es decreciente cuando la densidad poblacional es mayor a 3,534.51 hab/km^2. En cuanto a Ibagué y Medellín, a pesar de tener una densidad poblacional notablemente distinta, la tendencia en la cantidad de casos de dengue respecto al tiempo es creciente.

En la figura 4-7 se presenta el diagrama de dispersión separado en paneles según los valores que asume la precipitación promedio. Los intervalos de referencia son: [0 mm, 96.61 mm), [96.61 mm, 138.06 mm), [138.06 mm, 254.83 mm) y [254.83 mm, 419.11 mm).

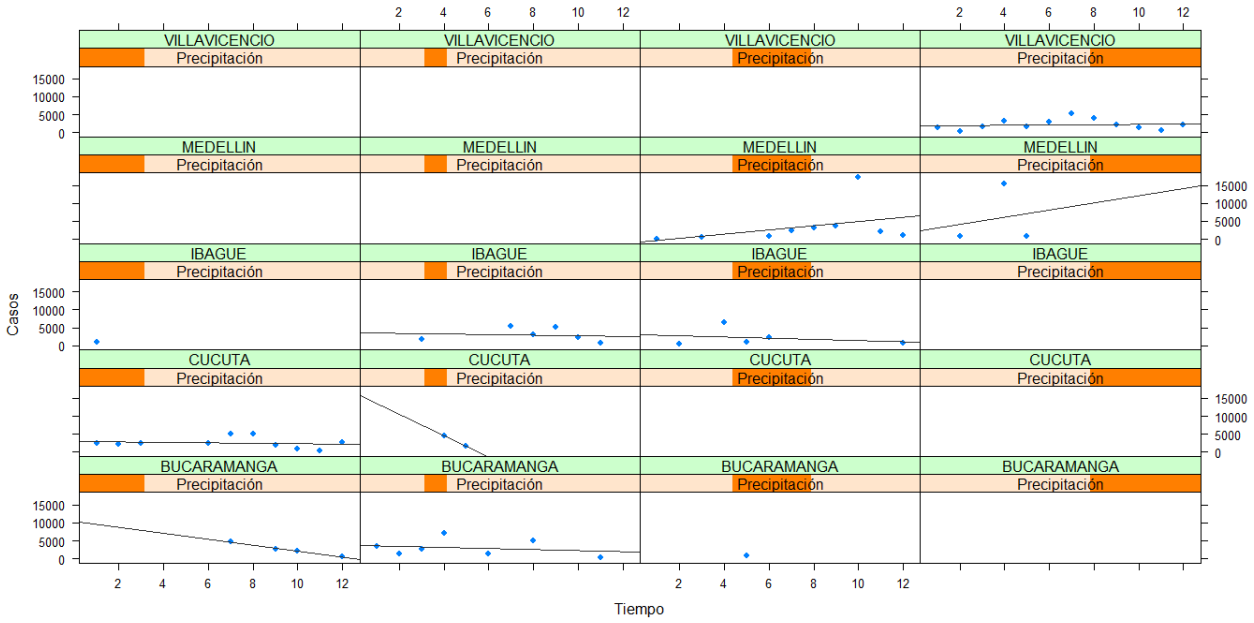

Figura 4-7.: Diagrama de dispersión según rango de valores de la variable Precipitación, clúster E.

Si bien el municipio con la mayor precipitación promedio anual en los 12 años considerados es Villavicencio, a lo largo del tiempo no se evidencia una tendencia creciente marcada, caso contrario a Medellín la cual, siendo la segunda ciudad con mayor precipitación anual promedio, como se ha mencionado tiene una tendencia creciente. Además, en Ibagué en los años en que la precipitación fue mayor a 138.06 mm, la tendencia es decreciente mientras que en Bucaramanga, los años en que la precipitación fue inferior a 96.61 mm la tendencia en la cantidad de casos es decreciente.

Finalmente, respecto a la temperatura, en la figura 4-8 se presenta el comportamiento de la variable de interés en el tiempo según la temperatura promedio. En el primer panel se presentan las observaciones con una temperatura promedio entre 21.8 °C y 23.01 °C, en el segundo panel entre 23.01 °C y 25.09 °C, en el tercer panel entre 25.09 °C y 25.26 °C y 27.46 °C.

En general, no se presentan tendencias en la cantidad de casos de dengue en el tiempo, al dividirlo en grupos según la temperatura promedio, a excepción de Medellín, para el cual todas las observaciones caen en el primer grupo, y tal como se ha mencionado previamente, presenta una tendencia creciente.

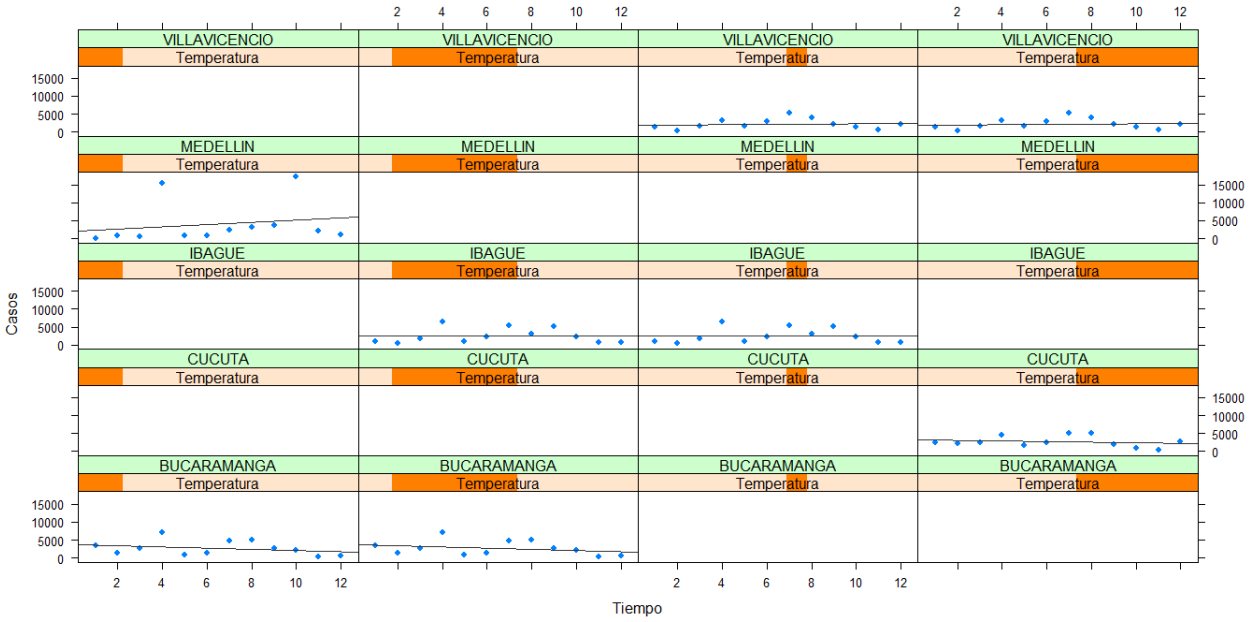

Figura 4-8.: Diagrama de dispersión según rango de valores de la variable Temperatura, clúster E.

4.2.2. Clúster D

El clúster D lo conforman las ciudades de Barranquilla, Neiva, Floridablanca, Armenia y Sincelejo. En la figura **4-9** se presenta el comportamiento a través del tiempo de los casos de dengue reportados en estas ciudades.

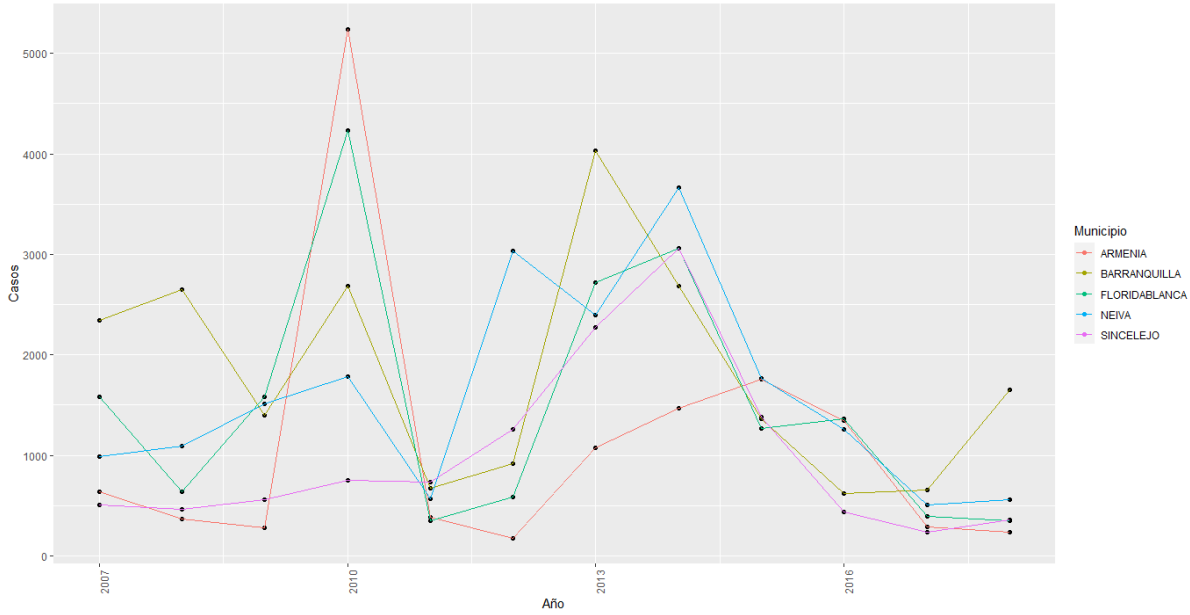

Figura 4-9.: Comportamiento de la cantidad de casos de dengue, clúster D.

En el año 2010 se presentó una cantidad de casos de dengue elevada en comparación con los demás años en los municipios de Armenia y Floridablanca. De igual manera, entre los años 2013 y 2014 la cantidad de casos fue alta y empezó a disminuir a partir del año 2015.

En la Tabla **4-3** se muestra el rango que tienen las variables explicativas para cada municipio.

Municipio	IRCA (%)		Densidad (hab/km^2)		Precipitación (mm)		Temperatura (°C)	
	Mín.	Máx.	Mín.	Máx.	Mín.	Máx.	Mín.	Máx.
Barranquilla	0	15.89	7024.82	7416.99	40.2	168.22	27.7	29
Neiva	2.5	52.22	210.85	230.13	47.75	181.32	26.3	30.7
Floridablanca	0	40.64	2773.77	3009.64	106.35	200.6	-	-
Armenia	0	2.12	1157.06	1185.47	133.87	345.025	-	-
Sincelejo	0.7	6.6	878.11	978.07	49.75	274.6	-	-

Tabla **4-3**.: Rango de las variables explicativas para cada municipio del clúster D.

Un inconveniente de este clúster es que infortunadamente no se cuenta con la temperatura promedio para todas las ciudades ya que Floridablanca y Sincelejo carecen de estaciones que

midan esta variable y, en el caso de Armenia, los datos no corresponden a los años de estudio.

Respecto al IRCA, Armenia es el municipio con la mejor calidad de agua dado que todas las mediciones tuvieron un IRCA inferior a 5 %, seguido de Barranquilla y Sincelejo que presentan mediciones con un nivel de riesgo bajo (entre 5.1 % y 14.1 %); mientras que los municipios de Neiva y Floridablanca presentan mediciones con un nivel de riesgo alto (entre 35.1 % y 80 %).

Neiva es el municipio con menor densidad poblacional, seguido de Sincelejo, mientras que Barranquilla es la ciudad con mayor densidad poblacional del clúster. Finalmente, respecto a la precipitación promedio, Barranquilla registra el valor promedio más bajo; por su parte, Armenia registra la precipitación promedio más alta del clúster (345.025 mm).

En la figura **4-10**, se presenta el diagrama de dispersión que relaciona la cantidad de casos respecto el tiempo, separado en cuatro grupos (paneles) según distintos rangos de la variable IRCA. El primer grupo toma valores del IRCA entre 0 % y 0.19 %, el segundo grupo toma valores de esta variable entre 0.19 % y 1.19 %, el tercer grupo entre 1.19 % y 12.9 % y finalmente el último grupo con valores de IRCA entre 12.9 % y 52.3 %. De tal manera que cada grupo cuenta con la misma cantidad de observaciones.

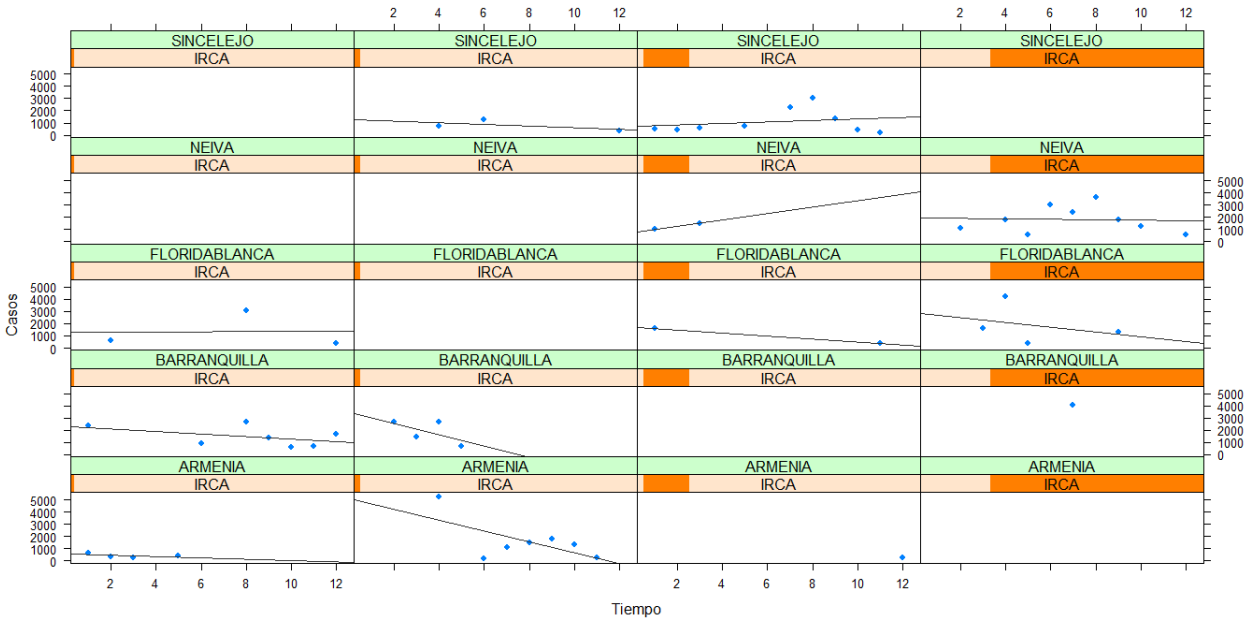

Figura 4-10.: Diagrama de dispersión según rango de valores de la variable IRCA, clúster D.

Respecto al primer grupo, las ciudades que tienen valores en este rango son Floridablanca, Barranquilla y Armenia, en los cuales se observa una tendencia decreciente en la cantidad de

casos respecto al tiempo, principalmente en Barranquilla; para el segundo grupo, se observa una tendencia decreciente en Sincelejo, Barranquilla y Armenia, en estos dos últimos una tendencia más fuerte. Para el tercer grupo la tendencia es creciente en Sincelejo y en Neiva mientras que en el último grupo, la tendencia es decreciente en Floridablanca a diferencia de Neiva, en donde la mayoría de las mediciones del IRCA son mayores a 9.4 % sin embargo no presenta una tendencia marcada.

Con relación a la densidad poblacional, en la figura **4-11**, se presenta la cantidad de casos *vs.* el tiempo separado en cuatro grupos según el intervalo que toma la densidad poblacional; el primer grupo corresponde a aquellas observaciones con una densidad inferior a 891.14 hab/km^2, se puede ver que todas las mediciones de Neiva se encuentran en este intervalo y algunas de Sincelejo, la tendencia es decreciente pero no tan marcada en Neiva. El segundo grupo está conformado por las observaciones que presentan una densidad entre 891.14 hab/km^2 y 1177.5 hab/km^2, los dos municipios que presentan densidad poblacional en este rango son Sincelejo y Armenia, el primero presenta una tendencia decreciente en el tiempo mientras el segundo una tendencia creciente. En el tercer grupo la densidad toma valores entre 1177.5 hab/km^2 y 2936.5 hab/km^2 en Floridablanca la tendencia es creciente mientras que en Armenia es decreciente y finalmente, el último grupo está conformado por aquellas mediciones de la densidad superiores a 2936.5 hab/km^2, tanto en Floridablanca como en Barranquilla la tendencia es decreciente, este último con todas las observaciones en este rango.

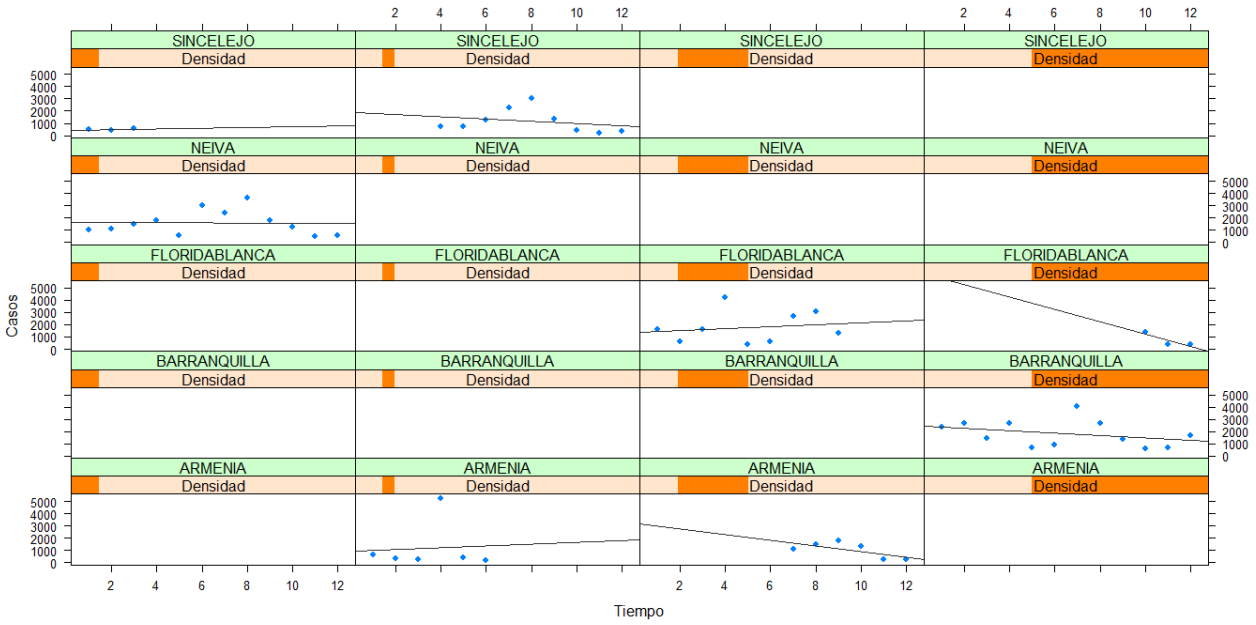

Figura 4-11.: Diagrama de dispersión según rango de valores de la variable Densidad poblacional, clúster D.

En cuanto a la variable precipitación, esta se dividió en los siguientes intervalos: [0 mm, 64.73 mm), [64.73 mm, 131.93 mm), [131.93 mm, 196.2 mm) y [196.2 mm, 591.4 mm). En Sincelejo, para los años con una precipitación promedio entre 64.73 mm y 131.93 mm la tendencia en la cantidad de casos de dengue es decreciente mientras que para años más lluviosos, con una precipitación promedio superior a 131.93 mm esta tendencia es creciente. Una situación similar se presenta en Neiva en donde los años con una precipitación promedio inferior a 64.73 mm generan una tendencia decreciente y en los años con una precipitación promedio entre 64.73 mm y 131.93 mm presentan una tendencia creciente. Contrario a este comportamiento, en Armenia para los años que registraron una precipitación promedio entre 131.93 mm y 196.2 mm la tendencia es creciente mientras que para años con una precipitación superior a 196.2 mm la tendencia fue decreciente.

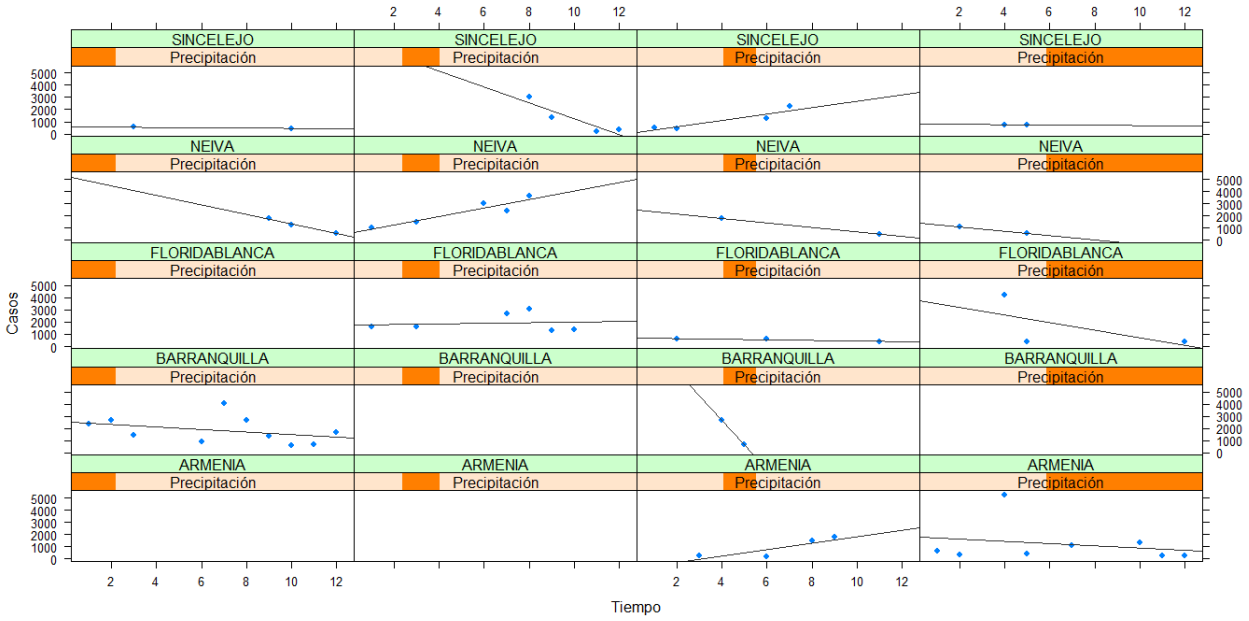

Figura 4-12.: Diagrama de dispersión según rango de valores de la variable Precipitación, clúster D.

Adicionalmente, respecto a la temperatura promedio no se presenta el diagrama xyplot dado que, por una parte no se cuenta con la información de todos los municipios y por otra la temperatura de Neiva y de Barranquilla presentan rangos similares.

4.2.3. Clúster C

El clúster C lo conforman los siguientes municipios: Cartagena, Valledupar, Santa Marta, Acacías, Itagüí, Pereira, Barrancabermeja, Girón, Piedecuesta, Palmira, Yopal y Soledad. En la figura 4-13 se presenta el comportamiento de la cantidad de casos de dengue a través del tiempo para los 12 municipios del clúster.

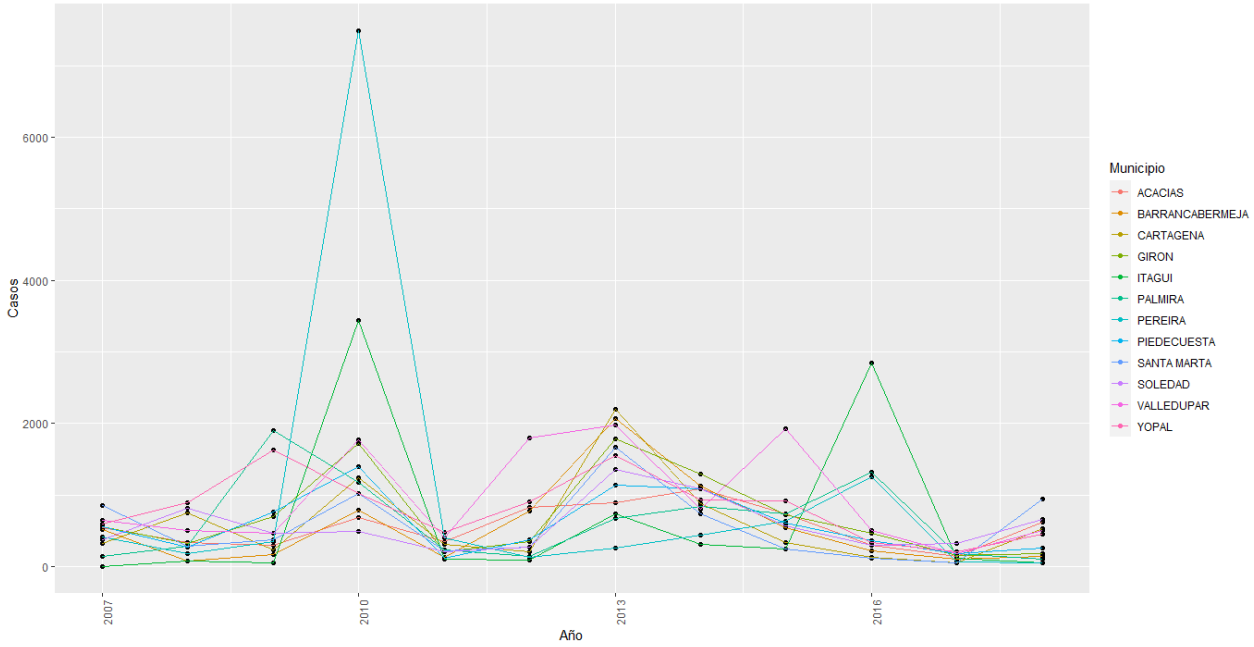

Figura 4-13.: Comportamiento de la cantidad de casos de dengue, clúster C.

En el año 2010 se registraron valores altos en la cantidad de casos de dengue reportados en los municipios que conforman el clúster, principalmente en Pereira e Itagüí, también se presentó una cantidad elevada de casos en el año 2013 en comparación con los demás periodos (exceptuando el año 2010), además, también se registraron picos importantes en Valledupar en el año 2015 y en Itagüí en el año 2016.

En la Tabla 4-4 se presentan los valores máximo y mínimo que toman las variables explicativas en cada uno de los 12 municipios; Yopal, Itagüí, Piedecuesta, Soledad y Acacías carecen de información sobre la temperatura, mientras que los municipios de Itagüí y Soledad carecen de mediciones respecto a la precipitación.

Se puede observar que, a excepción de Palmira y Acacías, todos los municipios presentan al menos una medición del IRCA inferior al 5 % es decir, no representan algún riesgo para el consumo humano; mientras que los municipios de Palmira y Acacías presentan al menos una medición del IRCA con un nivel de riesgo alto.

Los municipios de Yopal y Valledupar son los que tienen una menor densidad poblacional, de 60.78 hab/km^2 y 109.08 hab/km^2 respectivamente (para el año 2018); mientras que Itagüí y Soledad son los municipios con mayor densidad, de 16,279.06 hab/km^2 y 9,014.91 hab/km^2 para el año 2018.

Respecto a las variables climatológicas, Acacías registró la mayor precipitación en el periodo de tiempo considerado (426.45 mm) seguido de Barrancabermeja (398.1 mm) y de Valledupar (374.85 mm). Respecto a la temperatura, Pereira registró las temperaturas promedios más bajas (entre 21.9 °C y 22.3° C) y Valledupar las más altas (entre 28.5 °C y 30.2 °C).

Municipio	IRCA (%)		Densidad (hab/km^2)		Precipitación (mm)		Temperatura (°C)	
	Mín.	Máx.	Mín.	Máx.	Mín.	Máx.	Mín.	Máx.
Pereira	1.2	25.38	655.66	680.36	141.45	281.125	21.9	22.3
Valledupar	0	27.9	83.03	109.08	18.15	374.85	28.5	30.2
Yopal	2.4	29.7	41.67	60.78	217.425	360.8	-	-
Girón	0.06	16.9	304.72	337.69	0	171.5	24.4	24.7
Itagüí	0.03	4.43	14443.18	16279.06	-	-	-	-
Palmira	8.8	41	265.98	311.04	0	118.72	21.9	25.1
Cartagena	0	5.7	1271.15	1372.42	39.4	192.425	26.9	28.8
Piedecuesta	0.52	12	360.87	496	116.5	167.85	-	-
Soledad	0	1.29	7134.37	9014.91	-	-	-	-
Acacías	5.82	35.3	53.38	78.66	244.075	426.45	-	-
Santa Marta	2.3	17.3	175.81	203.92	12.2	88.675	27.2	30.4
Barrancabermeja	0	8.7	172.85	179.03	191.1	398.1	28.2	29.2

Tabla **4-4**.: Rango de las variables explicativas para cada municipio del clúster C.

Para mirar tendecias generales respecto a las variables explicativas, se empleó nuevamente la funcion xyplot pero no de forma separada (por municipio) como se presentó en los clústeres D y E sino incluyendo todas las observaciones sin distinción con el fin de analizar su comportamiento en conjunto.

En la figura **4-14** se presenta la tendencia en la cantidad de casos respecto al tiempo dividido en paneles según los valores que asume la variable IRCA. En el primer panel están las observaciones con mediciones del IRCA inferiores a 0.605 %; en el segundo, las mediciones entre 0.605 % y 4.305 %; en el tercero entre 4.305 % y 8.805 % y finalmente en el último panel entre 8.805 % y 41.005 %. La tendencia es decreciente en el primer y último panel, donde además, se encuentran algunos de los valores extremos mencionados anteriormente (los que posiblemente influyen en la línea de tendencia). Con respecto a los otros dos paneles no se evidencia una tendencia en el tiempo.

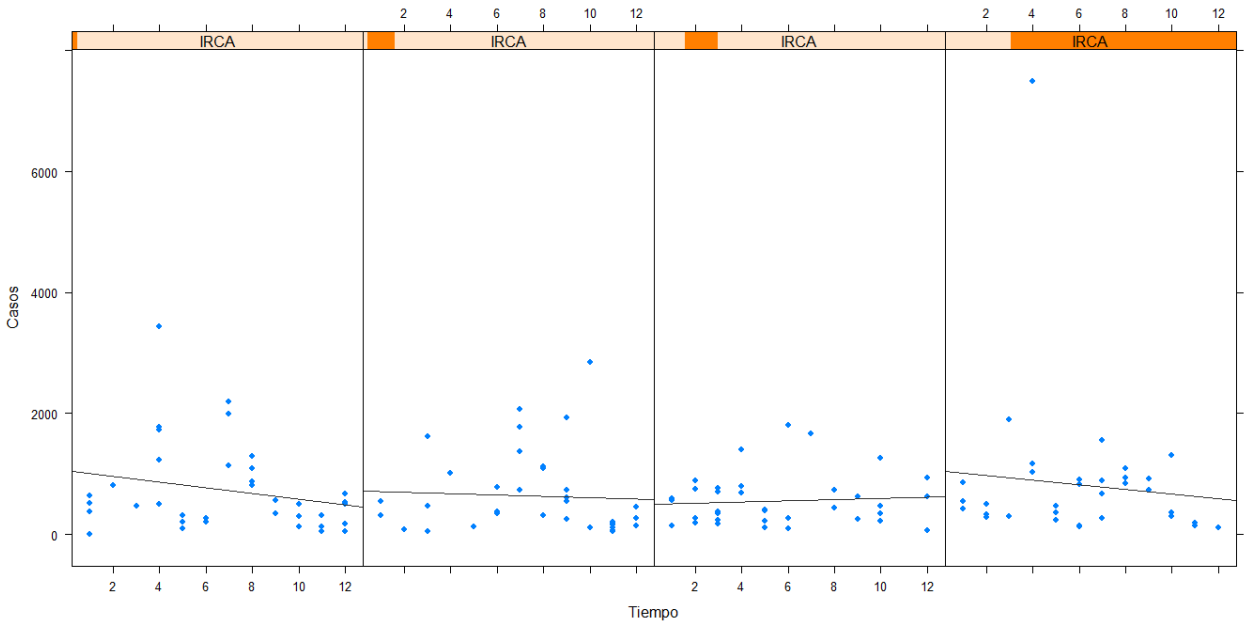

Figura 4-14.: Diagrama de dispersión según rango de valores de la variable IRCA, clústerC.

La variable densidad fue dividida en los siguientes intervalos: [41.67, 109.09) hab/km^2, [109.09, 305.9) hab/km^2, [305.9, 680.37) hab/km^2 y [680.37, 16279.07) hab/km^2. Con base en estos intervalos, se realiza el diagrama de dispersión de la cantidad de casos de dengue respecto al tiempo, los cuales se ilustran en la figura **4-15**. En el tercer panel se evidencia una tendencia decreciente en la cantidad de casos a través del tiempo mientras que en los demás paneles no se evidencian tendecias lineales fuertes.

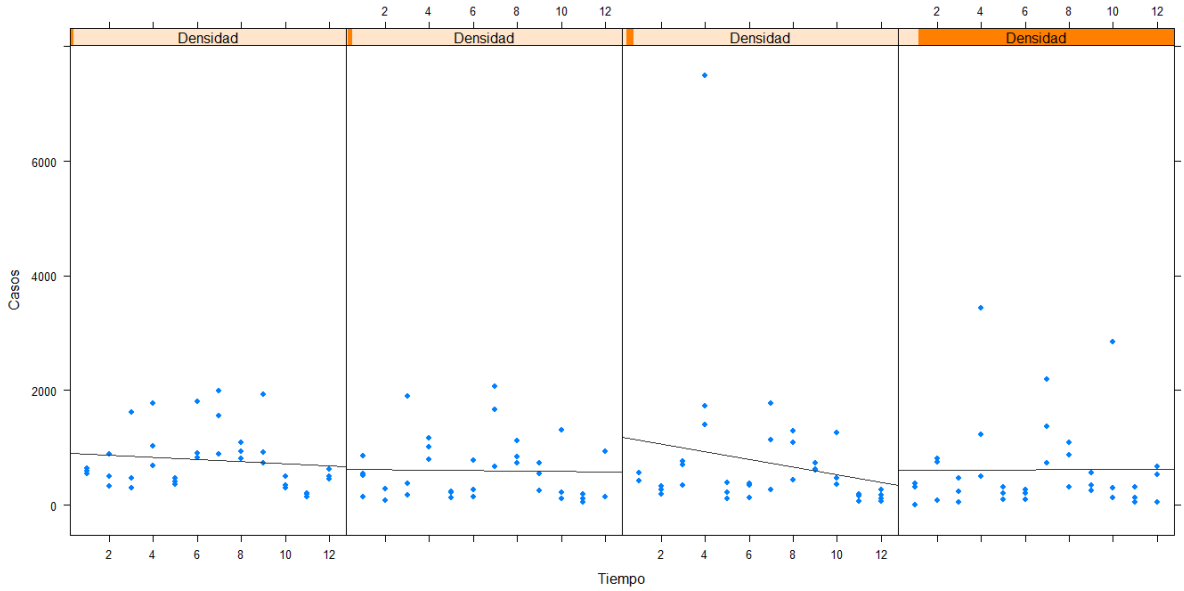

Figura 4-15.: Diagrama de dispersión según rango de la variable Densidad, clúster C.

Tomando como referencia la variable Precipitación, en la figura **4-16** se presenta el respectivo diagrama de dispersión de la cantidad de casos de dengue a través del tiempo para cada grupo. En el primer grupo se encuentran las observaciones en las cuales la precipitación promedio anual fue entre 0 mm y 85.05 mm; para el segundo grupo la precipitación toma valores entre 85.05 mm y 160.27 mm; en el tercer grupo, valores entre 160.27 mm y 270.17 mm y finalmente el último grupo corresponde a aquellas observaciones en las que la precipitación tomaba valores entre 270.17 y 426.50 mm. Si bien se presenta una tendencia lineal decreciente en los grupos 1,3 y 5, la tendencia más fuerte ocurre en el tercer grupo, es decir cuando la precipitación promedio anual fue entre 160.27 mm y 270.17 mm.

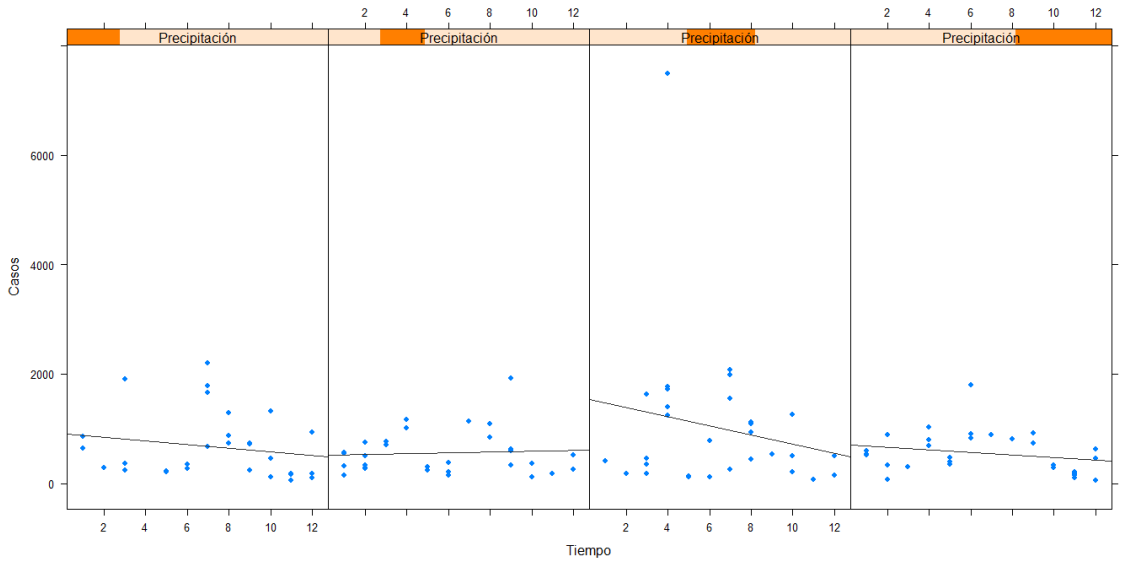

Figura 4-16.: Diagrama de dispersión según rango de la variable precipitación, clúster C.

De manera análoga al clúster D, para la variable Temperatura el anterior análisis no es diciente debido a la falta de información y a la similitud en el rango de la temperatura en los municipios que sí cuentan con dicha variable.

4.2.4. Clúster B

El clúster B lo conforman 61 municipios del país, los cuales se detallan en la Tabla **C-4** del Apéndice [C]. El comportamiento de los municipios considerados en este clúster se presenta en la figura **4-17**, dada la cantidad de municipios, es complejo identificar las trayectorias individuales, por lo cual se presentan sin el respectivo nombre.

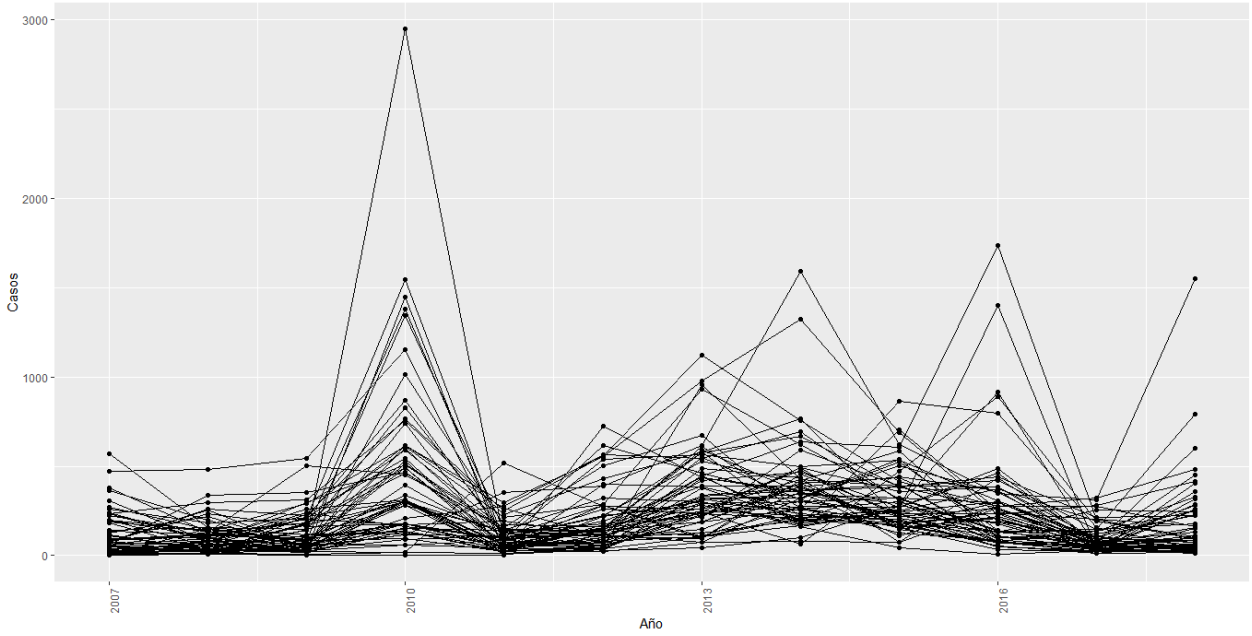

Figura 4-17.: Comportamiento de la cantidad de casos de dengue, clúster B.

Se puede observar que para el año 2010 se presentó nuevamente una cantidad elevada de casos de dengue en comparación con los demás años (el municipio con la mayor cantidad de casos en este año es Dosquebradas - Risaralda). Por otro lado, tambien se presentaron picos importantes desde el año 2013 en municipios como Fusagasugá (Cundinamarca), Montería (Córdoba) y Bello (Antioquia).

De los 61 municipios que conforman el clúster, se tienen registros de la temperatura promedio de 29 de ellos; la menor temperatura promedio fue de 20.7 °C obtenida en el municipio de Fusagasugá mientras que la temperatura promedio más alta fue de 29.5 °C registrada en el municipio de Aguachica (Cesar). Tomando los municipios que cuentan con esta variable, en la figura **4-18** se presenta el diagrama de dispersión de los casos de dengue respecto al tiempo separado en paneles según los siguientes intervalos de la temperatura promedio: [20.07 °C, 22.80 °C), [22.80 °C, 25.94 °C), [25.94 °C, 26.98 °C) y [26.98 °C, 29.35 °C). Se puede observar que en los primeros dos paneles no se presenta una tendencia lineal fuerte mientras que en el tercer panel (los años que la temperatura promedio tomó valores entre 25.94 °C y 26.97 °C) la tendencia lineal es creciente y fuerte mientras que para temperaturas promedio mayores la tendencia fue decreciente.

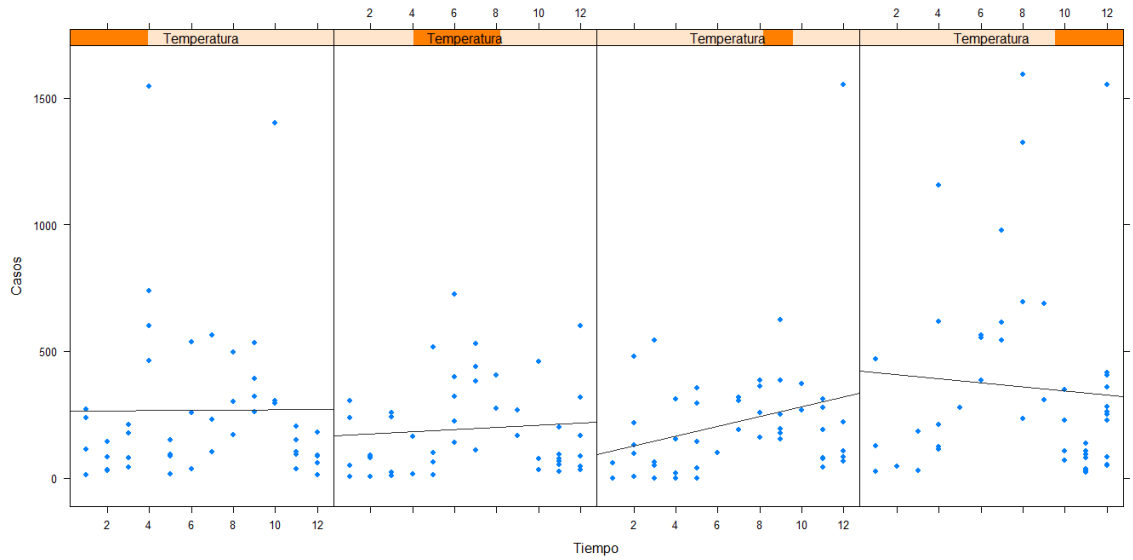

Figura 4-18.: Diagrama de dispersión según rango de la variable Temperatura, clúster B.

Respecto a la precipitación, se tiene información de 48 municipios en los que la precipitación promedio más baja fue de 0 mm en 11 municipios tales como Florencia (Caquetá), Arauca (Arauca) y Leticia (Amazonas), entre otros. La precipitación promedio más alta fue de 716.12 mm registrada en el municipio de Buenaventura (Valle del Cauca).

De igual manera, el municipio con la menor densidad poblacional en el periodo de tiempo considerado fue San Vicente del Caguán (Caquetá) con una densidad de 1.46 hab/km^2; por su parte, el municipio con la mayor densidad poblacional es Bello con una densidad de 3,677.92 hab/km^2.

Con relación al IRCA, en 25 municipios se registraron valores de 0 % lo que indica que la calidad del agua es apta para consumo humano, mientras que en los municipios de San José del Guaviare (Guaviare) y Santa Rosa del Sur (Bolívar) se registraron valores de IRCA del 100 %, esto es, no apta para consumo humano, con un nivel de riesgo alto.

Al realizar el diagrama de dispersión para las variables IRCA, densidad y precipitación promedio no se evidenciaron cambios significativos en la tendencia lineal de cada uno de los paneles. Este análisis se presenta en el Apéndice [A].

4.2.5. Clúster A

El clúster A lo conforman los municipios que registran la menor cantidad de casos de dengue, con la caractística que muchos de ellos presentan una cantidad importante de ceros, por ejemplo, en la ciudad de Bogotá de los 12 años considerados, 8 de ellos no registraron ningún caso. El comportamiento de los 1,038 municipios que pertenecen al clúster se presenta en la figura **4-19**.

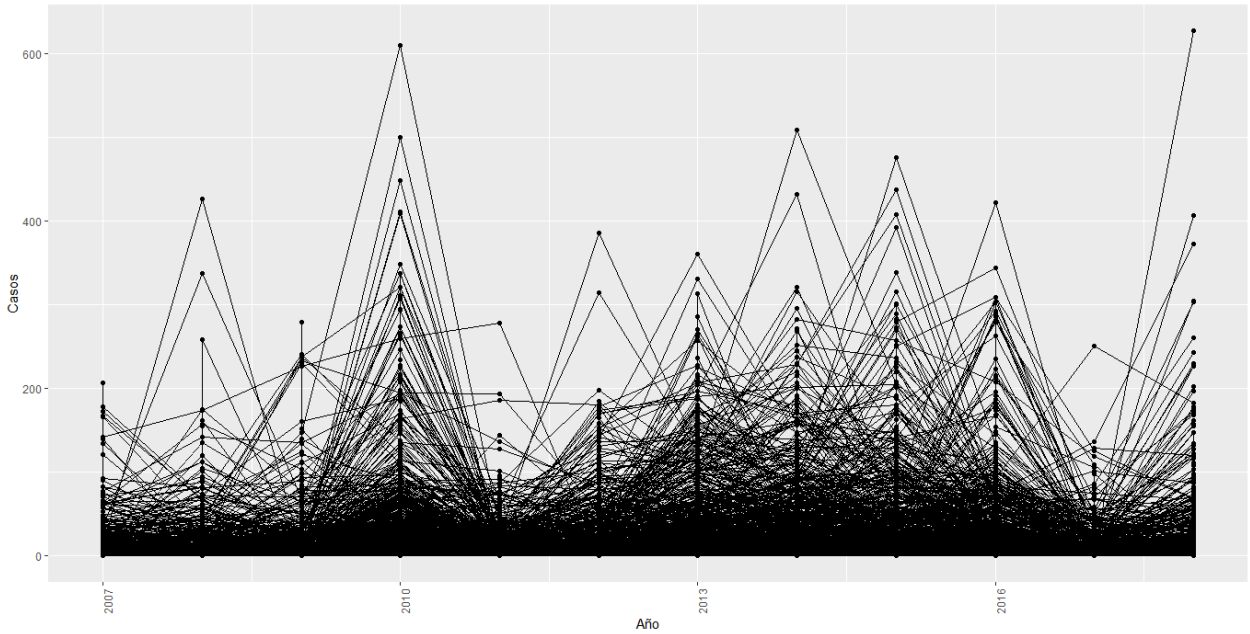

Figura 4-19.: Comportamiento de la cantidad de casos de dengue, clúster A.

Debido a la cantidad de municipios no es posible seguir trayectorias individuales, sin embargo se puede notar que muchos de las observaciones anuales fueron inferiores a 50 casos y, además, se presentó una cantidad elevada de casos en el año 2010 para una cantidad importante de municipios tal como sucedió en los anteriores clústeres; sin embargo, también se presentan algunos picos en la mayoría de los años de estudio.

Con relación a las variables explicativas, empezando con el IRCA, la menor medida obtenida fue de 0 % en al menos una de las mediciones anuales en 297 municipios; en contraste, la máxima medida obtenida fue de 100 % lo que representa que el agua distribuida es inviable sanitariamente, la cual se registró en por lo menos una medición de 22 municipios.

Por otra parte, la menor densidad poblacional registrada fue de 0.02 en el año 2007 en el municipio de Cacahual en Guanía (actualmente sigue siendo uno de los municipios con menor densidad poblacional del país). Otros municipios que registraron una densidad poblacional muy pequeña en el año 2007 son Puerto Santander (Amazonas) con una

densidad de 0.03 hab/km^2, Puerto Arica (Amazonas) con una densidad de 0.05 hab/km^2 y la Pedrera (Amazonas) cuya densidad fue de 0.07 hab/km^2. Cabe resaltar que algunos de estos municipios para dicho año eran considerados corregimientos. En contraste, hay municipios en este clúster con densidades poblacionales altas como Soacha (Cundinamarca), Sabaneta (Antioquia) y Bogotá cuyas densidades poblacionales para el año 2018 fueron de 3587.93 hab/km^2, 4670.8 hab/km^2 y 5491.67 hab/km^2 respectivamente.

En cuanto a la temperatura promedio, la menor temperatura se obtuvo en el municipio de Tona (Santander) con una temperatura promedio de 8.7 °C registrada en el año 2008. Mientras que Villavieja (Huila) registró una temperatura promedio de 30.4 °C en el año 2015. Además, de los 1,038 municipios, se cuenta con registros de 228 municipios.

Finalmente, respecto a la precipitación promedio se tiene información de 723 municipios, la menor precipitación fue de 0 mm la cuál se presentó en 73 municipios; en contraste, la mayor precipitación promedio fue de 1229.2 mm registrada en el municipio Lopez de Micay (Cauca).

Los distintos diagramas de dispersión (como los presentados en los anteriores clústeres) no permiten evidenciar cambios importantes en la tendencia al considerar diferentes intervalos de las variables explicativas. En el Apéndice [A] se realiza este análisis.

4.2.6. Municipios de estudio

Para el ajuste que se presentará en la siguiente sección se tuvieron en cuenta los municipios que cuentan con todas las variables explicativas consideradas, en total son 266 municipios de los cuales se seleccionaron 224 municipios del clúster A, 28 municipios del clúster B, junto con los municipios de Pereira, Valledupar, Girón, Palmira, Cartagena, Santa Marta y Barrancabermeja del clúster C, Barranquilla y Neiva del clúster D y Medellín, Bucaramanga, Cúcuta, Ibagué y Villavicencio del clúster E. El comportamiento de la cantidad de casos de dengue a través del tiempo para estos municipios se presenta en la figura **4-20**.

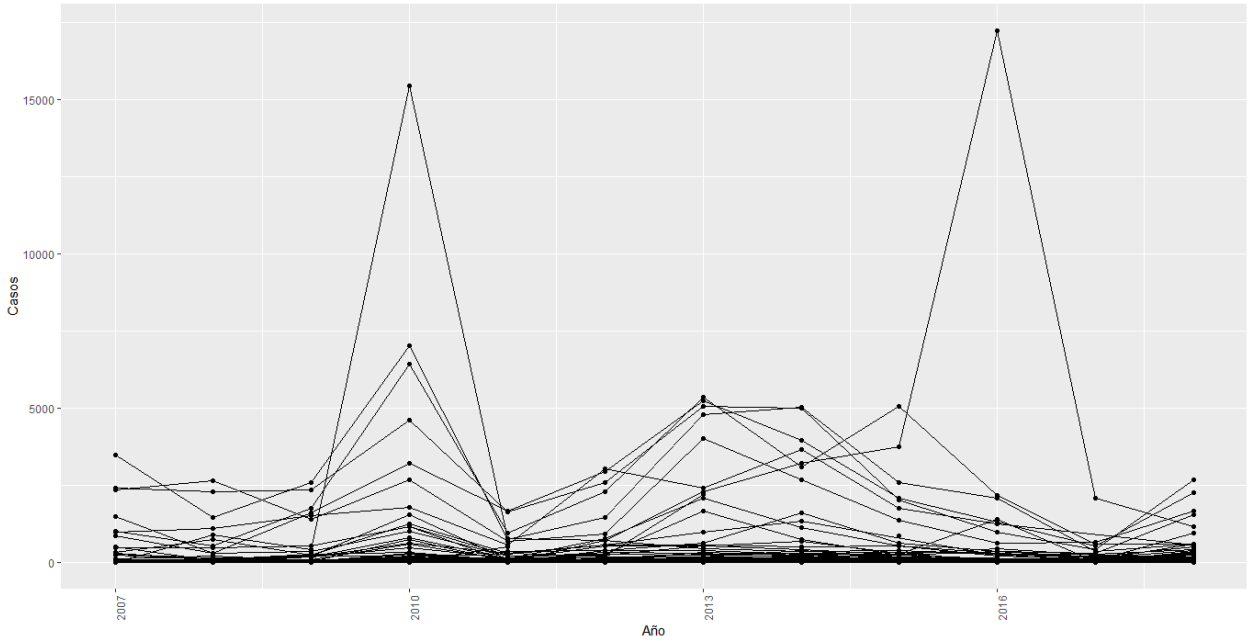

Figura 4-20.: Comportamiento de la cantidad de casos de dengue para los municipios de estudio.

Tal como se presentó previamente, las dos observaciones que superan los 15.000 casos anuales corresponden a la ciudad de Medellín y, además, se observa que en los años 2010 y entre los años 2012 y 2015 también hubo un aumento importante en la cantidad de casos de algunos municipios.

Con relación a las variables explicativas, empezando por el IRCA, 54 municipios presentaron al menos una medición del IRCA en 0 % mientras que 2 municipios registraron al menos una medición del 100 %. En total 14 municipios registraron al menos una medición superior al 80.1 % esto es, el agua distribuida es inviable sanitariamente; mientras que 146 municipios registraron al menos una medición inferior al 5 %, la cual es apta para consumo humano.

En la figura **4-21** se presenta el diagrama de dispersión de la variable de interés respecto

al tiempo, separado en grupos o paneles según el intervalo de valores que toma la variable IRCA. En el primer panel se presentan los casos de dengue en aquellos años en que IRCA presentó valores entre 0 % y 2.9 %. De manera análoga en el segundo panel para valores del IRCA entre 2.9 % y 10.9 %, el tercer panel para valores del IRCA entre 10.9 % y 32.7 % y finalmente el último panel para valores del IRCA entre 32.7 % y 100 %.
Si bien en cuanto a tendencia no se evidencia una relación fuerte, cabe destacar que los valores extremos se presentaron principalmente en el primer panel, es decir en donde el agua es apta para consumo humano.

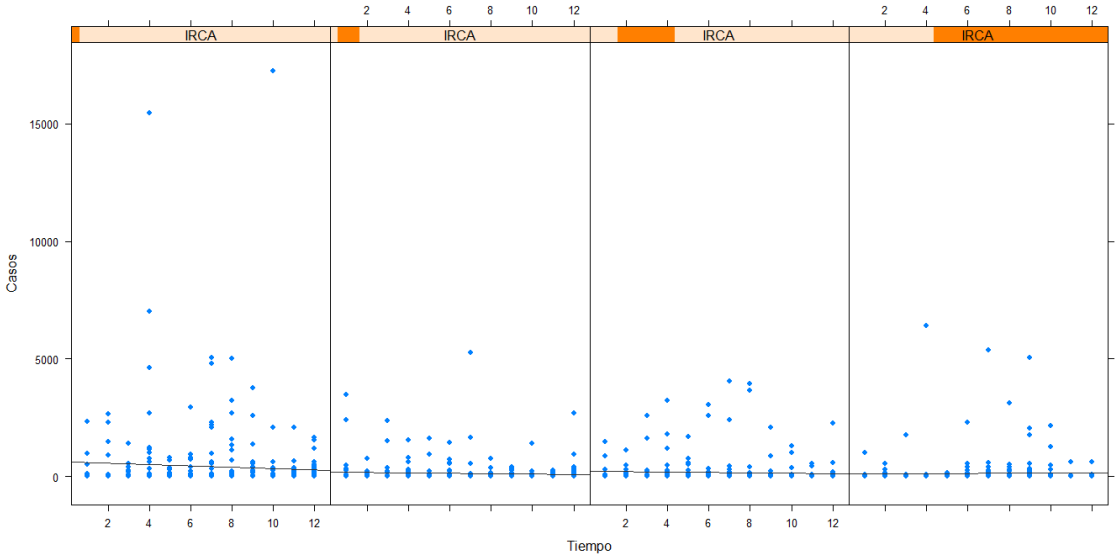

Figura 4-21.: Diagrama de dispersión según rango de la variable IRCA, municipios de estudio.

Por otra parte, el municipio con la menor densidad poblacional registrada es Solano (Caquetá) el cual para el año 2007 tenía una densidad de 0.21 hab/km^2. En cambio, la densidad poblacional más alta se registró en Barranquilla el año 2018 la cual fue de 7,416.99 hab/km^2. Asímismo, solo 10 municipios presentaron valores superiores a 1,000 hab/km^2 y 193 municipios presentaron al menos una medición inferior a 100 hab/km^2.

El diagrama de dispersión separado en paneles según intervalos que toma la variable Densidad se muestra en la figura **4-22**. Los intervalos correspondientes son: [0.2, 20.5) hab/km^2, [20.5, 50.4) hab/km^2, [50.4, 140.645) hab/km^2 y [140.6, 7417) hab/km^2.
Se puede observar que en el cuarto panel (los municipios con la mayor densidad poblacional) se presentan los casos más altos, y una tendencia levemente decreciente en el tiempo. En los demás paneles, el 98 % de las observaciones estuvieron por debajo de 500 casos anuales.

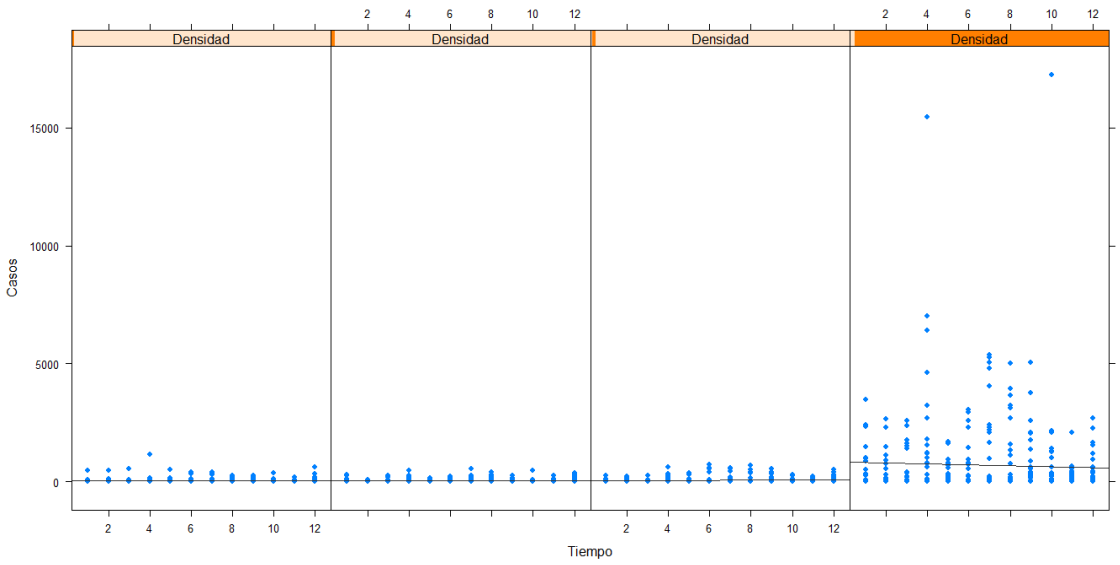

Figura 4-22.: Diagrama de dispersión según rango de la variable Densidad, municipios de estudio.

En cuanto a la precipitación promedio, el menor valor presentado fue de 0 mm (en 10 municipios), en contraste la precipitación promedio más alta se registró en el municipio Lloró (Choco) la cual fue de 795.2 mm. En total, 32 municipios tuvieron por lo menos una precipitación promedio anual superior a 200 mm. Finalmente, respecto a la temperatura promedio, la menor temperatura promedio anual se presentó en Tona (8.6° C) y la más alta en Neiva (30.7 °C).

En las figuras **4-23** y **4-24** se presentan los diagramas de dispersión de los casos de dengue respecto al tiempo separado en paneles según los intervalos de las variables precipitación y temperatura respectivamente.
Respecto a la precipitación promedio, los intervalos considerados son [0, 92.1) mm, [92.1, 130.68) mm, [130.68, 200.5) mm, y [200.5, 795.3) mm; no obstante, no se evidencian tendencias lineales fuertes y se presentan casos anuales superiores a 1,000 en cada uno de los paneles.

Por su parte, los intervalos de la variable temperatura son: [8.6, 16.9)°C, [16.9, 22.6)°C, [22.6, 26.4)°C y [26.4, 31)°C. Se puede observar que los casos más elevados de dengue ocurrieron principalmente en temperaturas superiores a 22.6 °C. El el segundo panel se resalta la presencia de Medellín que genera precisamente los valores más extremos. En cuanto a la tendencia lineal, esta es levemente decreciente en el tercer y cuarto panel.

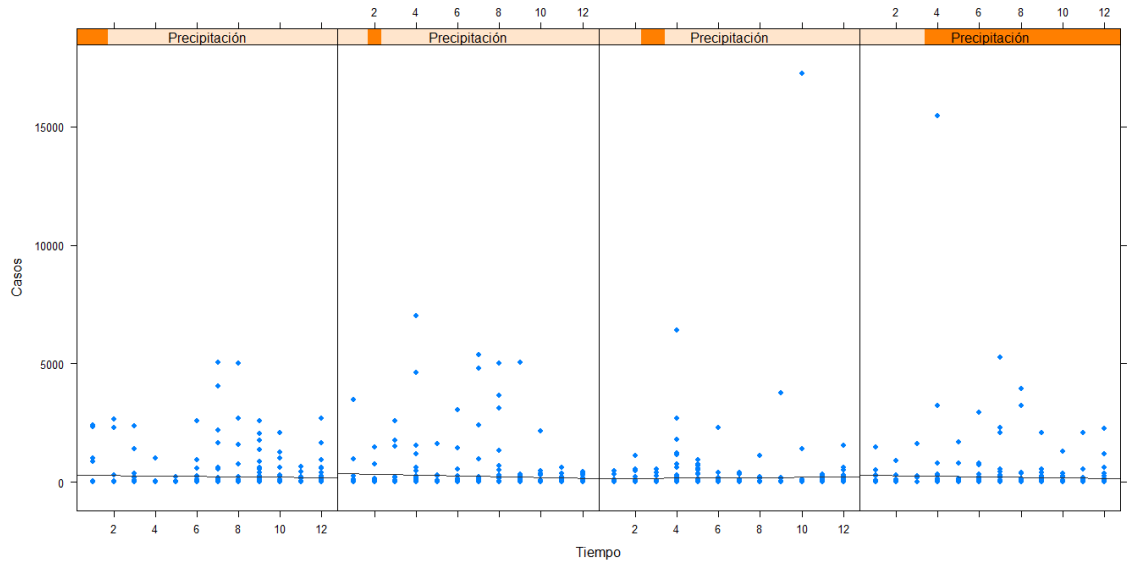

Figura 4-23.: Diagrama de dispersión según rango de la variable Precipitación, municipios de estudio.

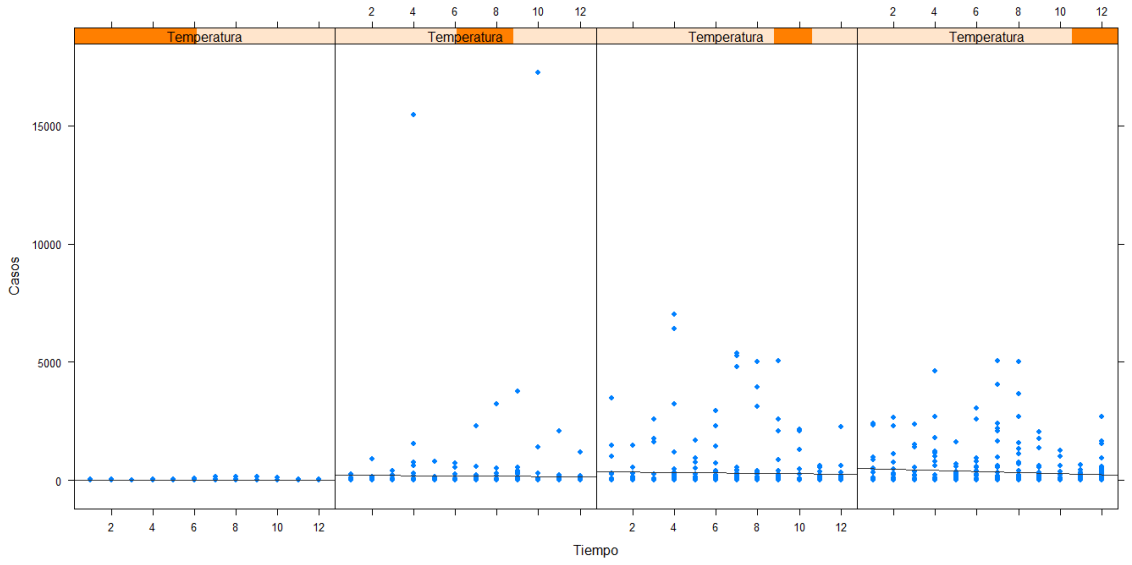

Figura 4-24.: Diagrama de dispersión según rango de la variable Temperatura, municipios de estudio.

5. Modelación longitudial de los casos de dengue

En esta sección se presenta el modelamiento de la cantidad de casos de dengue por municipio en Colombia. La variable respuesta corresponde a la cantidad de casos anuales de dengue registrados desde el año 2007 hasta el año 2018 para 266 municipios del país; las variables explicativas consideradas son: temperatura promedio anual (Tempr), precipitación promedio anual (Precpr), densidad poblacional (Dens), Índice de Riesgo de Calidad de Agua (IRCA) y tiempo (Tiempo), correspondiente a cada uno de los años en que se hizo la medición $(1, 2, ..., 12)$. Como se señaló anteriormente, es importante tener en cuenta que como no todos los municipios disponen de estaciones metereológicas, solo se incluyen en el modelo aquellos municipios que cuentan con las mediciones de las variables explicativas.

Finalmente, se incluye un término adicional en el modelo denominado variable *offset* (o de exposición) que en este caso corresponde al logaritmo de la población. La inclusión de dicha variable se debe a que el tamaño de la población es distinto para cada municipio. Por otra parte, las variables explicativas se centraron debido a los distintos rangos que estas asumen.

5.1. Modelo Poisson de Efectos Mixtos

En primer lugar, se ajustó un Modelo Poisson de efectos mixtos con todas las variables (modelo saturado), también se consideraron modelos eliminando una a una las variables explicativas; el modelo que presentó el menor AIC y menor BIC está dado por:

$$\log(\lambda_{ij}) = \log(n_{ij}) + \beta_0 + \beta_1 \text{Dens}_{ij} + \beta_2 \text{Precpr}_{ij} + \beta_3 \text{Tempr}_{ij} + \beta_4 \text{IRCA}_{ij} + \beta_5 \text{Tiempo} + \upsilon_i$$

Utilizando la función de enlace $\log(\cdot)$ con $i = 1, ..., 266$ indexando a los 266 municipios, $j = 1, ..., 12$ indicando el año correspondiente, $\log(n_{ij})$ la variable *offset*, υ_i correspondiente al intercepto aleatorio y se asume que $\nu_i \sim N(0, \sigma^2)$.

En este caso la estimación de σ^2 es de 3.26. En la Tabla **5-1** se presentan las respectivas estimaciones de los parámetros fijos del modelo y en la Tabla **5-4** se presenta la estimación del intercepto de los efectos aleatorios de 10 municipios.

Estimaciones	**Efectos fijos**					
	β_0	β_1	β_2	β_3	β_4	β_5
Media	-8.1001	0.9442	-0.1279	1.7292	-0.1558	-0.0704
Desv. Est.	0.1159	0.0349	0.0052	0.0372	0.0043	0.0026

Tabla **5-1**.: Estimación de la media y la desviación estándar de los parámetros fijos del modelo.

Municipio	**Efectos aleatorios**
	Estimación de υ_i
Medellín	-4.028
Ibagué	1.7704
Barranquilla	-7.0758
Neiva	0.8257
Cartagena	-2.0819
Santa Marta	-0.9018
Montería	-0.3908
Florencia	0.2478
Bogotá	-5.8854
Zipaquirá	-0.0656

Tabla **5-2**.: Estimación del intercepto aleatorio de 10 municipios.

Diagnóstico

Siguiendo los métodos de diagnóstico expuestos en la sección 3, se realizó el respectivo análisis de diagnóstico para el modelo ajustado. En primer lugar, en la figura **5-1** se ilustran los residuales escalados. Como se puede observar en la parte inferior del qqplot, los residuales caen fuera de la envolvente de simulación lo que implica que no se ajustan de la mejor manera, esto se corrobora con el p valor de la prueba ks ($p = 0.0011$) el cual sugiere que se rechaza la hipótesis nula de que los residuales siguen la distribución uniforme (cabe señalar que para un modelo especificado correctamente, se espera que asintóticamente los residuales escalados sigan la distribución uniforme).

Por otra parte, la prueba de dispersión (H_0: Equidispersión vs. H_1: Sobredispersión o Subdispersión) presenta un p valor de 0.89, esto es, no se rechaza la hipótesis nula de equidispersión. Además, si bien al observar la gráfica del panel derecho los residuales se encuentran alrededor de 0.5 (el equivalente a tener los residuales alrededor de 0 en los gráficos de residuales convencionales como los residuales de Pearson o estandarizados), la prueba de valores atípicos presenta un p valor de 0 lo que indica que el número de valores atípicos (cantidad de valores

que se encuentran fuera de la envolvente) es mayor de lo esperado, esto se evidencia con las estrellas resaltadas en rojo del panel derecho.

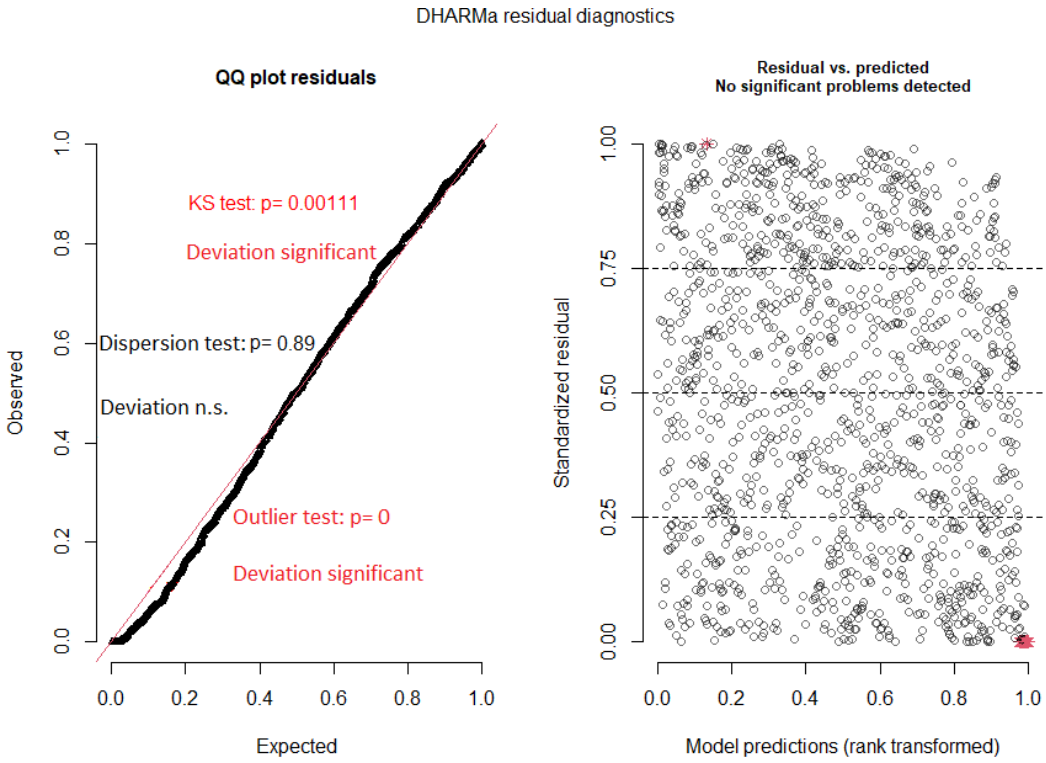

Figura 5-1.: qqplot para los residuales escalados (izq) y Residuales *vs.* valor predicho (der), modelo Poisson de efectos mixtos.

Se realizó el análisis de influencia entre grupos para ver si algún municipio resultaba ser más influyente en comparación con los demás; para ello, en la figura **5-2** se presenta la distancia de Cook así como la estimación de los parámetros al eliminar uno a uno cada municipio y estimando el modelo. Respecto a la distancia de Cook, los municipios que ejercen una mayor influencia son Medellín, Cúcuta y Bucaramanga, las tres ciudades pertenecientes al clúster E (el segundo clúster con los municipios que registraron la mayor cantidad de casos).

De igual manera, se puede ver que al remover el municipio de Medellín y reajustando el modelo, la estimación del intercepto, la densidad, la precipitación y la tempertura cambian notablemente, principalmente la densidad que cambiaría de signo. De igual manera, al eliminar los municipios de Florencia y Cúcuta y reajustando el modelo, la estimación de la precipitación disminuye; mientras que al eliminar uno a uno municipios como Cúcuta, Fusagasugá, Montería entre otros, la estimación de la temperatura cambia considerablemente; Por su parte, elimanando uno a uno municipios como Cúcuta, Villavicencio, Santa Rosa del Sur, Barranquilla e Ibagué y reajustando el modelo, la estimación del IRCA cambia.

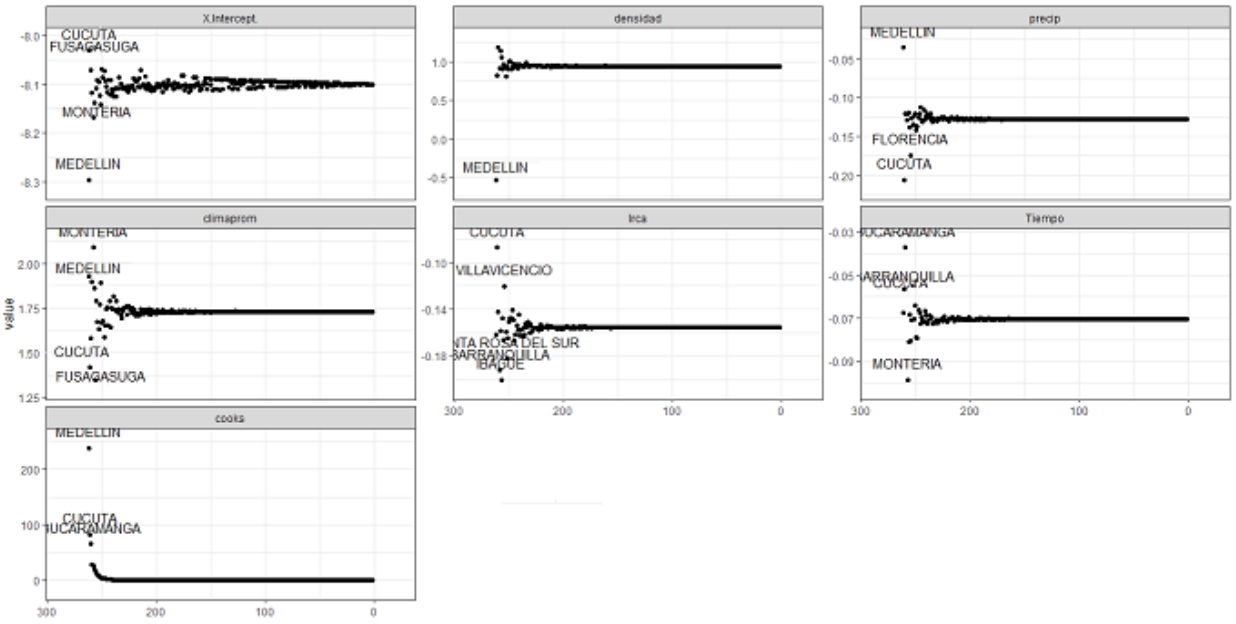

Figura 5-2.: Comportamiento de la estimación de los efectos fijos del modelo al eliminar uno a uno cada municipio y reajustando el modelo. Distancia de Cook. Modelo Poisson de efectos mixtos.

Con el fin de encontrar un mejor ajuste, se eliminaron uno a uno los municipios más influyentes y se hacía el respectivo análisis de diagnóstico; sin embargo, los resultados no fueron satisfactorios. Finalmente, a pesar de que en el modelo Poisson de efectos mixtos no se evidenció presencia de sobredispersión, en el siguiente apartado se ajusta un modelo binomial negativo con el fin de comparar resultados.

5.2. Modelo binomial negativo de efectos mixtos

Se consideró en segundo lugar un Modelo binomial negativo de efectos mixtos. El mejor modelo en términos de menor AIC y menor BIC que se obtuvo es el siguiente:

$$\log(\lambda_{ij}) = \log(n_{ij}) + \beta_0 + \beta_1 \text{Dens}_{ij} + \beta_2 \text{Precpr}_{ij} + \beta_3 \text{Tempr}_{ij} + \upsilon_i$$

Utilizando la función de enlace $\log(\cdot)$ con $i = 1, ..., 266$, $j = 1, ..., 12$, $\log(n_{ij})$ la variable *offset*, υ_i el intercepto aleatorio y se asume que $\nu_i \sim N(0, \sigma^2)$.

Para este modelo la estimación de σ^2 es de 0.758. En las Tablas **5-3** y **5-4** se presentan las respectivas estimaciones de los parámetros fijos del modelo y la estimación del intercepto de los efectos aleatorios de los 10 municipios considerados anteriormente.
A diferencia del modelo Poisson de efectos mixtos, y, a pesar de que se consideró inicialmente el modelo saturado, las variables IRCA y Tiempo no resultaron ser estadísticamente

Estimaciones	**Efectos fijos**			
	β_0	β_1	β_2	β_3
Media	-6.1392	-0.4565	-0.0891	0.5190
Desv. Est.	0.0792	0.0776	0.0451	0.0696

Tabla **5-3**.: Estimación de la media y la desviación estándar de los parámetros fijos del modelo binomial negativo.

significativas y, además, la estimación de la densidad resultó ser negativa, caso contrario al modelo Poisson. En cuanto a la estimación de la precipitación promedio y la temperatura promedio, el signo se matuvo igual al obtenido en el modelo Poisson.

Municipio	**Efectos aleatorios**
	Estimación de υ_i
Medellín	1.3003
Ibagué	0.2098
Barranquilla	1.8578
Neiva	0.0782
Cartagena	-1.2379
Santa Marta	-1.4076
Montería	-1.2559
Florencia	-1.5432
Bogotá	-3.2429
Zipaquirá	-0.5456

Tabla **5-4**.: Estimación del intercepto aleatorio de 10 municipios, modelo binomial negativo.

Diagnóstico

En la figura **5-3** se presentan los residuales escalados. Se puede observar que el p valor de la prueba ks sugiere que no se rechaza la hipótesis nula de que los residuales siguen la distribución esperada. Por otra parte, la prueba de dispersión sugiere que no hay presencia de sobredispersión ($p = 0{,}518$), mientras que la prueba de valores atípicos presenta un p valor de 0.0029 lo que indica que el número de valores atípicos es mayor de lo esperado. Estos resultados nuevamente indican que hay problemas con la especificación del modelo.

Con relación al análisis de influencia, en la figura **5-4** se presenta la distancia de Cook así como la estimación de los efectos fijos del modelo al eliminar cada municipio y reajustar el modelo.

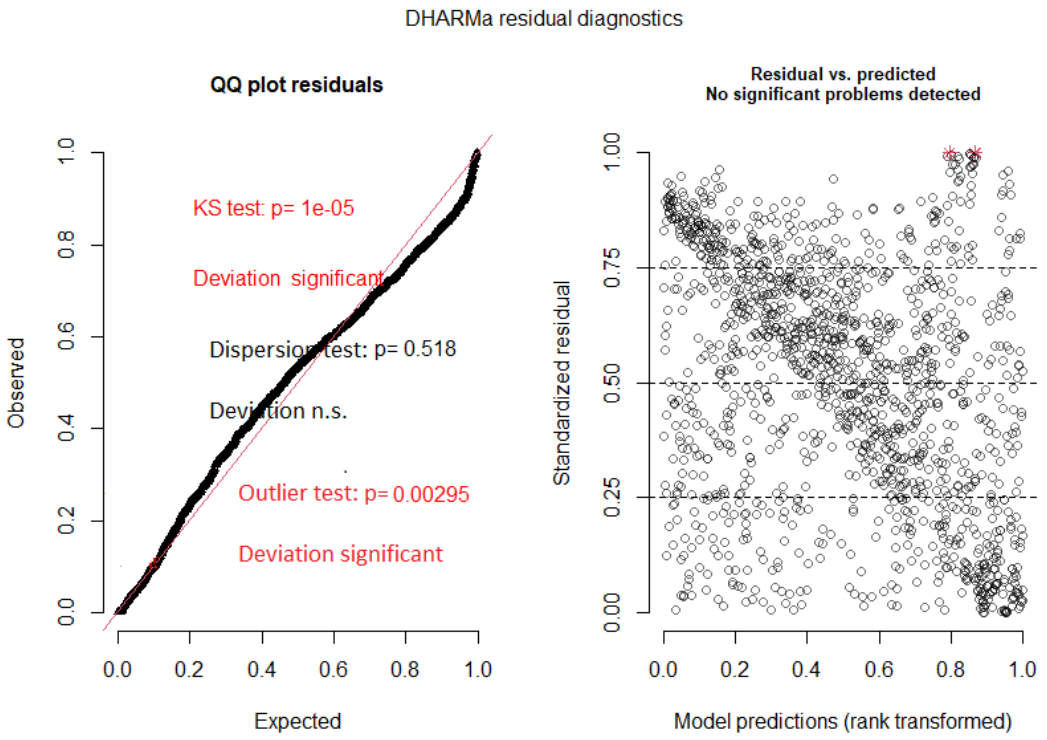

Figura 5-3.: qqplot para los residuales escalados (izq) y Residuales *vs.* valor predicho (der), Modelo binomial negativo de efectos mixtos.

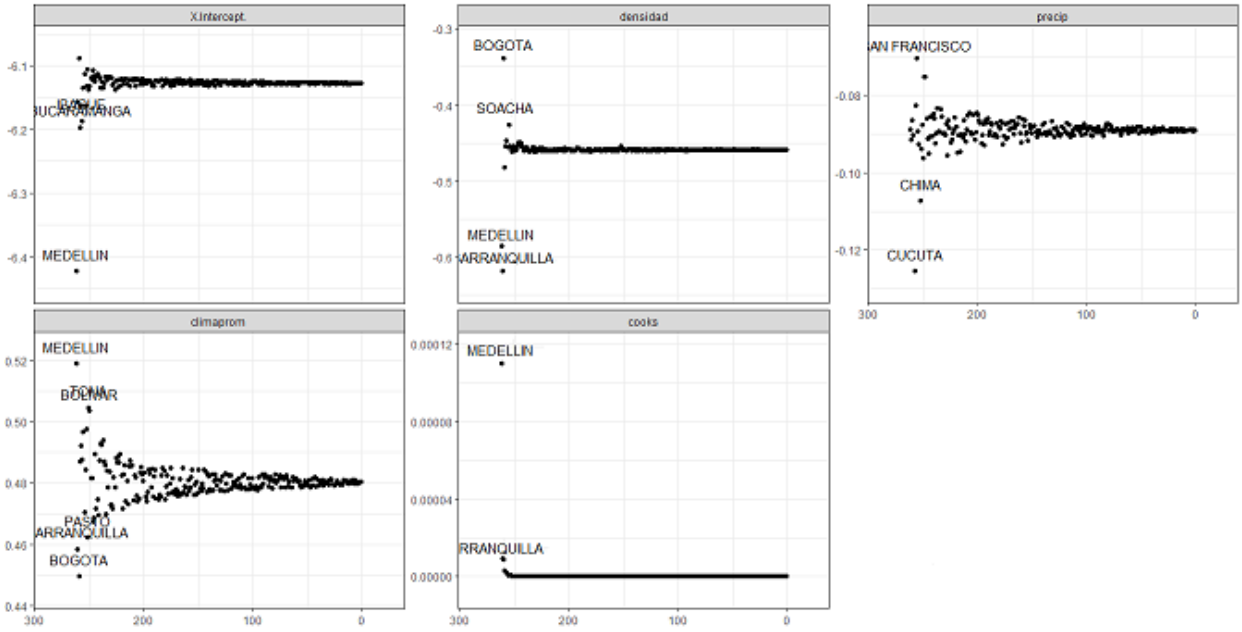

Figura 5-4.: Comportamiento de la estimación de los efectos fijos del modelo al eliminar uno a uno cada municipio y reajustando el modelo. Distancia de Cook. Modelo binomial negativo de efectos mixtos.

En cuanto al intercepto, se puede observar que al eliminar cudades como Ibagué, Bucaramanga y en especial Medellín, la estimación del parámetro cambia notablemente; respeto a la variable densidad poblacional, los municipios más influyentes son Bogotá, Soacha Medellín y Barranquilla, esto es, al eliminar cada municipio y reajustar el modelo el parámetro estimado asociado con la variable explicativa densidad presenta un cambio importante. Finalmente, respecto a las variables precipitación y temperatura, se evidencia muchos municipios influyentes en la estimación de los parámetros asociados con estas variables, en los que se destacan municipios como Chima, San Francisco, Cúcuta, Medellín, Bolívar, Pasto, Barranquilla y Bogotá.
A partir de la distancia de Cook, el municipio con mayor influencia es Medellín, seguido de Barranquilla, Bogotá, Bucaramanga, Cúcuta e Ibagué.

5.3. Modelo Poisson inflado de ceros (ZIP) de efectos mixtos

Se ajustó un Modelo Poisson inflado de ceros de efectos mixtos con todas las variables explicativas en las dos partes del modelo. El mejor modelo en términos de BIC y AIC es el siguiente:

$$\begin{aligned} \log(\lambda_{ij}) &= \log(n_{ij}) + \beta_0 + \beta_1 \text{Dens}_{ij} + \beta_2 \text{Precpr}_{ij} + \beta_3 \text{Tempr}_{ij} + \beta_4 \text{IRCA}_{ij} \\ &\quad + \beta_5 \text{Tiempo}_{ij} + \upsilon_i \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \omega_0 + \omega_1 \text{Precpr}_{ij} + \omega_2 \text{Tempr}_{ij} + b_i \end{aligned} \tag{5.3.1}$$

donde $i = 1, ..., 266$, $j = 1, ..., 12$, $\log(n_{ij})$ corresponde a la variable *offset*, υ_i y b_i los interceptos aleatorios de las partes Poisson y Bernoulli respectivamente y se asume que $\nu_i \sim N(0, \sigma_1^2)$ y $b_i \sim N(0, \sigma_2^2)$.

En este caso las estimaciones de σ_1^2 y σ_2^2 son respectivamente 3.26 y 2.26. Para la parte Poisson todas las variables resultan ser significativas mientras que para la parte Bernoulli resultaron ser significativas la precipitación y la temperatura.

En las Tablas **5-5** y **5-6** se ilustran la estimación de los parámetros fijos del modelo y la estimación del intercepto de los efectos aleatorios para los mismos municipios presentados anteriormente.

Estimaciones	Efectos fijos								
	Poisson						Bernoulli		
	β_0	β_1	β_2	β_3	β_4	β_5	ω_0	ω_1	ω_2
Media	-8.0709	1.7855	-0.0781	1.0463	-0.1102	-0.1514	-3.4407	-1.1656	0.5360
Desv. Est.	0.0441	0.0332	0.0026	0.0332	0.0051	0.0043	0.1721	0.1095	0.0660

Tabla **5-5**.: Estimación de los parámetros fijos del modelo ZIP de efectos mixtos.

Municipio	Efectos aleatorios	
	Estimación de υ_i	Estimación de b_i
Medellín	-4.6638	-0.6115
Ibagué	1.7001	-0.2844
Barranquilla	-7.8399	-0.1231
Neiva	0.7463	-0.1502
Cartagena	-2.2644	-0.1018
Santa Marta	-0.9671	-0.1031
Montería	-0.4492	-0.1900
Florencia	0.5584	2.5657
Bogotá	-5.1167	2.3436
Zipaquirá	0.1185	0.1271

Tabla **5-6**.: Estimación del intercepto aleatorio de 10 municipios, modelo ZIP.

Diagnóstico

En la figura **5-5** se ilustran los gráficos de los residuales escalados, se puede observar que el p valor de la prueba ks (p = 0,0671) sugiere que no se rechaza la hipótesis nula de que los residuales siguen la distribución esperada; por otra parte, la prueba de dispersión arroja un p valor de 0.828 lo cual indica la no presencia de sobredispersión y, finalmente, la prueba de valores atípicos presenta un p valor de 0.38231 esto es, el número de valores atípicos no es mayor ni inferior a lo esperado.

Respecto al análisis de influencia, en la figura **5-6** se presenta la distancia de Cook. Se puede observar que el comportamiento es similar al obtenido con el modelo Poisson de efectos mixtos. Medellín sigue siendo el municipio más influyente seguido de Cúcuta y Bucaramanga (los tres municipios pertenecientes al clúster E). Al eliminar Medellín y reajustar el modelo, la estimación de los parámetros asociados con las variables temperatura, densidad y precipitación cambian notablemente, donde además, se destaca que el parámetro estimado asociado con la variable densidad poblacional cambia de signo. De igual manera, al eliminar Cúcuta la estimación del parámetro asociado a las variables temperatura, precipitación e IRCA cambia considerablemente aunque conservan el signo.

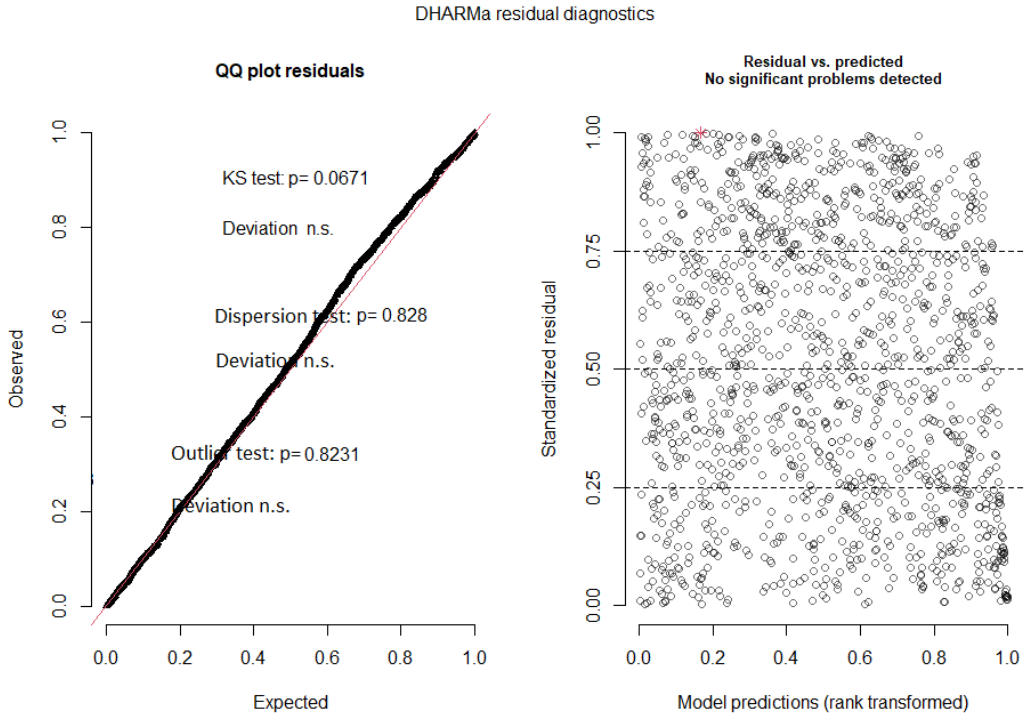

Figura 5-5.: qqplot para los residuales escalados (izq) y Residuales *vs.* valor predicho (der), modelo ZIP.

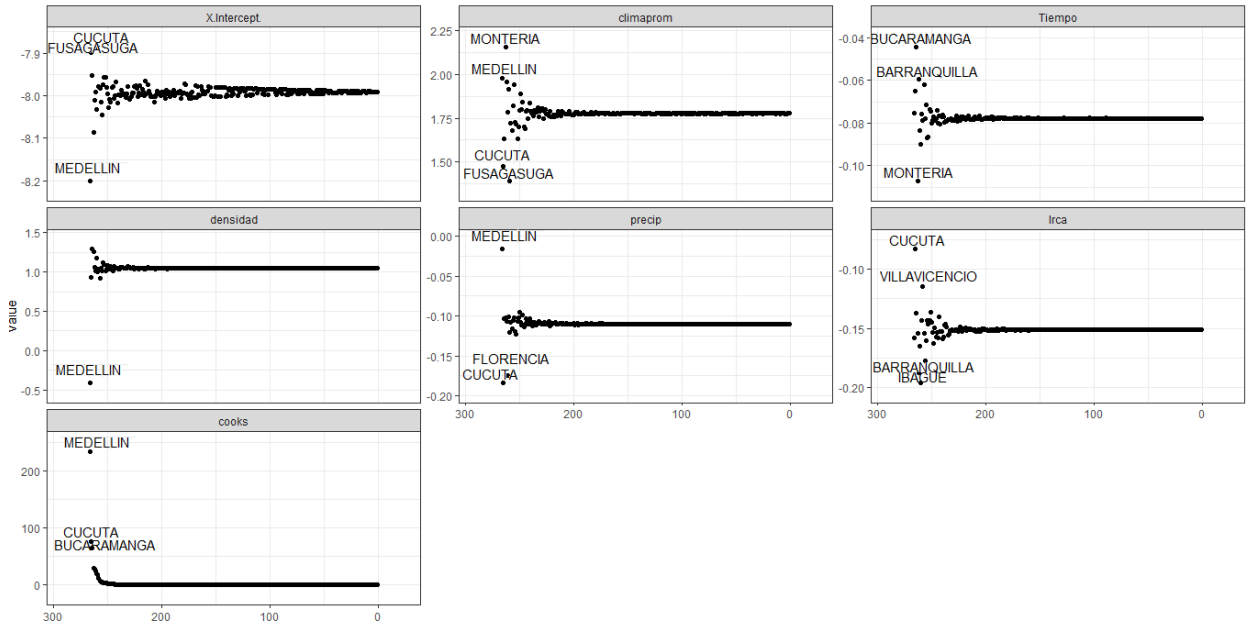

Figura 5-6.: Estimación de los parámetros al eliminar uno a uno cada municipio, distancia de Cook. Modelo ZIP de efectos mixtos.

Por su parte, el municipio de Fusagasugá es influyente principalmente en la estimación del intercepto y de la temperatura. Otras ciudades influyentes son Barranquilla, Ibagué y Montería quienes tienen incidencia especialmente en la estimación de la temperatura, el tiempo y el IRCA. Finalmente, con relación a la distancia de Cook, se puede ver que Medellín, Cúcuta y Bucaramanga presentan una distancia de Cook bastante alta, de 234.37, 75.88 y 64.65 respectivamente así como los municipios de Montería (28.26), Barranquilla (27.34), Florencia (23.78) e Ibagué (19.94).
Tendiendo en cuenta el anterior análisis y los altos valores obtenidos en la distancia de Cook para los municipios descritos previamente, se procedió a eliminar dichos municipios y se ajustó nuevamene el modelo 5.3.1. Las respectivas estimaciones de los parámetros se presentan en la Tabla **5-7**. Se puede ver que no se presenta cambios significativos (como cambios de signo) entre estos valores estimados y los presentados en la Tabla **5-5**.

Estimaciones	**Efectos fijos**								
	Poisson						**Bernoulli**		
	β_0	β_1	β_2	β_3	β_4	β_5	ω_0	ω_1	ω_2
Media	-7.9912	1.0453	-0.1101	1.7765	-1.1513	-0.0777	-2.4789	-0.90861	0.5215
Desv. Est.	0.0075	0.0299	0.0027	0.0091	0.0063	0.0051	0.1549	0.1944	0.1265

Tabla **5-7**.: Estimación de los parámetros fijos al eliminar valores influyentes.

En la figura **5-7** se presentan los residuales escalados. El p valor de la prueba ks es 0.13114, esto es, no rechaza la hipótesis nula de que los residuales siguen la distribución esperada. Además, no hay presencia de sobredispersión ($p = 0{,}89$) y no presenta exceso de valores atípicos ($p = 0{,}37605$).

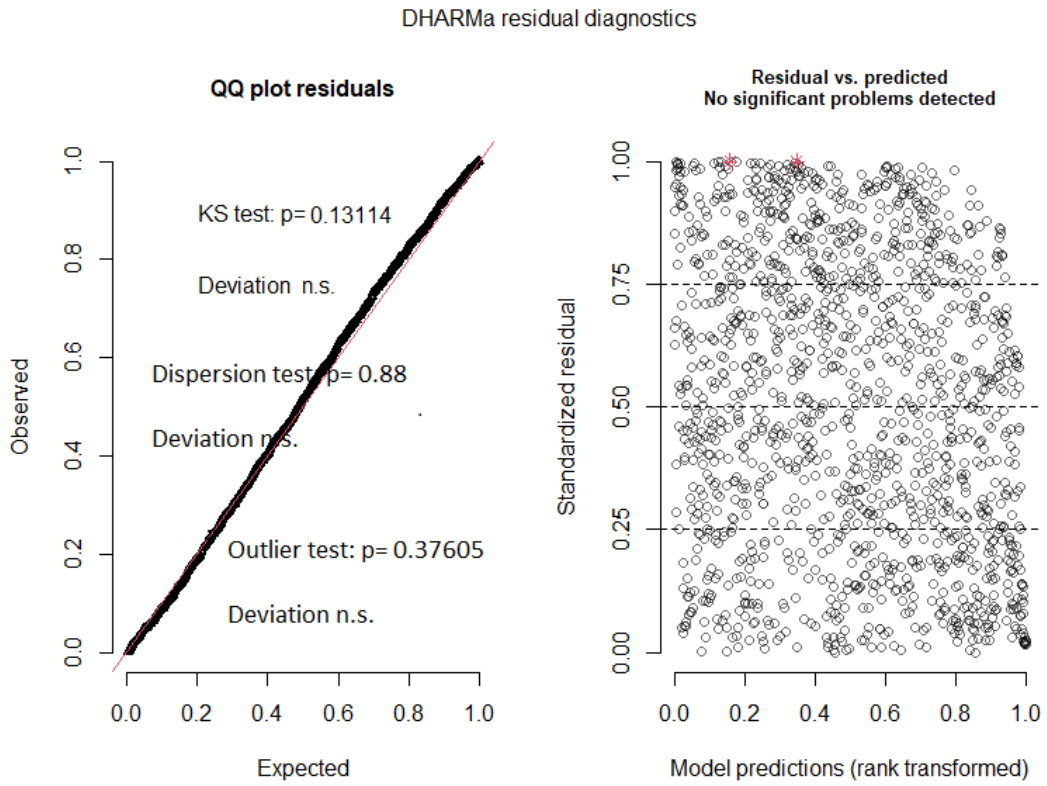

Figura 5-7.: qqplot para los residuales escalados (izq) y Residuales vs valor predicho (der), al eliminar valores influyentes.

Interpretación de resultados

De los modelos considerados el modelo que presentó mejores resultados en el análisis de diagnóstico fue el modelo Poisson inflado de ceros (ZIP) de efectos mixtos. Respecto a la estimación, el modelo Poisson y el modelo ZIP de efectos mixtos mostraron el mismo signo para los parámetros estimados asociados a las variables explicativas; mientras que el modelo binomial negativo de efectos mixtos presentó resultados diferentes en cuanto a que solo resultaron ser significativas la densidad, la precipitación y la temperatura y, además, el signo del parámetro estimado β_3 asoaciado a la densidad, era contrario al obtenido con los otros dos modelos. También se consideró el modelo binomial negativo inflado de ceros (ZINB por sus siglas en inglés) de efectos mixtos pero no resultaron significativas las estimaciones de la parte Bernoulli.

La estimación del parámetro β_3 asociado con la variable Temperatura promedio, mostró un efecto positivo sobre la cantidad de casos de dengue, de este modo, cuanto mayor es la temperatura promedio, mayor es la cantidad de casos, siendo los municipios con temperaturas superiores a 22.6 °C más susceptibles a incrementos en la cantidad de casos de

dengue. Para poner en contexto, de los 23 municipios del país que conforman los clústeres C, D, E y F que, como se ha mencionado previamente son los que presentan la mayor cantidad de casos de dengue, 14 tienen información sobre esta variable y de estos, 12 tienen todos sus registros por encima de 22.6 °C (esto es, el 88 % de las observaciones), mientras que Medellín y Pereira tuvieron una medición mínima de 21.8 °C y 21.9 °C, es decir próximas al valor de corte.

Caso contrario ocurre con la variable Precipitación, puesto que la estimación del parámetro β_2 asociada a esta variable presentó una asociación negativa, esto es, disminuciones en la precipitación promedio anual, están asociados con incrementos en los casos de dengue. En el diagrama de dispersión por intervalos de valores de esta variable, se pudo observar que hay una concentración alta de observaciones en las cuales se presentó una cantidad importante de casos de dengue, principalmente con precipitaciones inferiores a 130.68 mm. No obstante, tomando como referencia los 23 municipios con más casos, 21 tienen registros de esta variable y solamente el 42 % de las observaciones de estos municipios registraron un precipitación inferior a 130.68 mm.

Por su parte, la densidad mostró una relación positiva sobre la cantidad de casos de dengue, esto es, cuanto mayor es la densidad poblacional mayor es la cantidad de casos registrados. Aquellos municipios con densidades superiores a 140.6 hab/km^2 son los que presentaron una gran cantidad de casos. Tomando como referencia los municipios con más casos, el 87 % de las observaciones presentaron densidades poblacionales superiores a este valor, solamente los municipios de Valledupar, Acacías y Yopal no superaron este umbral. Además, 7 municipios (Barranquilla, Bucaramanga, Cali, Floridablanca, Itagüí, Medellín y Soledad) registraron densidades poblacionales superior a 3,000 hab/km^2.

Respecto al Índice de Riesgo de Calidad del Agua, la estimación del parámetro β_4 asociado con esta variable, es negativa. En consecuencia, disminuciones en el IRCA, se asocian con incrementos en la cantidad de casos de dengue. Para ilustrar esto, municipios como Medellín, Bucaramanga, Barranquilla, Cartagena o Armenia registraron todas las mediciones de esta variable, por debajo del 16 % es decir representan niveles de riesgo bajo en la calidad de agua; sin embargo, son ciudades que registraron una cantidad de casos alta.

Finalmente, el tiempo también presentó un efecto negativo sobre la cantidad de casos de dengue registrados, esto quiere decir que con el paso de los años ha ido disminuyendo la cantidad de casos de dengue.

Adicionalmente, cabe mencionar que se realizaron ajustes de modelos por clúster pero no resultaron ser los más óptimos, debido entre otros aspectos a la falta de información que presentaban algunos municipios al interior de cada clúster o a la similitud que presentaban

especialmente en los rangos climatológicos, lo que hacía que, por ejemplo, la temperatura no resultara ser significativa en ningún modelo considerado por clústeres.

6. Conclusiones y recomendaciones

6.1. Conclusiones

En el presente trabajo se presentó una aplicación de los Modelos Lineales Generalizados de efectos Mixtos (GLMM) específicamente de los modelos Poisson y ZIP, en los cuales la variable de interés corresponde a la cantidad de casos de dengue anuales por municipio. Estos modelos se ajustaron siguiendo la metodología clásica. Además, se realizó el respectivo análisis de influencia, mediante medidas como los DFBETAS y la distancia de Cook y el análisis de residuales, siguiendo la metodología propuesta por [Hartig, 2020] para los modelos GLMM.

En cuanto a la cantidad de casos de dengue en Colombia, se encontró que temperaturas promedio anual altas, superiores a 22.6 °C, así como densidades poblacional altas (superiores a 140.6 hab/km^2, menores niveles de precipitación (menores a 130.68 mm) y niveles de riesgo del agua bajos, estaban asociados con una mayor cantidad de casos de dengue y, en general, con el paso de los años (tomando como referencia del año 2007 al año 2018) los casos han disminuido.

Estos resultados coinciden con los factores ambientales, sociales y biológicos que permiten la proliferación del vector, presentados en el capítulo 1. Un aspecto a resaltar tiene que ver con la calidad del agua, ya que se encontró que a mejores niveles de agua, se espera una mayor cantidad de casos, este resultado se puede asociar con lo mencionado por [Ochoa et al., 2015] quién afirma que el vector pone sus huevos en agua límpia. Otra característica biológica importante es que el vector prospera principalmente con temperaturas entre 15 °C y 40 °C, en investigaciones previas se encontró que la temperatura adecuada para su transmisión eran aquellas superiores a 30 °C. En el presente estudio se encontró que, para los municipios con mayor cantidad de casos que contaban con esta variable, el 88 % de las observaciones registraron temperaturas promedio anuales superiores a 22.6 °C.

En cuanto a la precipitación, en investigaciones previas como en [Rúa et al., 2013] encontraron un efecto positivo de esta variable respecto a la cantidad de casos de dengue para la ciudad de Medellín, mientras que en [Cassab et al., 2010] no se observó asociación estadística entre las variables climatológicas y la cantidad de casos en Montería. Sin embargo, cabe resaltar que estos trabajos se centraron específicamente en los municipios mencionados,

pero al tomar una mayor cantidad de municipios, es posible que dicho efecto sea negativo como se ilustra en el presente estudio.

Los anteriores resultados sirven para tener un panorama general de algunas caraterísticas que presentan los municipios y que resultan ser favorables para la proliferación del dengue. Además, sirven para que municipios con estas características puedan tomar medidas sanitarias más contundentes, ya que si bien aspectos como la calidad del agua se cumplen en algunos casos, quizás no se hace un seguimiento a aspectos como aguas estancadas, abastecimiento de agua potable a todas las regiones y en otra vía a aspectos como el rápido crecimiento poblacional y la poca planificación de zonas urbanas.

Finalmente, otro aspecto a tener en cuenta es la calidad hospitalaria (o incluso la ausencia de hospitales) de los municipios cercanos a las ciudades principales con las características descritas, puesto que este hecho puede hacer que los habitantes de dichos municipios prefieran ser atendidos en ciudades principales (Cali, Medellín, Bucaramanga, etc.) lo que influye en la alta cantidad de casos que se registran en estan ciudades, y que puede influir en el hecho de que no se estén tomando las medidas necesarios en los municipios que también lo requieren.

6.2. Recomendaciones

Para futuras investigaciones se pueden considerar modelos espacio-temporales usuales o modelos de datos longitudinales correlacionados [Cepeda, 2011].

A. Anexo: Diagramas de dispersión clústeres A y B

En la figura **A-1**, se presenta el diagrama de dispersión separado en paneles según los siguientes rangos de la variable IRCA: [0 %, 2.3 %), [2.3 %, 8.7 %), [8.7 %, 22.2 %) y [22.2 %, 100.05 %). A pesar de la diferencia que existe en los rangos, la tedencia prácticamente no cambia de un grupo a otro. Sin embargo, un aspecto a destacar es que los valores extremos se encuentran en los años con un IRCA inferior a 22.2 %

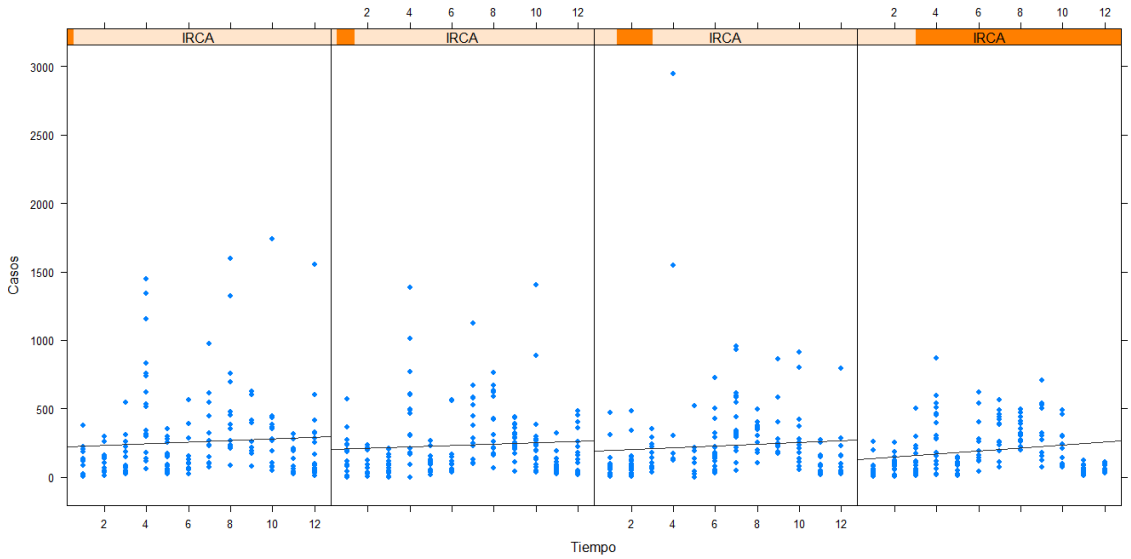

Figura A-1.: Diagrama de dispersión según rango de la variable IRCA, clúster B.

De manera análoga, en la figura **A-2** se presenta el diagrama de dispersión respectivo dividido según los siguientes rangos de la variable densidad poblacional: [1.45, 49.08) hab/km^2, [49.08, 140.88) hab/km^2, [140.88, 267.09) hab/km^2 y [267.09, 3677.92) hab/km^2. Tal como ocurrió con la variable IRCA, la tendencia creciente es similar de un grupo a otro, aunque se resalta que los valores extremos se presentan en el grupo con la densidad poblacional más alta.

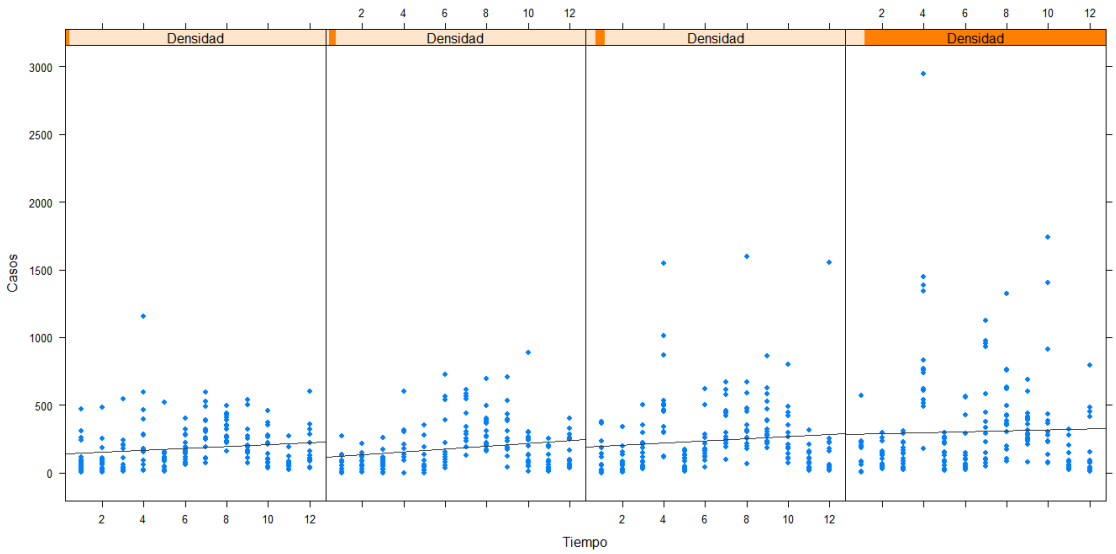

Figura A-2.: Diagrama de dispersión según rango de la variable Densidad, clúster B.

En la figura **A-3**, se presenta el respectivo diagrama de dispersión según los intervalos que toman la variable precipitación: [0, 97.38)mm, [97.38, 133.48)mm, [133.48, 212.58)mm y [212.58, 716.14). Se puede observar que la tendencia es creciente en los paneles 1, 2 y 4 pero es más fuerte en los primeros dos paneles.

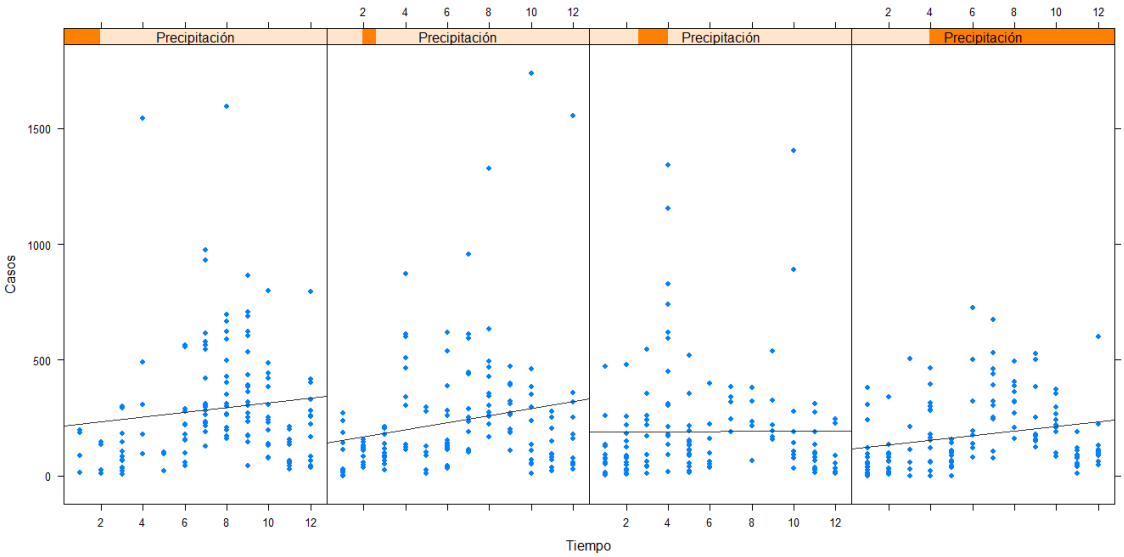

Figura A-3.: Diagrama de dispersión según rango de la variable precipitación, clúster B.

Para el clúster A, en la figura **A-4** se presenta el diagrma de dispersión de los casos de dengue respecto al tiempo separado según intervalos de la variable IRCA, el primer intervalo es de 0 % a 5.3 %, el segundo de 5.3 % a 19.6 %, el tercer intervalo es de 19.6 % a 42.1 % y finalmente de 42.1 % a 100 %. No se observan cambios significativos de un panel a otro.

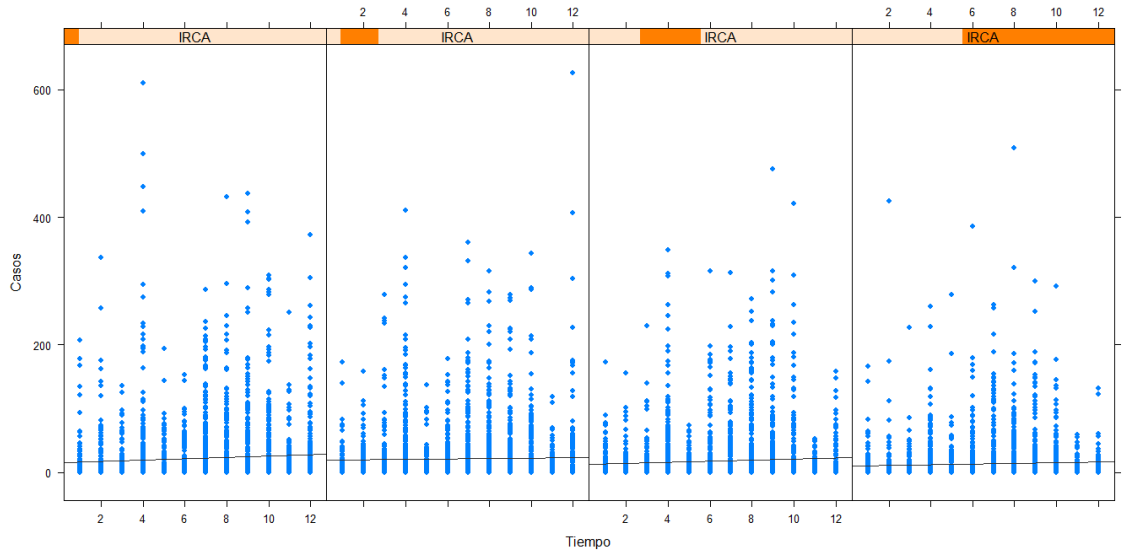

Figura A-4.: Diagrama de dispersión según rango de la variable IRCA, clúster A.

Tomando como referencia la variable densidad, esta se dividió en los siguientes intervalos: [0, 18.8) hab/km^2, [18.8, 40.2) hab/km^2, [40.2, 76.8) hab/km^2 y [76.8, 5491.7) hab/km^2, y se construyó el correspondiente diagrama de dispersión. De manera análoga como ocurrió con el IRCA, no se evidencian cambios significativos de un panel a otro como se presenta en la figura **A-5**.

En la figura **A-6**, se presenta el diagrama de dispersión de la variable precipitación, la cuál se dividió en los siguientes intervalos: [0, 90.6)mm, [90.6, 137), [137, 212.4)mm y finalmente [212.4, 1229)mm. Se observa una leve tendencia lineal en el último panel, es decir en las observaciones en las cuales la precipitación fue superior a 212.6 mm. En los demás paneles no se evidencia tendencias fuertes.

Finalmente, respecto a la variable Temperatura, se dividió en los siguientes intervalos: [10.6, 16.1)°C, [16.1, 21.2)°C, [21.2, 26.1)°C y de [26.1, 30.5)°C. Para las observaciones del primer panel y tercer panel, no hay tendencia en los casos registrados, en el segundo panel se observa una leve tendencia lineal decreciente mientras que en el último panel (temperatura promedio mayor a 26.1 °C) la tendencia es creciente.

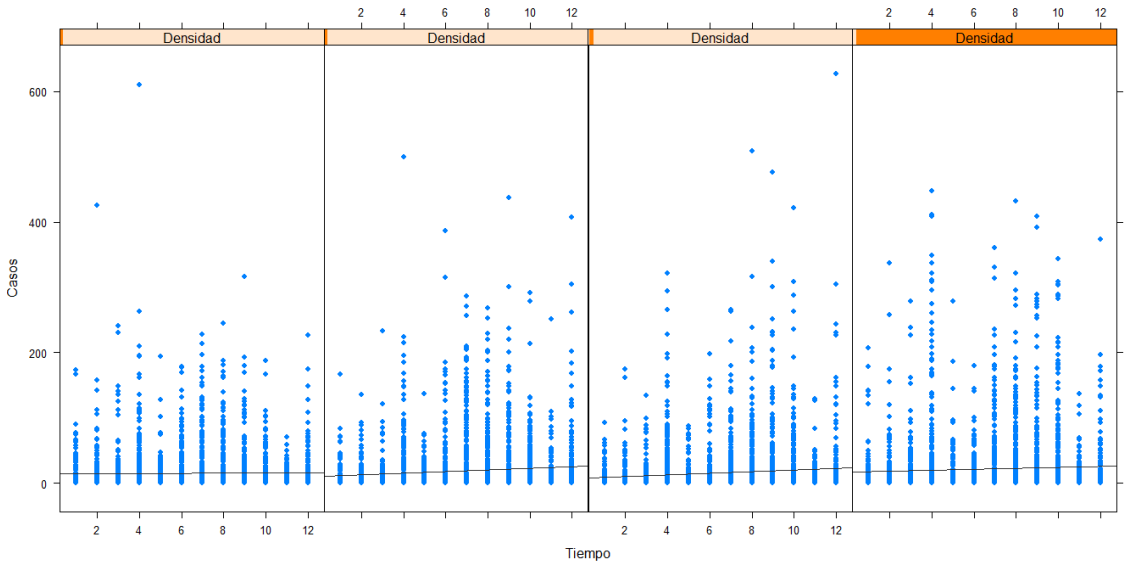

Figura A-5.: Diagrama de dispersión según rango de la variable densidad, clúster A.

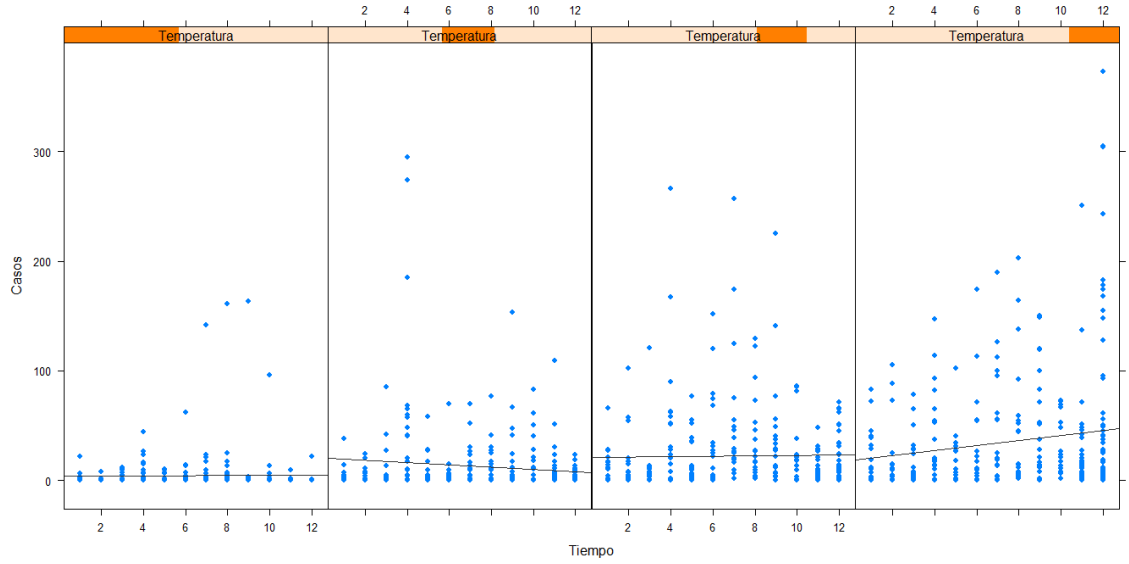

Figura A-7.: Diagrama de dispersión según rango de la variable Temperatura, clúster A.

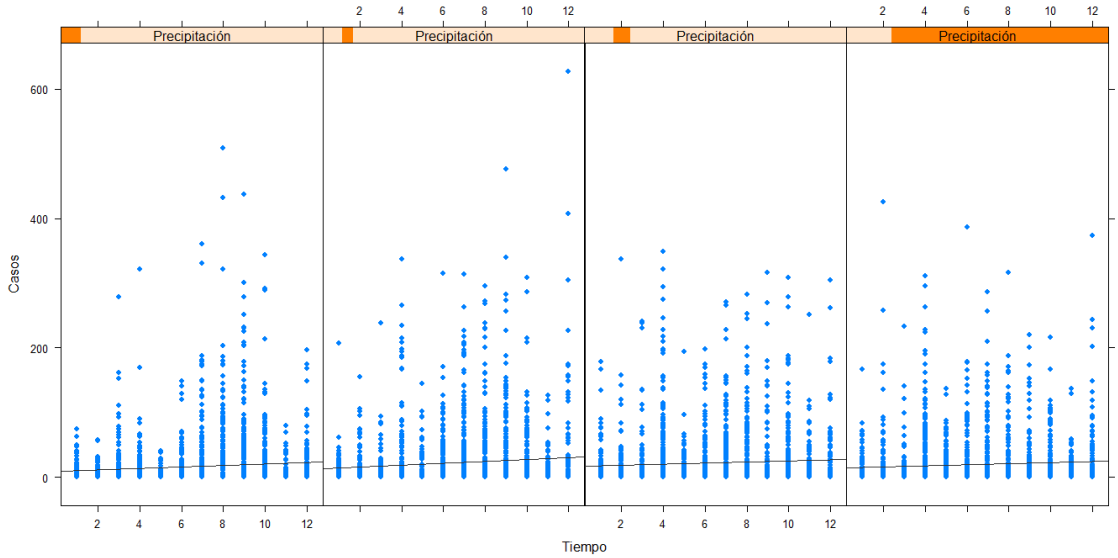

Figura A-6.: Diagrama de dispersión según rango de la variable precipitación, clúster A.

B. Anexo: Código de R

En el presente anexo se ilustra parte del código empleado en el proyecto.

El código de R que se implementó para ajustar la metodología k-means para datos longitudinales es el siguiente:

```
library("kml")
library(clv)
library(cluster)
library(class)
library(reshape)
library(longitudinalData)
library(rgl)
library(rgl)
library(dplyr)
library(ggplot2)
library(tidyr)
library(tidyverse)
library(tigerstats)
library(glmmTMB)
library(DHARMa)

mun2dengue=rename(mun2dengue,c(Nombre="id"))
denguecld <- cld(mun2dengue, timeInData = 3:50)

kml(denguecld, nbRedraw = 2, toPlot = "both")
kml(denguecld)

kml(denguecld, 4, parAlgo = parALGO(distance = function(x, y)
  + cor(x, y), saveFreq = 10))

X11(type="Xlib")
choice(denguecld)
  plotAllCriterion(denguecld)
```

El código empleado para graficar el comportamiento de los municipios, así como la dispersión entre la variable respuesta y el tiempo, separado según intervalos de valores quue asume una variable explicativa es:

```
ggplot(muncE, aes(Periodo,Casos, group = Municipio))+
  geom_point(aes(colour = factor(Municipio))) +
  geom_point() +
  geom_line(aes(colour = factor(Municipio)))+
  theme(axis.text.x = element_text(angle = 90))+
  labs(x="Periodo",y="Casos",color="Municipio")

irca <- equal.count(Irca, number = 2, overlap = 0)
xyplot(Casos ~ Tiempo | irca * Municipio,
       data = CE, pch = 19, type = c("p","r"),
       layout = c(2,5))

Dens<- equal.count(densidad, number = 2, overlap = 0)
xyplot(Casos ~ Tiempo | Dens * Municipio,
       data = CE, pch = 19, type = c("p","r"),
       layout = c(2,5))

prec<- equal.count(precip, number = 2, overlap = 0)
xyplot(Casos ~ Tiempo | prec * Municipio,
       data = CE, pch = 19, type = c("p","r"),
       layout = c(2,5))

cl<- equal.count(climaprom, number = 2, overlap = 0)
xyplot(Casos ~ Tiempo | cl * Municipio,
       data = CE, pch = 19, type = c("p","r"),
       layout = c(2,5))
```

El ajuste de los modelos propuestos Poisson de efectos mixtos, binomial negativo de efectos mixtos y ZIP de efectos mixtos se realizaron mediante el siguiente código:

```
mod1 <- glmmTMB(Casos~densidad+precip+climaprom+Tiempo+Irca+ offset(log(Población))+
                (1 |ID), data =ps2 , family = poisson)

mod2 <- glmmTMB(Casos~densidad+precip+climaprom+ offset(log(Población))+
                (1 |ID), data =ps2, family = poisson)
```

```
mod3 <- glmmTMB(Casos~densidad+precip+climaprom+Tiempo+Irca+ offset(log(Población))+
                (1 |ID), data =ps2, ziformula =~precip+climaprom+(1 |ID),
                family = poisson)
```

El código empleado para el análisis de diagnóstico es el siguiente:

```
testDispersion(mod1)
simulationOutput <- simulateResiduals(fittedModel = mod1,plot = T)

source(system.file("other_methods","influence_mixed.R", package="glmmTMB"))
owls_nb1_influence_time <- system.time(
  owls_nb1_influence <- influence_mixed(mod1, groups="ID")
)
car::infIndexPlot(owls_nb1_influence)

inf <- as.data.frame(owls_nb1_influence[["fixed.effects[-ID]"]])
inf <- transform(inf,
                 ID=rownames(inf),
                 cooks=cooks.distance(owls_nb1_influence))
inf$ord <- rank(inf$cooks)
if (require(reshape2)) {
  inf_long <- melt(inf, id.vars=c("ord","ID"))
  gg_infl <- (ggplot(inf_long,aes(ord,value))
              + geom_point()
              + facet_wrap(~variable, scale="free_y")
              ## n.b. may need expand_scale() in older ggplot versions ?
              + scale_x_reverse(expand=expansion(mult=0.15))
              + scale_y_continuous(expand=expansion(mult=0.15))
              + geom_text(data=subset(inf_long,ord>24),
                          aes(label=ID),vjust=-1.05)
  )
  print(gg_infl)
}
```

C. Anexo: Municipios por Clúster

- **Cluster E**

MUNICIPIO	DEPARTAMENTO
VILLAVICENCIO	META
MEDELLIN	ANTIOQUIA
CUCUTA	NORTE SANTANDER
BUCARAMANGA	SANTANDER
IBAGUE	TOLIMA

Tabla C-1.: Municipios que conforman el clúster E.

- **Cluster D**

MUNICIPIO	DEPARTAMENTO
NEIVA	HUILA
ARMENIA	QUINDIO
FLORIDABLANCA	SANTANDER
SINCELEJO	SUCRE
BARRANQUILLA	ATLANTICO

Tabla C-2.: Municipios que conforman el clúster D.

- **Cluster C**

MUNICIPIO	DEPARTAMENTO
CARTAGENA	BOLIVAR
VALLEDUPAR	CESAR
SANTA MARTA	MAGDALENA
ACACIAS	META
ITAGUI	ANTIOQUIA
PEREIRA	RISARALDA
BARRANCABERMEJA	SANTANDER
GIRON	SANTANDER
PIEDECUESTA	SANTANDER
PALMIRA	VALLE DEL CAUCA
YOPAL	CASANARE
SOLEDAD	ATLANTICO

Tabla C-3.: Municipios que conforman el clúster C.

- **Cluster B**

MUNICIPIO	DEPARTAMENTO
MONTERIA	CORDOBA
ESPINAL	TOLIMA
BELLO	ANTIOQUIA
GIRARDOT	CUNDINAMARCA
DOSQUEBRADAS	RISARALDA
VILLA DEL ROSARIO	NORTE SANTANDER
LOS PATIOS	NORTE SANTANDER
MELGAR	TOLIMA
CALARCA	QUINDIO
GARZON	HUILA
ARAUCA	ARAUCA

MUNICIPIO	DEPARTAMENTO
CARTAGO	VALLE DEL CAUCA
TULUA	VALLE DEL CAUCA
AGUAZUL	CASANARE
PITALITO	HUILA
MONTENEGRO	QUINDIO
SAN JOSE DEL GUAVIARE	GUAVIARE
AGUACHICA	CESAR
BUGA	VALLE DEL CAUCA
OCAÑA	NORTE SANTANDER
FUSAGASUGA	CUNDINAMARCA
GRANADA	META
QUIMBAYA	QUINDIO
SAN GIL	SANTANDER
PUERTO LOPEZ	META
RIOHACHA	GUAJIRA
CHAPARRAL	TOLIMA
LIBANO	TOLIMA
ENVIGADO	ANTIOQUIA
NILO	CUNDINAMARCA
LA PLATA	HUILA
FLORENCIA	CAQUETA
SARAVENA	ARAUCA
MALAMBO	ATLANTICO
SAN VICENTE DEL CAGUAN	CAQUETA
PUERTO ASIS	PUTUMAYO
YUMBO	VALLE DEL CAUCA
LA TEBAIDA	QUINDIO
GUAMO	TOLIMA
MAGANGUE	BOLIVAR
VILLETA	CUNDINAMARCA
SOCORRO	SANTANDER
VALLE DEL GUAMUEZ	PUTUMAYO
ORITO	PUTUMAYO
MARIQUITA	TOLIMA
FLORIDA	VALLE DEL CAUCA
LETICIA	AMAZONAS
LEBRIJA	SANTANDER
LA DORADA	CALDAS

MUNICIPIO	DEPARTAMENTO
SANTA ROSA DEL SUR	BOLIVAR
LERIDA	TOLIMA
MAICAO	GUAJIRA
SAMPUES	SUCRE
LA MESA	CUNDINAMARCA
TIBU	NORTE SANTANDER
CANDELARIA	VALLE DEL CAUCA
AGUSTIN CODAZZI	CESAR
MOCOA	PUTUMAYO
SAN MARCOS	SUCRE
BUENAVENTURA	VALLE DEL CAUCA
TUMACO	NARIÑO

Tabla C-4.: Municipios que conforman el clúster B.

Bibliografía

[Breslow, 1984] Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, 33:38–44.

[Cameron and Trivedi, 2013] Cameron, A. and Trivedi, P. (2013). *Regression Analysis of Count Data*. Cambridge: Cambridge University.

[Cassab et al., 2010] Cassab, A., Morales, V., and Mattar, S. (2010). Factores climáticos y casos de Dengue en Montería, Colombia. 2003-2008. *Revista de Salud Pública*, 13(1):115–128.

[Castrillón et al., 2015] Castrillón, J., Castaño, J., and Urcuquí, S. (2015). Dengue en Colombia: diez años de evolución. *Revista chilena de infectología*, 32(2):142–149.

[Cepeda, 2011] Cepeda, E. (2011). Generalized spatio-temporal models. *SORT-Statistics and Operations Research Transactions*, 35(2):165–178.

[Crawley, 2013] Crawley, M. (2013). *The R book*. John Wiley Sons.

[Dobbie and Welsh, 2001] Dobbie, M. and Welsh, A. (2001). Theory & Methods: Modelling Correlated Zero-Inflated Count Data. *Australian and New Zealand Journal of Statistics*, 43:431–444.

[Días and Morales, 2012] Días, L. and Morales, M. (2012). *Análisis Estadístico de Datos Multivariados*. Universidad Nacional de Colombia.

[Fitzmaurice et al., 2011] Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*. Jhon Wiley and Sons.

[Genolini et al., 2015] Genolini, C., Alacoque, X., Sentenac, M., and Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65(4):1–34.

[Gutierrez et al., 2020] Gutierrez, H., Medina, S., Zapata, C., and Chua, J. (2020). Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country. *Tropical Medicine and Infectious Disease*, 5:156.

[Hall, 2000] Hall, D. (2000). Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, 56:1030–1039.

[Hartig, 2020] Hartig, F. (2020). Dharma: residual diagnostics for hierarchical (multi-level/mixed) regression models. *University of Regensburg*.

[Hedeker and Gibbons, 2006] Hedeker, D. and Gibbons, R. (2006). *Longitudinal Data Analysis*. Wiley-Interscience.

[Heilbron, 1989] Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. Technical report, Department of Epidemiology and Biostatistics, University of California, San Francisco.

[Hossain and Howlader, 2015] Hossain, S. and Howlader, H. (2015). Estimation Techniques for Regression Model with Zero-inflated Poisson Data. *International Journal of Statistics and Probability*, 4:64–76.

[Hur et al., 2002] Hur, K., Hedeker, D., Henderson, W., Khuri, S., and Daley, J. (2002). Modeling Clustered Count Data with Excess Zeros in Health Care Outcomes Research. *Health Services & Outcomes Research Methodology*, 3:5–20.

[King, 1989] King, G. (1989). Event Count Models for International Relations: Generalizations and Applications. *Health Services & Outcomes Research Methodology*, 33(2):123–147.

[Lambert, 1992] Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometric*, 34(1):1–14.

[Lawless, 1987] Lawless, J. (1987). Negative Binomial and Mixed Poisson Regression. *Canad. J. Statist*, 15:209–225.

[López et al., 2012] López, L., A, M., Olivar, G., and Betancourt, J. (2012). Modelo matemático para el control de la transmisión del Dengue. *Revista de Salud Publica*, 14(3):512–523.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. CRC press.

[Medina and Ramos, 2017] Medina, Y. and Ramos, J. (2017). Modelo matemático que explica mejor la afectación e identifica el patrón relevante en la difusión para el dengue en la zona urbana del municipio de Neiva. *Entornos*, 30(2):121–131.

[Mendez et al., 2012] Mendez, J., Usme-Ciro, J., Domingo, C., and et al (2012). Phylogenetic reconstruction of dengue virus type 2 in colombia. *Virology Journal*, 9:64.

[Min and Agresti, 2005] Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5:1–19.

[Molenberghs and Verbeke, 2005] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer series in Statistics.

[Monod, 2014] Monod, A. (2014). Random Effects Modeling and the Zero-Inflated Poisson Distribution. *Communications in Statistics—Theory and Methods*, 43:664–680.

[Mullahy, 1986] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–355.

[Nieuwenhuis et al., 2012] Nieuwenhuis, R., Grotenhuis, M., and Pelzer, B. (2012). influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *The R Journal*, 4(2):38–47.

[Ochoa et al., 2015] Ochoa, M., Casanova, M., and Díaz, M. (2015). Análisis sobre el dengue, su agente transmisor y estrategias de prevención y control. *Revista Archivo Médico de Camagüey*, 19(2):189–202.

[Pinheiro and Bates, 1995] Pinheiro, J. and Bates, D. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

[Press et al., 1992] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C: The art of scientific computing*. Cambridge University Press (second ed.).

[Rose et al., 2013] Rose, C., Martin, S., Wannemuehler, K., and Plikaytis, B. (2013). On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics*, 16:463–481.

[Rúa et al., 2013] Rúa, G., Suarez, C., Chauca, J., Ventosilla, P., and Almanza, R. (2013). Modelado del efecto de la variabilidad climática local sobre la transmisión de dengue en Medellín (Colombia) mediante análisis de series temporales. *Biomédica*, 33(1):142–152.

[Siddiqui, 1996] Siddiqui, O. (1996). *Modeling clustered count and survival data with an application to a school based smoking prevention study*. Tesis de doctorado, University of Illinois at Chicago.

[Simmons et al., 2012] Simmons, C., Farrar, J., Chau, N., and Wills, M. (2012). Dengue. *New England Journal of Medicine*, 366(15):1423–1432.

[Snijders and Bosker, 1999] Snijders, T. and Bosker, R. (1999). *Multilevel analysis, an introduction to basic and advanced multilevel modelling*. SAGE Publications.

[Velandia and Castellanos, 2011] Velandia, M. and Castellanos, J. (2011). Virus del dengue: estructura y ciclo viral. *Infectio*, 15(1):33–43.

[Vélez et al., 2006] Vélez, S., Núñez, C., and Ruiz, D. (2006). Hacia la construcción de un modelo de simulación de la transmisión del dengue en Colombia. *Revista EIA*, 5:22–43.

[Yau and Lee, 2001] Yau, K. and Lee, A. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, 20:2907–1920.

[Zhu et al., 2015] Zhu, H., Luo, S., and DeSantis, S. (2015). Zero-inflated count models for longitudinal measurements with heterogeneous random effects. *Statistical Methods in Medical Research*, 0:1–16.

[Álvarez and Vargas, 2019] Álvarez, A. and Vargas, R. (2019). Dengue: presentación e importancia de factor activación de plaquetas en la evolución de la fase crítica. *Revista Médica Sinergia*, 4(11):e294.